

# Data Science

# Building a Complex Model for Speaker Recognition Using Speech Signal Segmentation

Gani Esen<sup>1</sup>, and Marat Nurtas<sup>1</sup>

<sup>1</sup> International Information Technology University, Almaty, Kazakhstan

## Abstract

Speech segmentation is an essential stage in designing automatic speech recognition systems. It is a difficult problem, as speech is immensely variable. Our aim was to design an algorithm that could be employed at the stage of automatic speech recognition. This would allow some issues with speech signal parametrization to be avoided. The intent of this thesis was to develop a new unsupervised automatic speech signal segmentation algorithm. It used second-order statistics of a speech signal to perform segmentation without knowing anything about the phonetic content of the utterances. The proposed method began with a nonlinear model based on deep Gaussian processes. The benefits of these methods, which are all generative in the sense that novel signals can be sampled from the underlying models, allow us to consider how well the extent to which they encode important perceptual characteristics of sound. First, we introduced a massive audio-visual speaker recognition dataset derived from open-source media. Second, we developed and compared Convolutional Neural Network (CNN) models and training strategies that can effectively recognize identities from voice under various conditions. To account for the properties of human hearing, inter-phoneme boundaries were detected using statistics defined on the mel spectrum mel spectrum determined from the reflection coefficients. The thesis presents the structure of the algorithm, defines its properties, lists parameter values, describes detection efficiency results, and compares them with those for another algorithm. The obtained segmentation results were satisfactory.

## Keywords

Convolutional neural network (cnn), Segmentation, Deep Gaussian processes, Mel spectrum

## 1. Introduction

Speaker recognition is the process of self-acting identification of the speaker based on personal information included in speech waves. This method uses the speaker's identity by using his voice and ensures access control to these services, which include voice kit, database access services, information services, voice mail, security management for sensitive information areas, and a variety of other areas where security is valued [1].

Speech is a complex sign formed by a series of transformations that occur at several different levels: semantic, linguistic, articulatory and acoustic [2]. Differences in these transformations are reflected in differences in the acoustic qualities of the speech signal. In addition, there are speaker differences that are thought to be the result of a combination of anatomical differences inherent in the vocal tract and different individuals' learned speech abilities. In speaker recognition, all these differences are provided for and applied in order to distinguish between speakers. Computers can recognise human voices, receive commands, dictate sentences containing words in their memory, and recognise context in sentences partially [2]. Furthermore, based on the above developments, you can control your computer with a human voice and control all of your home's lights, as well as the air conditioning and TV, by connecting your computer to certain electronic devices in your home. Several articles on the topic of speech recognition can be found in research in this area. Today, there is a great deal of interest in the field of speaker recognition, owing to the high demand for security and voice access applications.

There are several subfields of speaker recognition. However, they can be divided into three major categories: speaker identification, speaker verification, and speaker classification. For the first field, the system attempts to recognise an unknown person's voice among many other people's voices [3]. During the speaker check, the voice of the speaker whose identity is being verified and the identity voice are compared, and the system ends the process by accepting or rejecting the decision based on

the threshold value [4]. The goal of speaker verification is to verify a claimed identification based on voice signal measurements. Entry control to limited locations, access to privileged information, cash transfer, credit card authorization, voice banking, and other transactions can all benefit from speaker verification [5]. There are two types of voice authentication: one verifies a talker's identity based on the content of the spoken password or pass-phrase, such as a personal identification number (PIN), social security number, or mother's maiden name; the other verifies a talker's identity based on the content of the spoken password or pass-phrase, such as a personal identification number (PIN), social security number, or mother's maiden name. The test phrase in the former scenario may be out in the open and even shared by other talkers in the population, whereas the password information in the later case is considered to be known only by the authorized talker [8]. In order to classify speakers, certain characteristics such as age, gender, and so on are used. In addition, speaker recognition can be implemented in two ways: text independent and text dependent. The system handles the voice characteristics in the first case. In the second case, the system is interested not only in the voice characteristics, but also in the spoken words [6].

The following chapters are about getting started, how to build a simple and representative system of self-acting determination of the determiner. This system of defining definitions fulfills the overarching goal of identifying the definer, which forms the broad scope of the definition of the definer. The developed system has the potential to be used in a variety of security applications. Examples may find themselves needing to say their own credit card number over the telephone in order to get a random person or to gain access to laboratories with the support of voices [7]. The system has the ability to increase the auxiliary degree of security by verifying the singing properties of the input expression with the support of the system's self-acting definition of the proposed, most likely the one we describe.

## **2. Analytical Review**

Over the past several decades people in Kazakhstan were habituated to typing text to search online. However, it is very usual for them to speak than to type or click providing a sky-high trend of voice search and making it more significant. With the outbreak of the COVID-19 pandemic in Kazakhstan, technologies that avoid physical contact like voice-activated assistants help out in reducing or eliminating touch points and thereby increase safety. With the expanding dominance of smartphones, voice search has been immovably growing. The presence of voice interfaces will open access to public services to people with disabilities, as well as to the population living in remote regions and having the only way to access them in the form of phones. On the other hand, the introduction of voice contact centers can reduce the cost of operators required to serve the population, since, the system can operate automatically without human intervention [9]. In addition, speech recognition systems can also be used offline to analyze and collect information on audio data, including monitoring telephone conversations to prevent terrorist attacks and sabotage.

During the peak of the coronavirus pandemic in March 2020, many residents of Kazakhstan were left without work. The Government of Kazakhstan organized payments in the amount of 42,500 tenge as financial support. These payments were made through the electronic portal Egov.kz. More than a million people have tried to receive these payments, as a result of which the electronic portal was overloaded and despite the fact that a huge number of operators were involved. However, many people could not get advice and a solution to the problem. Thus, for people who use this platform, voice recognition and voice assistant allows to recognize speech in the Kazakh language and it will significantly reduce the time and labor costs, it will optimize the system for obtaining advice and assistance, as well as will effectively build a communication with people with disabilities. Since service such as Egov.kz in a high demand we had to find out whether people needed voice recognition feature.

## **3. Implementation Part**

### **3.1. Survey results**

According to a study by the Egov.kz portal, more than 62% of the total population of Kazakhstan

use electronic public services (according to 2021 data). Considering the population of the Republic of Kazakhstan (about 19,545,535 people according to 2022 data), the potential audience may be 12 million 118 thousand 231 people, but this is the number of users of the Egov portal. Of this number of users, many simply go to personal accounts and receive the service online, so at the very beginning, approximately 1% of users will use this service, and this in turn is about 120 thousand people. In general, this is a large segment, which in the future can bring good income, if this social project is successfully promoted.

With an increase in users of speech recognition services, registration for various public services through the portal will only become more relevant and more convenient for users every day. Since our application is highly specialized, it has a significant usability advantage. Therefore, there is only one risk in this project - it is the accuracy of speech recognition, but we solved this problem in the technical part. According to our calculations, the new service can attract the attention of new users and improve their user experience.

After conducting a survey, where people between 25-40 years old participated it was found out that more than 75% of people were using Egov and a voice assistant [10] (Figure 1).



Figure 1: Survey result in Pie chart form

The survey showed that vast majority of people who participated in survey were endorsing the idea of having a voice assistant.

### 3.2. Feature extraction

It is essential to convert the speech waveform into a set of features, or rather feature vectors, for subsequent analysis (at a much lower information rate). The signal-processing front end is what it's called. The voice signal is a quasi-stationary signal, which means it varies slowly in time. Figure 2 depicts an example of a speech signal.

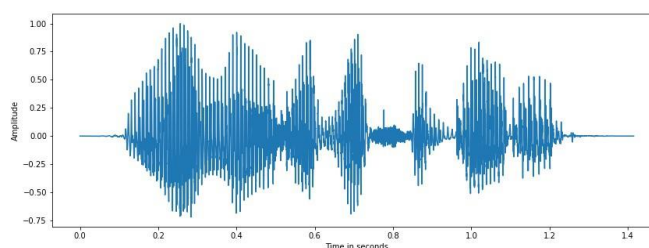


Figure 2: Speech Signal [1]

The properties of the voice signal are fairly steady when analyzed over a short period of time (for

example, between 5 and 100 ms). Signal properties, on the other hand, tend to change over longer periods of time (on the order of 1/5 second or more). This reflects the many spoken utterances that are being said. As a result, the most popular method of characterizing a speech signal is short-time spectrum analysis [11]. There are numerous approaches for parametrically encoding the voice signal for the purpose of speaker recognition. Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), Cepstral Coefficients using DCT, and other techniques are among them. It was employed MFCC to the project.

### 3.3. Truncation

The wavread command's default sampling frequency is 44100 Hz. When a two-second audio clip is captured, the number of samples created is roughly 90000, which is far too many to process. As a result, we can truncate the signal by choosing a specific threshold value. When the signal gets above the value while traversing the time axis in a positive direction, we can indicate the start of the signal. Similarly, we may get the signal's termination by performing the above process in the other way. The consequence of truncating a voice signal is seen in (Figure 3) [5].

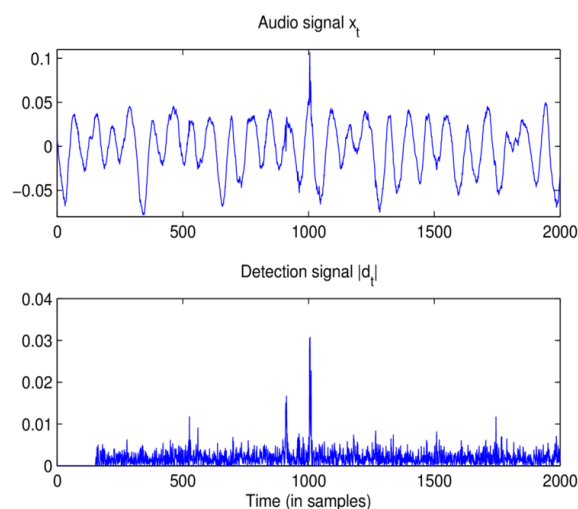


Figure 3: Truncated version of the audio signal

### 3.4. Frame Blocking

The continuous voice signal is segmented into N-sample frames in this step, with neighboring frames separated by M samples of a value M less than N. The first N samples make up the first frame. The second frame follows the first by M samples and overlaps it by N - M samples, and so on. This technique is repeated until all of the speech has been captured in one or more frames. The values of M and N have been set to N = 256 and M = 128 correspondingly. The frame output of the truncated signal is shown in (Figure 4).

### 3.5. Frame Blocking

The continuous voice signal is segmented into N-sample frames in this step, with neighboring frames separated by M samples of a value M less than N. The first N samples make up the first frame. The second frame follows the first by M samples and overlaps it by N - M samples, and so on. This technique is repeated until all of the speech has been captured in one or more frames. The values of M and N have been set to N = 256 and M = 128 correspondingly. The frame output of the truncated signal is shown in (Figure 4).

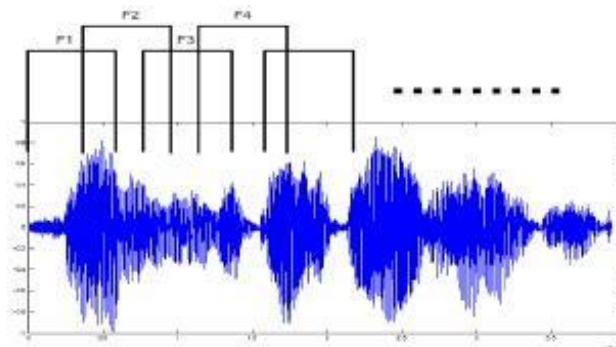


Figure 4: Frame Output

### 3.6. Windowing

To recognize certain types of patterns Windowing is used. In Figure 5 we can observe how Hamming window works.

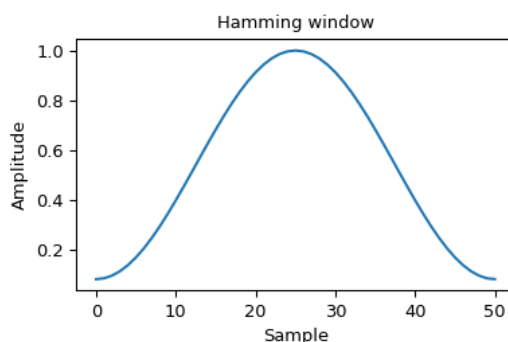


Figure 5: Hamming window

The following step is to window each individual frame to reduce signal discontinuities at the start and finish of each frame. The idea is to use the window to taper the signal to zero at the start and end of each frame to reduce spectral distortion. The signal is the outcome of windowing if the window is defined as  $w(n)$ ,  $0 \leq n < N - 1$ , where  $N$  is the frame length.

### 3.7. Mel frequency wrapping

In the pre-processing step frame blocking, windowing, and FFT were applied to the continuous speech signal, as shown in the block diagram (Figure 7). The signal's spectrum is the result of the last stage. Human perception of frequency content of sounds for speech signals does not follow a linear scale, according to psychophysical studies. A subjective pitch is measured on a scale called the "mel scale" for each tone with an actual frequency  $f$ . The mel-frequency scale has linear frequency spacing below 1 KHz and logarithmic frequency spacing above 1 KHz.

The Mel Frequency Scale given by

$$F_{mel} = \left( \frac{1000}{\log(2)} \right) \times \log \left( \frac{1 + f}{1000} \right). \quad (1)$$

For fetching data from collections it is better to use for each since it enumerates the children's elements in the DataSnapshot to their query. Surprisingly when try to use a map to get data from the events database, it does not collect data correctly. The employment of a filter bank spaced equally on the mel-scale is one method of replicating the subjective spectrum. The frequency response of the filter bank is triangular band pass. A constant mel frequency interval determines the spacing and bandwidth.

K is equal to 20 mel spectrum coefficients. In the frequency domain, using this filter bank essentially means applying triangle-shape windows to the spectrum. Each filter can be thought of as a histogram bin in the frequency domain (with bins that overlap). A mel-spaced frequency bank in Figure 7. Example of Mel-spaced frequency bank shown in Figure 7.

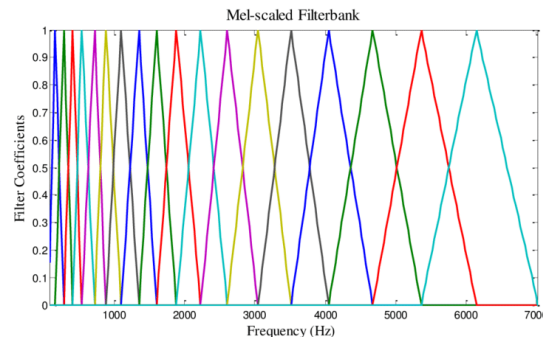


Figure 6: Hamming window

### 3.8. MFCC (mel frequency cepstral coefficients)

The block diagram of an MFCC processor is shown in Figure 7. MFCCs are based on the known fluctuation of the critical bandwidths of the human ear with frequency. The phonetically relevant aspects of speech have been captured using filters spaced linearly at low frequencies and logarithmically at high frequencies. The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic frequency spacing beyond 1000 Hz.

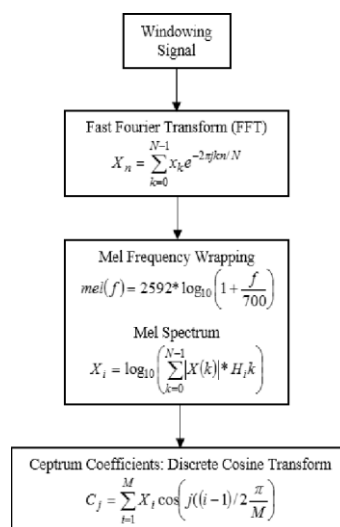


Figure 7: MFCC[3]

Mel Frequency Cepstral Coefficients (MFCC) are perception-based coefficients that represent audio. It is derived from the audio clip's Fourier Transform (FFT) or Discrete Cosine Transform (DCT). The main difference between the FFT/DCT and the MFCC is that the frequency bands in the MFCC are positioned logarithmically (on the Mel scale), which more closely approximates the human auditory system's response than the FFT/linearly DCT's spaced frequency bands. This enables more efficient data processing.

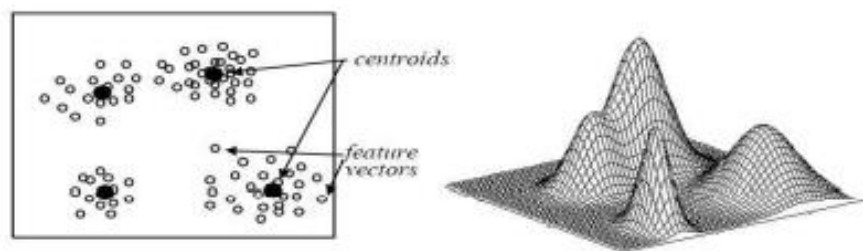
## 4. GMM (Gaussian mixture modeling)

This is one of the non-parametric strategies for identifying speakers. After clustering, feature vectors are shown in d-dimensional feature space and mimic Gaussian distribution. It means that each related cluster can be seen as a Gaussian probability distribution, with the probability values best representing the cluster's attributes. The only challenge is efficient feature vector categorization. Two facts motivate the use of Gaussian mixture density for speaker identification.

They are:

- A set of acoustic classes is represented by individual Gaussian classes. These acoustic classes represent information about the vocal tract.
- Gaussian mixture density approximates the distribution of feature vectors in multi-dimensional feature space using a smooth approximation.

In Figure 8 a feature space related to Gaussian model centroids and feature vectors are represented.



**Figure 8:** A feature space and the related Gaussian model are shown in the GMM model

Gaussian mixture density uses a smooth approximation to approximate the distribution of feature vectors in multi-dimensional feature space.

In Figure 9 the Gaussian mixture density is the weighted necessary sum of the M component densities, and it is directed by the equation [4]:

$$P(x | \lambda) = \sum_{k=1}^M w_k \mathcal{N}(x | \mu_k, \Sigma_k)$$

M Components
Weights
Gaussian density

**Figure 9:** Gaussian mixture density

Thus, we assign powers to any Gaussian density, so that the sum of the weights is equal. GMM Features:

- weights associated with each component;
- mean vectors of any component;
- covariance matrices of any component.

Following the acquisition of the feature vectors, the next step is to categorize them into different Gaussian components. However, we don't know the mean or co-variance of the components at first. As a result, proper vector classification is impossible. The Expectation Maximization technique is used to maximize the classification process for a given set of feature vectors. This is how the algorithm works:

- initial values for  $\mu_i$ ,  $\Sigma_i$  and  $w_i$  are assumed.
- then, using the following formula, we repeatedly calculate the next values of mean, covariance, and mixture weights, maximizing the probability of classification of the collection of T feature vectors.

Formulas are used:

Mixture Weights:

$$\rho = \frac{1}{T} \sum_{t=1}^T \rho(i) | x_t, \lambda \quad (2)$$

Means:



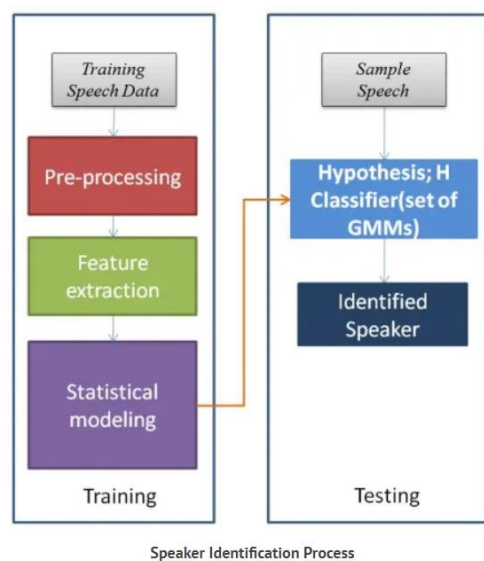
$$\mu_i = \frac{\sum_{t=1}^T \rho(i|x_{t,\lambda})x_t}{\sum_{t=1}^T \rho(i|x_{t,\lambda})} \quad (3)$$

Variances:

$$\sigma_i^2 = \frac{\sum_{t=1}^T \rho(i|x_{t,\lambda})x_t^2}{\sum_{t=1}^T \rho(i|x_{t,\lambda})} - \mu_i^{-2} \quad (4)$$

GMM assumes that the vector space is divided into certain components depending on the clustering. Feature vectors and frames the distribution of feature vectors in each component, which should be Gaussians. Since initially we do not know which vector belongs to which component a for optimal classification, a likelihood-maximization algorithm is used. For testing we calculated the posterior probability of the test utterance and the reference speaker maximizing The Gaussian distribution is called the unknown speaker identity. The result of this procedure is given below: There is a lot of information that can be extracted from a speech sample, for example, who is the speaker, what is the gender of the speaker, what is the language being spoken, with what emotion has the speaker spoken the sentence, the number of speakers in the conversation.

It will illustrate the same with a naive approach using Gaussian Mixture Models (GMM). GMM-UBM (Gaussian Mixture Model – Universal Background Model) using MAP (Maximum Aposteriori) adaptation is one of the successful conventional technique to implement speaker identification. I-vectors based speaker identification is the state-of-the-art technique implemented in lot of voice biometric products. A GMM will take as input the MFCCs and derivatives of MFCCs of the training samples of a speaker and will try to learn their distribution, which will be representative of that speaker. A typical speaker identification process can be shown by flow in Figure 10.



**Figure 10:** Process of Speaker Identification

Figure 11 illustrates the process of speaker identification. First of all, our audio samples are getting trained, then we apply pre-processing and then we extract feature and apply GMM modeling.

While testing when the speaker of a new voice sample is to be identified, first the 40-dimensional feature (MFCCs + delta MFCC) of the sample will be extracted and then the trained speaker GMM models will be used to calculate the scores of the features for all the models. Speaker model with the maximum score is predicted as the identified speaker of the test speech. Having said that we will go through the python implementation of the following steps:

- 40-Dimensional Feature Extraction;
- Training Speaker Models;

- Evaluating Performance on test set.

MFCC extraction pseudocode

DEFINE FUNCTION extract\_features(audio,rate):

```
"""extract 20 dim mfcc features from an audio, performs CMS and combines
delta to make it 40 dim feature vector"""
SET mfcc_feat TO mfcc.mfcc(audio,rate, 0.025, 0.01,20,appendEnergy TO True)
SET mfcc_feat TO preprocessing.scale(mfcc_feat)
SET delta TO calculate_delta(mfcc_feat)
SET combined TO np.hstack((mfcc_feat,delta))
```

RETURN combined

## 5. Training the model

### 5.1. Loading a dataset and training speaker models

For this project KSC dataset was used. KSC is a massive dataset. From that dataset 3333 audio samples were used. This dataset contains 30 speakers from different regions of Kazakhstan with different age and type of device that was used to record audios. Training of the model was done in Jupyter Notebook. Link to the source code [11].

Dataset contains 30 speaker classes. And this is how it looks inside (Figure 11).

metadata_df_train_data									
Unnamed: 0	uttID	deviceID	Gender	Age	Region	Device_Type	Headphones	Speaker_class	
0	0	5f5af4cf7bb2c.wav	12793	1	25	East	phone	0	0
1	1	5f5af4dc52b1f.wav	12793	1	25	East	phone	0	0
2	2	5f5af4f705da7.wav	12793	1	25	East	phone	0	0
3	3	5f5af4fe3e8a2.wav	12793	1	25	East	phone	0	0
4	4	5f5af5084a3be.wav	12793	1	25	East	phone	0	0
...	...	...	...	...	...	...	...	...	...
3328	3328	5f634471ef03e.wav	22108	0	21	South	computer	0	29
3329	3329	5f6344df10511.wav	22108	0	21	South	computer	0	29
3330	3330	5f6345096eabd.wav	22108	0	21	South	computer	0	29
3331	3331	5f634524979a8.wav	22108	0	21	South	computer	0	29
3332	3332	5f63462c63f49.wav	22108	0	21	South	computer	0	29

3333 rows x 9 columns

Figure 11: KSC dataset

The next step is to divide the dataset to test, train and validation datasets. Also unnecessary columns such as Unnamed and DeviceID were dropped, shown in the Figure 12.

```
X_train, y_train, X_valid, y_valid, X_test, y_test = train_valid_test_split(df, target = 'Speaker_class',
                                                                    train_size=0.6, valid_size=0.1, test_size=0.3)

print(X_train.shape), print(y_train.shape)
print(X_valid.shape), print(y_valid.shape)
print(X_test.shape), print(y_test.shape)

(1999, 13)
(1999,)
(333, 13)
(333,)
(1001, 13)
(1001,)

(None, None)

X_train.drop(['Unnamed: 0', 'deviceID'], axis=1, inplace=True)
```

Figure 12: Splitting dataset

A Gaussian mixture model is a probabilistic clustering model for representing the presence of sub-populations within an overall population. The idea of training a GMM is to approximate the probability distribution of a class by a linear combination of ‘k’ Gaussian distributions/clusters, also called the components of the GMM. The likelihood of data points (feature vectors) for a model is given by following equation:

$$P(X|\lambda) = \sum_{k=1}^k \omega_k P_k(X|\mu_k, \Sigma_k) \quad (5)$$

Initially, it identifies k clusters in the data by the K-means algorithm and assigns equal weight  $w=1/k$  to each cluster. ‘k’ gaussian distributions are then fitted to these k clusters. The parameters  $\mu$ ,  $\sigma$  and  $w$  of all the clusters are updated in iterations until the converge. The most popularly used method for this estimation is the Expectation Maximization (EM) algorithm.

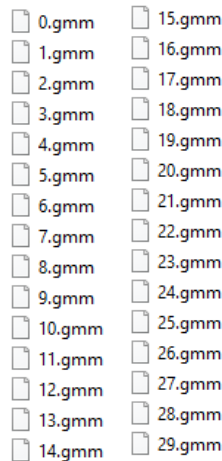
Python’s sklearn.mixture package is used to learn a GMM from the features matrix containing the 40 dimensional MFCC and delta-MFCC features.

Once code is run it identifies every speaker and corresponding audio of them as you can see below on the Figure 13.

```
62 0 62
+ modeling completed for speaker: 0.gmm with data point = (39195, 40)
124 1 124
+ modeling completed for speaker: 1.gmm with data point = (116380, 40)
137 2 137
+ modeling completed for speaker: 2.gmm with data point = (129787, 40)
57 3 57
+ modeling completed for speaker: 3.gmm with data point = (43942, 40)
35 4 35
+ modeling completed for speaker: 4.gmm with data point = (23402, 40)
116 5 116
+ modeling completed for speaker: 5.gmm with data point = (68169, 40)
20 6 20
+ modeling completed for speaker: 6.gmm with data point = (15134, 40)
71 7 71
+ modeling completed for speaker: 7.gmm with data point = (45988, 40)
55 8 55
+ modeling completed for speaker: 8.gmm with data point = (35088, 40)
56 9 56
+ modeling completed for speaker: 9.gmm with data point = (43441, 40)
45 10 45
+ modeling completed for speaker: 10.gmm with data point = (25963, 40)
59 11 59
+ modeling completed for speaker: 11.gmm with data point = (41196, 40)
58 12 58
+ modeling completed for speaker: 12.gmm with data point = (44429, 40)
51 13 51
+ modeling completed for speaker: 13.gmm with data point = (35410, 40)
49 14 49
+ modeling completed for speaker: 14.gmm with data point = (33979, 40)
214 15 214
+ modeling completed for speaker: 15.gmm with data point = (155962, 40)
63 16 63
+ modeling completed for speaker: 16.gmm with data point = (36571, 40)
32 17 32
+ modeling completed for speaker: 17.gmm with data point = (21722, 40)
68 18 68
+ modeling completed for speaker: 18.gmm with data point = (49645, 40)
66 19 66
```

**Figure 13:** GMM models for speakers

The code is run once for each speaker and train\_file is variable which has text filename containing path to all the audios for the respective speaker as you see in Figure 14. Also, the “speaker\_models” directory was created, where all the models will be dumped after training.



**Figure 14:** Saved GMM models in directory

As we can see in the folder “speaker\_model” 30 gmm trained models were created for the future speech recognition. It greatly aids with more accurate training models, therefore our model will be more accurate in detecting and verifying the right speaker.

First of all, we check on test data. You can see the results of test in Figure 15. The dataset was divided into 3 parts and test data makes up for approximately 1000 audio samples.

```
----- -- --
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29
Testing Audio : 29 --
detected as - 29

print (error, total_sample)
accuracy = ((total_sample - error) / total_sample) * 100
print ("The Accuracy Percentage for the current testing Performance with MFCC + GMM is : ", accuracy, "%")

21.0 1001.0
The Accuracy Percentage for the current testing Performance with MFCC + GMM is : 97.9020979020979 %
```

**Figure 15:** Accuracy of our model

GMM model has successfully done the job. The trained model has tested around 1000 audios and made mistake only in 21 audio samples and got the very high accuracy which is equal to 97.9%.

## 5.2. Creating CNN model

In Figure 18 we can see the implementation of CNN.

```
class CNNNetwork(nn.Module):
    def __init__(self):
        super().__init__()

        self.conv1 = nn.Sequential(
            nn.Conv2d(
                in_channels=1,
                out_channels=16,
                kernel_size=3,
                stride=1,
                padding=2
            ),
            nn.ReLU(),
            nn.MaxPool2d(kernel_size=2)
        )
        self.conv2 = nn.Sequential(
            nn.Conv2d(
                in_channels=16,
                out_channels=32,
                kernel_size=3,
                stride=1,
                padding=2
            ),
            nn.ReLU(),
            nn.MaxPool2d(kernel_size=2)
```

Figure 16: CNN for speaker identification

Overall training time and accuracy shown in the Figure 17.

Classifier	Train accuracy	Test accuracy	Training time
CNN	70.1%	68.2%	01:56:43

Figure 17: CNN accuracy

### 5.3. Model comparisons

There were used another models for speaker identification, but most of them showed low accuracy as you can see in the Figure 20.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis

classification_models = [
    KNeighborsClassifier(),#(3),
    SVC(kernel='linear'),#, C=0.025),
    SVC(kernel='rbf'),
    DecisionTreeClassifier(),#max_depth=5),
    RandomForestClassifier(),#max_depth=5, n_estimators=10, max_features=1),
    AdaBoostClassifier(),
    QuadraticDiscriminantAnalysis()]

scores = []
for model in classification_models:
    model.fit(X_train_scaled, y_train)
    score = model.score(X_test_scaled, y_test)
    model_name = type(model).__name__
    if model_name=='SVC' and model.kernel=='rbf': model_name+=' RBF kernel'
    scores.append((model_name,(f'{100*score:.2f}%'))))
# Make it pretty
scores_df = pd.DataFrame(scores,columns=['Classifier', 'Accuracy Score'])
scores_df.sort_values(by='Accuracy Score',axis=0,ascending=False)
```

	Classifier	Accuracy Score
4	DTW	82%
5	Vector Quantization	74.28%
3	SVC	68.19%
0	SVC RBF kernel	52.70%
1	RandomForestClassifier	48.81%
2	KNeighborsClassifier	47.92%
6	DecisionTreeClassifier	34.00%
7	AdaBoostClassifier	31.70%

Figure 18: Other model's accuracy

Classifier	Train accuracy	Test accuracy	Training time
SVC RBF kernel	52.70%	49.21%	0:04:43
RandomForestClassifier	48.61%	39.3%	0:02:22
KNeighborsClassifier	47.92%	31%	0:03:20
SVC	68.19%	65.21%	0:05:07
DTW	82%	81.3%	0:18:59
Vector Quantization	74.28%	70.7%	0:42:08
<b>GMM</b>	<b>100%</b>	<b>97%</b>	<b>0:08:13</b>
DecisionTreeClassifier	34.00%	27.56%	0:16:11
AdaBoostClassifier	31.70%	11.3%	0:02:14
CNN	70.1%	68.2%	01:56:43

Figure 19: All models

The overall time spent, and accuracy of all models used for speaker identification is shown in Figure 19.

## 6. Conclusion

Speech recognition is a critical component of complicated speaker recognition systems, and it is often done using ML techniques and deep neural networks. There is a possibility to mix multiple speaker identification technologies, algorithms, and tools to increase the performance of your speech processing system, depending on the complexity of the work at hand.

This paper has given a thorough overview of deep learning-based on speaker recognition. The relationship between various subtasks and speaker verification and identification, and compiled a list of common difficulties were examined. In addition to that the loss functions of end-to-end speaker verification for feature learning from the standpoint of various training sample construction methods in particular were examined. It was investigated three types of deep learning-based domain adaptation methods, as well as several speech preprocessing methods that deal with domain mismatch and background noise, respectively, for robust speaker recognition. To summarize, deep learning has significantly improved the performance of speaker recognition. So it was done the best to summarize the recent rapid progress of deep learning-based speaker recognition; hopefully, this serves as a knowledge resource and contributes to the growth of the research community. Although deep learning-based speaker recognition has been a huge success, there are still many issues that need to be addressed.

Despite recent advancements, there is still a great deal of work to be done. Although incremental improvements in all areas of speaker recognition are being made on a daily basis, channel and audio type mismatch appears to be the most significant barrier to achieving perfect speaker recognition results. It should be noted that ideal results are asymptotes that will almost certainly never be achieved. At its core, as the population of speakers in the database grows, variation within speakers outnumbers variation between speakers. This is the most common source of error in large-scale speaker recognition. In fact, most other directions can be considered trivial if large-scale speaker identification approaches acceptable results. However, this is a complex problem that will undoubtedly take much more time to perfect, if at all possible. Meanwhile, it appears to be in the early stages of large-scale identification.

Considering the various classifications and modules of an automatic speaker recognition system in this diploma work. For each 30 second speech frame, a set of mel-frequency cepstrum coefficients (MFCC) is computed. An acoustic vector is a set of coefficients. As a result, each input speech utterance is converted into a series of acoustic vectors. Despite the fact that many loss functions have been proposed, there is a lack of a strong theoretical foundation for the success of the loss functions, as well as theoretical guidance that could lead to better loss functions. Although the verification losses for end-to-end speaker verification closely match the verification process, their full potential has not yet been realized. So choosing the GMM method of speaker identification. are motivated by the fact that a speaker's vocal tract information follows a Gaussian distribution and that a Gaussian model approximates the feature space as a smooth surface. GMM's accuracy for the same data set was 89

percent, indicating its high efficiency. The computational efficiency suffers slightly as the number of components used increases. While striving for high accuracy, however, these flaws may be compromised. were examined.

## 7. References

- [1] Beginner's guide to Speech Analysis. Published on: January 3, 2019. URL: <https://towardsdatascience.com/beginners-guide-to-speech-analysis-4690ca7a7c05>.
- [2] What is the Fast Fourier Transform? Published on: December 30, 2019. URL: <https://towardsdatascience.com/fast-fourier-transform-937926e591cb>.
- [3] Infant's cry sound classification using Mel-Frequency Cepstrum Coefficients feature extraction and Backpropagation Neural Network. doi: 10.1109/ICSTC.2016.7877367.
- [4] The process of training a Gaussian mixture model using EM. Published on: January 3, 2020. URL: <https://towardsdatascience.com/expectation-maximization-for-gmms-explained-5636161577caD>.
- [5] A method for recovering audio signals corrupted by impulsive noise such as clicks, splashes, or scratches. Published on 13 November, 2015. URL: [https://www.researchgate.net/figure/Example-of-detection-signal-for-real-life-audio-signal-of-classical-music-corrupted-by\\_fig3\\_284277095](https://www.researchgate.net/figure/Example-of-detection-signal-for-real-life-audio-signal-of-classical-music-corrupted-by_fig3_284277095).
- [6] Dipanjan Nandi, Debadatta Pati, K. Sreenivasa Rao – Sub-segmental, Segmental and Supra-segmental Analysis of Linear Prediction Residual Signal for Language Identification – Bangalore, India 2014. doi:10.1109/SPCOM.2014.6983974.
- [7] I. Mohamed Kalith, Samantha Thelijjagoda, David Asirvatham – Isolated to Connected Tamil Digit Speech Recognition System Based on Hidden Markov Model – 2016. URL: [https://www.researchgate.net/publication/299562718\\_Isolated\\_to\\_Connected\\_Tamil\\_Digit\\_Speech\\_Recognition\\_System\\_Based\\_on\\_Hidden\\_Markov\\_Model](https://www.researchgate.net/publication/299562718_Isolated_to_Connected_Tamil_Digit_Speech_Recognition_System_Based_on_Hidden_Markov_Model).
- [8] A. Lakshmi, Hema A Murthy – Syllable based Continuous Speech Recognition for Tamil Language – January 2016. doi: 10.1007/s10772-009-9058-0.
- [9] Kuldeep Kumar, Rajesh Aggarwal – Automatic Speech Recognition System for Isolated and Connected Words for Hindi Language by using HTK – 2012. doi: 10.1504/IJCSYSE.2012.044740.
- [10] Link to survey. URL: [https://docs.google.com/forms/d/e/1FAIpQLSerH1JuG1Rc1dG8yRUIJw8otr98EGV2CwWzoK6HrAt-DE5LfA/viewform?usp=sf\\_link](https://docs.google.com/forms/d/e/1FAIpQLSerH1JuG1Rc1dG8yRUIJw8otr98EGV2CwWzoK6HrAt-DE5LfA/viewform?usp=sf_link). Link to the source code. URL: <https://github.com/ganiesenov/Kazakh-Speech-Identification>.
- [11] Ms. VimalaC – Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM – 2012. doi: 10.1016/J.PROENG.2012.01.968.
- [12] K.P. Unnikrishnan, J.J. Hopfield, D.W. Tank – Speaker-Dependent Speech Recognition Using a Neural Network with TimeDelayed Connections 26 – March 3, 1991. doi: 10.1109/78.80888.
- [13] A. Desai Vijayendra – MFCC and Real Cepstral Coefficient (RC) Conjugate Gradient (CG) Algorithm and Levenberg-Marquardt (LM) Algorithm–2016. doi: 10.1016/j.procs.2016.07.259.
- [14] Yogesh K. Gedam, Sujata S Magare – Frame length-20ms Pre-Emphasis- FIR filter Hamming Window Mel Frequency Wrapping Discrete Cosine Transform (DCT) VQ DTW– March 2014. URL:[https://www.researchgate.net/publication/270899179\\_Development\\_of\\_Automatic\\_Speech\\_Recognition\\_of\\_Marathi\\_Numerals-A\\_Review](https://www.researchgate.net/publication/270899179_Development_of_Automatic_Speech_Recognition_of_Marathi_Numerals-A_Review).
- [15] James Lee-Thorp, Joshua Ainslie – Multi-Layer Feed-Forward Networks Fast Fourier Transform (FFT) Discrete Fourier Transform (DFT) – 2014. URL: <https://doi.org/10.48550/arXiv.2105.0382>.
- [16] Voice recognition trends in 2022 — URL: <https://www.analyticsinsight.net/top-5-voice-recognition-trends-to-look-out-for-in-2022/>.
- [17] Richard Dufour, Richard Dufour, Richard Dufour – Spontaneous Speech Characterization and Detection in Large Audio Database – 2009. URL: <https://cyberleninka.org/article/n/676880>.
- [18] Egov statistics for 2017 year. URL: [https://egov.kz/cms/ru/articles/ecology/waste\\_reduction\\_recycling\\_and\\_reuse](https://egov.kz/cms/ru/articles/ecology/waste_reduction_recycling_and_reuse).

# Forecasting the Potential Scenarios of CO<sub>2</sub> Emissions in Kazakhstan Using Deep Learning (DL) Predictive Models

Kamshat Asmaganbetova<sup>1</sup>, Zhenis Otarbay<sup>2</sup>, Almaz Turginbekov<sup>1</sup>, Damir Karimov<sup>2</sup>,  
Olzhas Sayakov<sup>2</sup>, and Bakhtiyar Kalzhan<sup>2</sup>

<sup>1</sup> Astana IT University, Astana, Kazakhstan

<sup>2</sup> Nazarbayev University, Astana, Kazakhstan

## Abstract

Today, carbon dioxide emissions will play an essential part in the world's energy transformation and environmental revolution for the next decades. The mass quantities of this gas are released into the atmosphere daily, significantly affecting the climate conditions and environmental sustainability—the major contributor of carbon dioxide emission into the atmosphere as fossil fuel burning. Therefore, Kazakhstan, being one of the largest oil producers globally, needs to develop an innovative approach to power progress and cleaner energy solutions to minimize the effect of carbon dioxide emissions on the environment. The project's primary purpose is to forecast the scenarios of carbon dioxide emissions in Kazakhstan using artificial intelligence (AI) predictive models for the following years. With the current technological advances, AI-based techniques are widely implemented for the construction of predictive models. In our case, the input parameters for these predictive models are based on currently available data for carbon dioxide emissions in Kazakhstan. Based on the survey and simulated prediction results, the trend is downward for CO<sub>2</sub> emissions, as the overall attitude of the population in Kazakhstan related to CO<sub>2</sub> emissions is optimistic. The project outcome's significance relies on increasing the energy demand and developing low-carbon energy transformation, solving climate change problems. Consequently, forecasting and monitoring the carbon dioxide emission is the first step to improve the current environmental situation in the industry and consider environmental-friendly society shortly.

## Keywords

Carbon dioxide, emissions, Kazakhstan, prediction, AI, artificial intelligence

## 1. Introduction

Energy demand is continuously increasing worldwide due to the population increase and improvements in quality of life. Energy is used in every aspect of life – lighting, cooling, heating, cooking, etc. Traditional methods of energy production involve the burning of fossil fuels. Between the ages of 2009 and 2018, coal was responsible for emitting the most that is 42% of CO<sub>2</sub> to the atmosphere. Oil took the second place with 34% of produced emissions, followed by 19% that came from natural gas and the rest from other fossils [10]. There are two significant drawbacks of using fossil fuels: 1) they are finite, and 2) they emit adverse chemicals that diffuse with the air. Moreover, the increased concentration of CO<sub>2</sub> in the air leads to climate change. Over the years, energy economics and environmental issues caused by CO<sub>2</sub> emissions have become more prominent [13]. The negative effect of fossil fuels on the environment has been recognized in many countries. Therefore, numerous proposals and policies have been revised to shorten CO<sub>2</sub> emissions. Optimization of power plant operating conditions, increasing efficiency, and replacing fossil fuel sources with renewable ones are encouraged [1]. Carbon dioxide, a component of greenhouse gases, has the most significant impact on climate change. “Fossil fuels account for over 80% of energy consumption across the world” [25]. Besides climate, CO<sub>2</sub> influences social, political, and economic issues. Several studies show a relationship between CO<sub>2</sub> emissions, energy consumption, and Gross Domestic Product (GDP) [9]. Since there is a dependency between energy systems and the environment, forecasting carbon dioxide emissions is essential.

There are several ways to predict carbon dioxide emissions with AI models, and the most precise regression models are listed in the text. Each regression model is different and applicable depending on the predicted variables' linearity, complexity, and non-linearity. One of the famous technologies is the

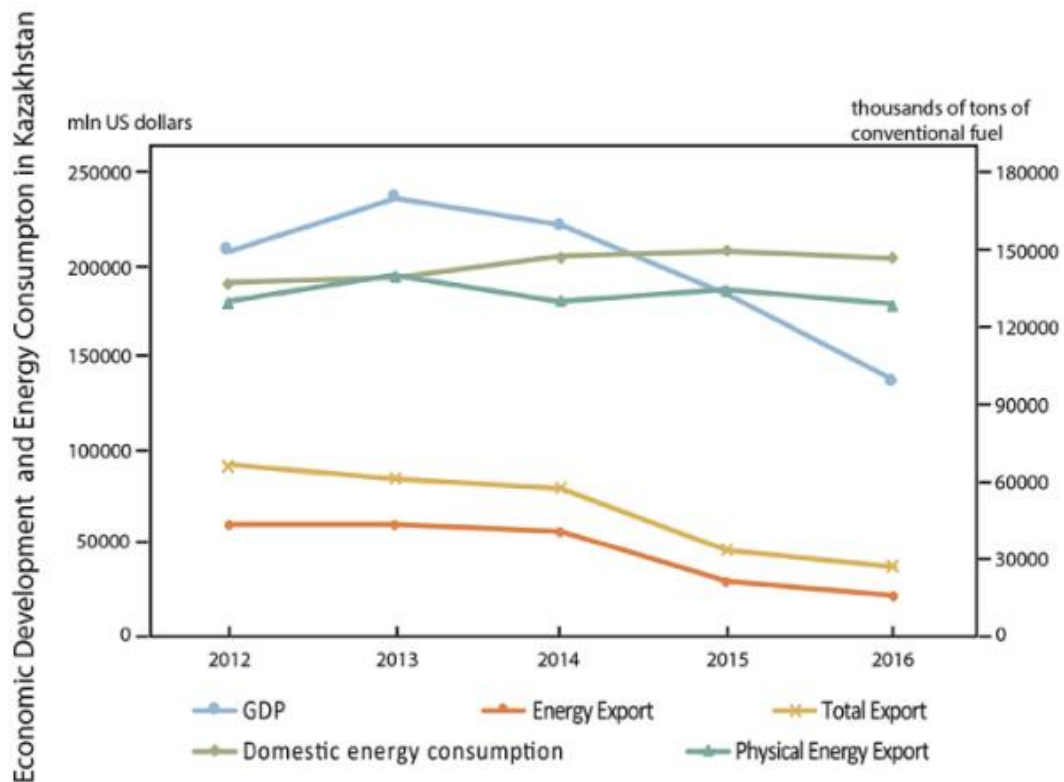


SVM (support vector machine) method. Several researchers used SVM to predict carbon emissions and rates [29]. SVM is a tool for “classification and data forecasting” to predict heat load users for district heating systems [26], [27]. [5] used an adaptive neuro-fuzzy interference system, and a multi-layer perceptron artificial neural network was used to predict CO<sub>2</sub> emissions, which are intelligence approaches. Artificial Neural Network is one more tool for estimation done by Ahmadi et al. (2019) [1]. There are several models for forecasting the CO<sub>2</sub> emissions using the Python scikit library based on Jupyter notebooks, such as Multiple variables that have additional parameters: quantity of cylinders and amount of consumed fuel that refer to the linear function; Non-linear regression, which generates curves that can include either exponential, sigmoidal, or logarithmic functions; Single variable linear regression module which has only a variable called engine size.

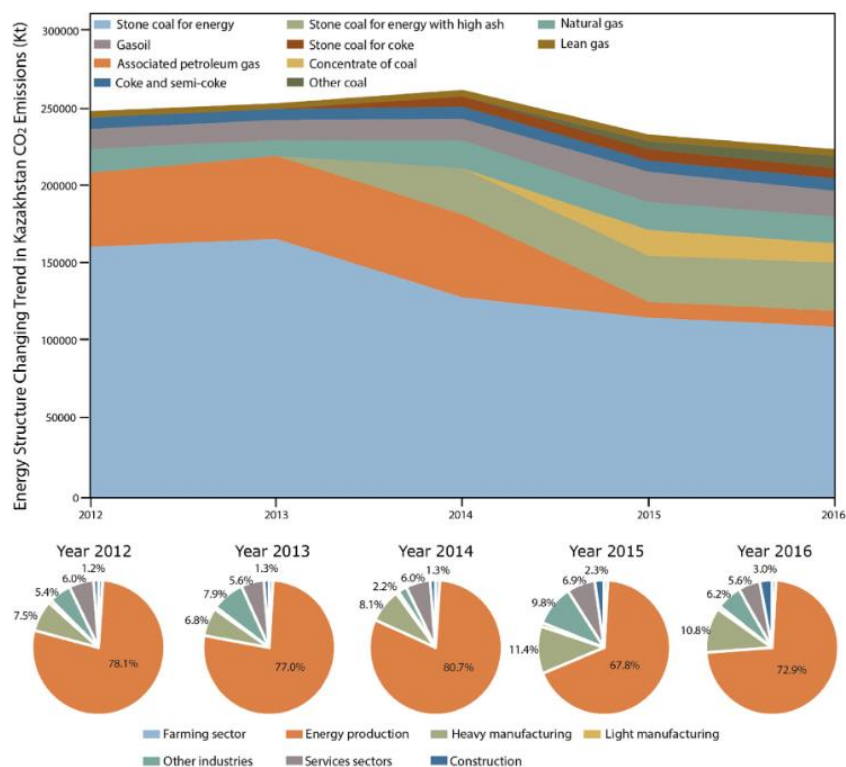
## **2. Literature Review**

The threat of global warming became one of the fundamental problems for humankind in the XXI century. Therefore, the leaders of major global economies, including 192 parties, initiated the ratification process of the Kyoto Protocol in 2005 [16], [20], [21]. This protocol is based on principles on shared responsibility to reduce the emission of greenhouse gases significantly. CO<sub>2</sub> gas is one of the most significant constituents among the greenhouse gases, and it has the most critical harmful impact on the environment, respectively [22], [19], [15]. However, one of the Kyoto Protocol’s limitations did not affect the developing countries such as China, India, which have become the major CO<sub>2</sub> contributors nowadays [17]. Consequently, a new ambitious treaty has been risen, which deals with the obligatory constant reports to monitor the progress in reducing CO<sub>2</sub> emission. This treaty, known as The Paris Agreement, was adopted in 2015 by 196 parties [31]. Thus, Kazakhstan, being of the parties that signed this agreement, is committed to fulfilling the obligations for CO<sub>2</sub> emissions, which includes a 25% reduction in greenhouse gas emissions by the end of 2030 [33]. Kazakhstan is the largest energy-oriented country globally, exporting fossil fuels and minerals products worth 22 billion US dollars for more than 190 countries in the world [12].

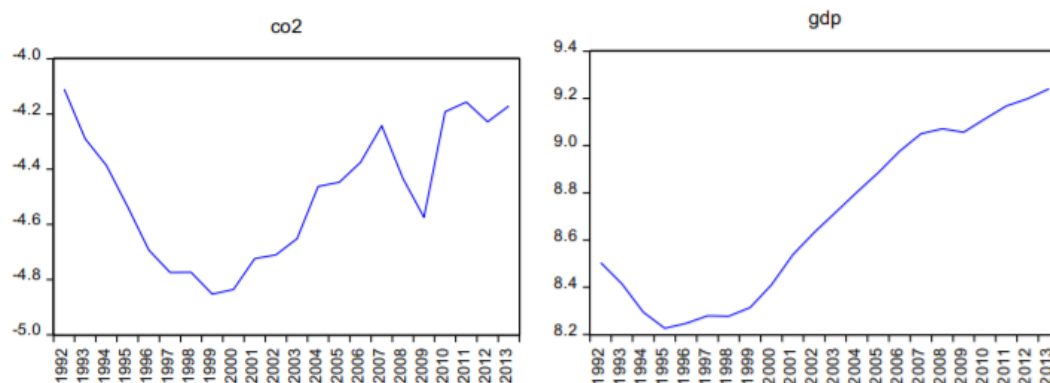
A significant part of the export is the coal, oil, and gas, which are the primary sources responsible for CO<sub>2</sub> emission into the atmosphere [8], [11], [24]. Many institutes and academics have already done substantial research in this direction for Kazakhstan. For instance, Wang and coworkers (2019) provided a comprehensive overview of Kazakhstan’s leading contributors to emissions from 2012 to 2016, which gives a clear picture of the current situation in terms of CO<sub>2</sub> emissions [32]. Based on Figure 1, Kazakhstan’s economy is strongly driven by energy export, and the oil prices significantly affected the economic slowdown in 2015 under these circumstances.



**Figure 1:** Economic development and energy consumption in Kazakhstan from 2012 to 2016 (Wang, 2019)



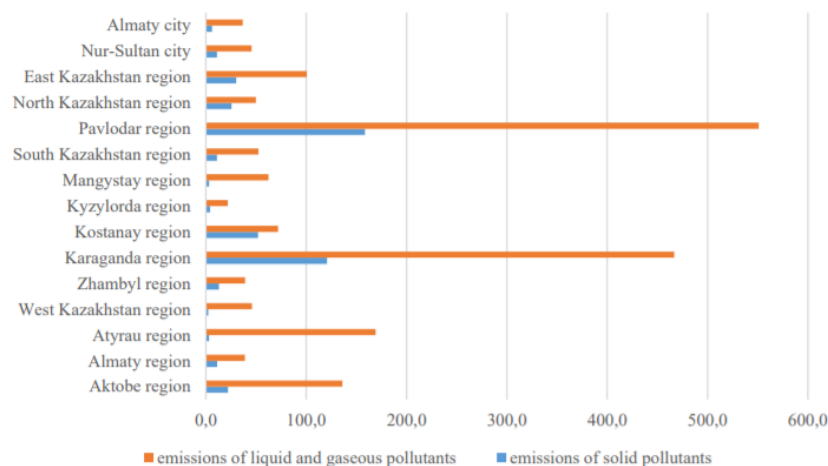
**Figure 2:** Segregation of CO2 emissions by economic sectors in Kazakhstan from 2012 to 2016



**Figure 3:** Time profile for CO2 emissions and GDP in Kazakhstan

Hasanov and coworkers (2019) [14] found another interesting point about the gross domestic product and carbon dioxide emission. Figure 3 illustrates the relationship between the about the gross domestic product and carbon dioxide emission in Kazakhstan. The general tendency for GDP per capita has increased as CO2 per capita also increased from 1992 to 2013. However, some drops in CO2 emission for the period 1997-2000 are due to the weakening of industrial sections, including the shutdowns of many process factors across the country.

Sansyzbayeva and coworkers (2020) [30] conducted a more detailed analysis of the emissions from cities in Kazakhstan. The infographics about the released emissions in 2018 for Kazakhstani cities by region are provided in Figure 4. Pavlodar and Karaganda have higher emissions of pollutants than all combined cities and account for more than half of all emissions in the whole country. Most industrial enterprises, factors a, and plants are concentrated in these regions, which negatively affect the environment due to gaseous, liquid, and solid emissions released into the atmosphere.



**Figure 4:** Emission of pollutants by regions across the country in 2018

This project aims to perform the CO2 emissions prediction for Kazakhstan using ATNN, RNN models and compare the data with prediction data of Deep learning. In the case of an ERP implementation, the project may simply include a current state assessment, selection and implementation. Most organizations are closer to this option, because most companies are not ready to change everything at once.

### 3. Methodology

As long as the project requires precise prediction, residual MSE, mean absolute error, R squared value, etc., can be used to precisely estimate the behavior of CO2 propagation. Normalization, global

averaging are planned to be used as the CO2 dataset is big data. We project to use Carbontracker implemented in Python to track and forecast the power usage and carbon dioxide emission of deep learning-based models. The model is expected to be realized as multithreaded software. It will apply particular features to gather energy consumption and for the carbon emission in actual time for simultaneous effectiveness and not interfere with the main program in the training stage.

Carbontracker allows us to forecast the overall period, power, and carbon track of training a deep CNN model. These forecasts are built on user-due various tested epochs with a default of one. We predict the carbon emission of, for instance, the manufacture of electrical equipment at the time of the forecasted period applying the necessary APIs. The expected CO2 is then used to indicate the carbon emission tracking. The upcoming will be our preparatory study, and we will apply a primitive linear model for forecasting.

We want to use the NVIDIA Quadro P2000 GPU. We are most likely to apply classical machine learning algorithms such as LDA, NN, DSLVQ classifiers to predict the upcoming level of CO2. Other than carbon tracker PyTorch most likely could be our subsequent monitoring deep learning algorithm to improve the algorithm's performance and show the results on top [4].

Hybrid models, including CNN and LSTM, can be another option for us to accelerate the learning process and obtain higher results [3]. LSTM can be further improved, and we may plan to use attention mechanisms and transformer networks if the results are successful when involving LSTM [23], [6]. Applying model selection using those state-of-the-art deep learning approaches may most likely optimize the performance of the prediction and may compete with state-of-the-art.

The dataset is in time series. Then, RNN, ATTN, and another RNN are used to classify the dataset. There are some dense layers and activation layers. Adam optimizer and Keras are used as the deep learning framework and mean square error method. The model was fit with 256 batch sizes, train, and tests. 20% of the dataset was left for validation, use and 100 epochs were run in overall. Predictions from 2018 to 2023 could be made using this algorithm, and the model shows accurate results. Therefore, the model in this algorithm is a prediction one. Also, we put year (like Transformer architecture does) and country encodes to RNN; we hope this can help the model better understand some trends.

Specifically, six values were passed to the model to predict the next one, and steps to realize this project was next:

1. Information about countries was collected;
2. Unique tokens were created for country income groups;
3. Dictionary with unique values was created: token;
4. Application an income group map;
5. Data for Kazakhstan;
6. Showing a missing CO2 level per Year;
7. Getting rid of trash columns;
8. Showing missing CO2 values by country;
9. Dropping countries that do not have enough data;
10. Observation and creation of CO2 array levels;
11. Design of array with country income;
12. Conversion to a NumPy array has three inputs [batch size, timestamps, features] and three parts: CO2 level, Year, country income token;
13. Country income token put in embedding then concentrate with other data.

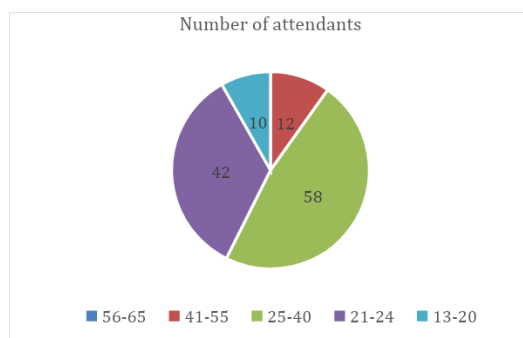
Bidirectional LSTM layers and second bidirectional LSTM layers were used to simulate the results in the prediction analysis. Then, the values for each year were used to predict the next successive year, respectively. As a result, CO2 values in 2019 were used to the following value for CO2 in 2020. Finally, the last valid values and predicted values were concatenated for next year.

## 4. Results and discussion

This section contains two parts; results obtained by public survey and results from simulations of codes that include predictions of CO2 emissions in Kazakhstan.

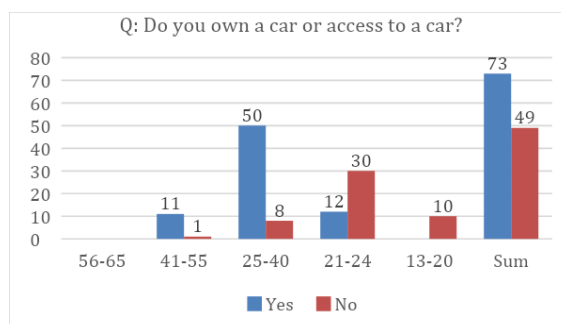
## 4.1. Survey results

This survey contained 27 questions; some are related to personal information and some about the thoughts and predictions associated with CO2 emissions, global warming, and electric cars. In this part, only some of the results are highlighted, which were thought to be necessary. Figure 5 demonstrates how many people attended and their ages categorize them. In total, 122 people attended, and the majority were between the ages of 25 and 40. The next popular category was between 21 and 24. Noteworthy, in this survey, there was no person older than 55 years. The reason may be that fewer local, more senior people speak English, and the survey had to be translated into the local language to reach more of the population in the region.



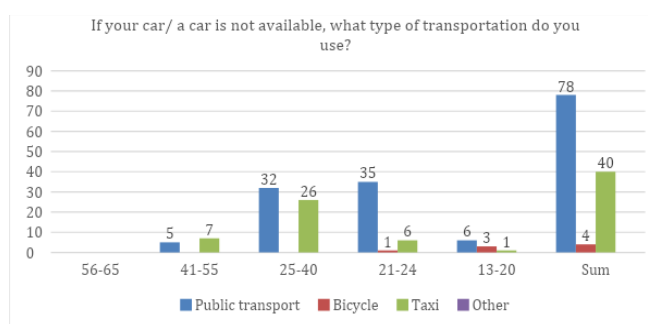
**Figure 5:** Age distribution of attendants

The following substantial question was whether people have their cars. According to the survey, about 60% of attendants have their vehicle or access to a car. Most of the car owners are between the ages of 25 and 40, which follows the logic. More detail can be seen in figure 6.



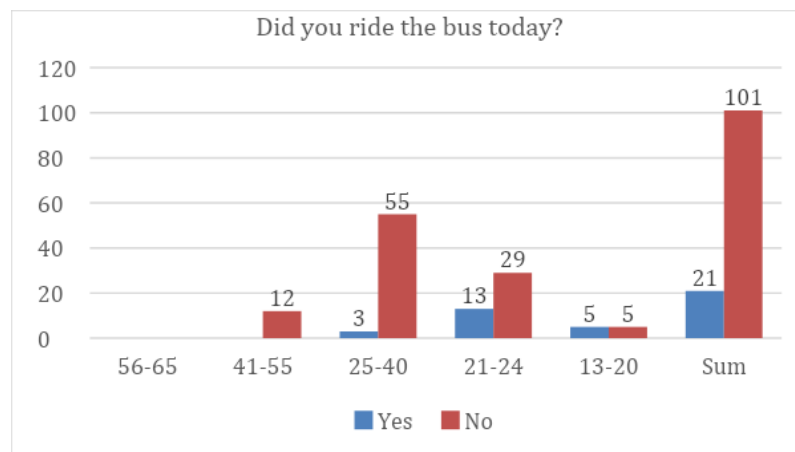
**Figure 6:** Statistics of age distribution who own a car or access to a car

In Figure 7, data about popular transport among attendants is demonstrated. It shows that public transportation is more popular than taxis and bicycles. However, between the ages 25-40, the numbers are very close to each other. The reason for this might be the financial wealth of people in this category.



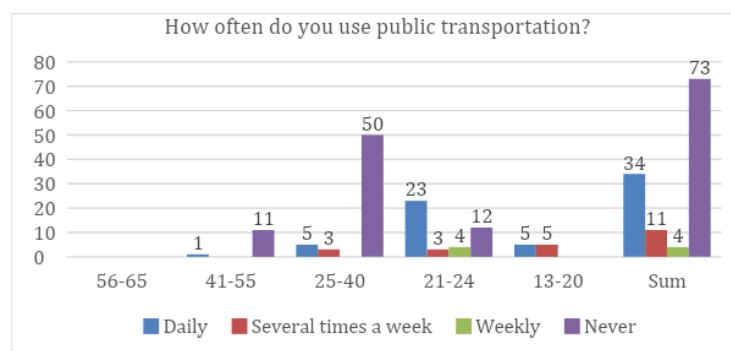
**Figure 7:** Popular transport type

The following substantial question in the questionnaire was related to riding a bus on the current day. From the picture, it could be seen that most of the survey attendants were not using it, and only about 17 percent used it. Such results are related to that most of the attendants have their cars. Consequently, this data does not show the whole picture in our country. That is why the number of attendants should be increased, to see the objective results.



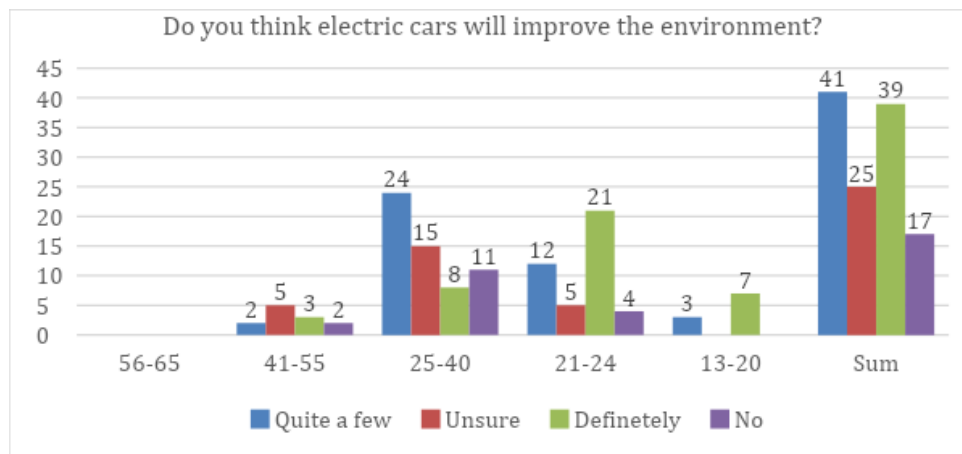
**Figure 8:** Statistics of how many attendants used the bus on a current day

The next question was related to the frequency of public transport use. Again, this data does not show the whole picture in the country. According to this data, most of the respondents answered that they never use public transport.



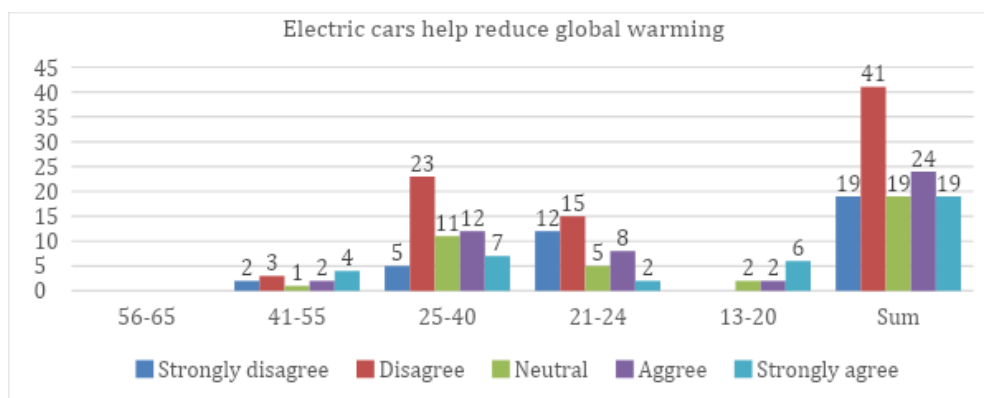
**Figure 9:** Statistics of how frequently attendants use public transport

The next question was related to electric cars if they can improve the environment. Most people think that this technology might improve it, some of them are unsure, and few believe it will not improve.



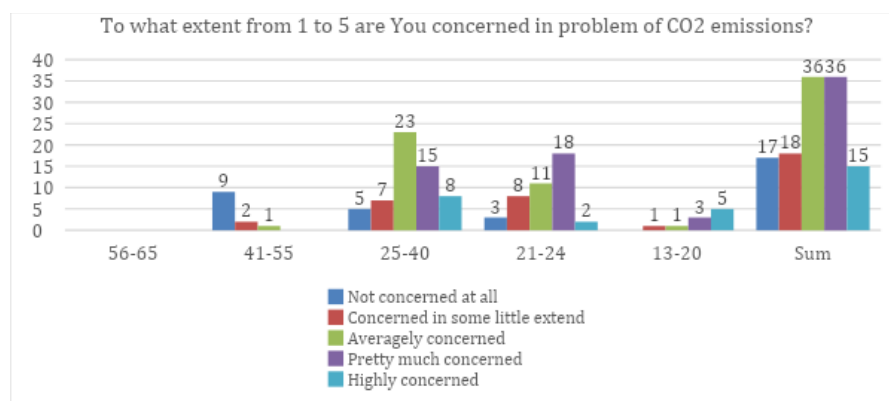
**Figure 10:** Statistics of what attendants think about electric cars effect on the environment

The next question was related to electric cars if they can reduce global warming. Noteworthy, this time, most of the attendants said that they disagreed with this statement. The reason could be that in Kazakhstan, the source of electricity is coal, which is one of the main reasons for global warming [18]. If analyzing according to the ages, all of the attendants between the ages of 13-20 agreed with the statement. In contrast, most of the 25-40 categorized attendants disagree with this statement.



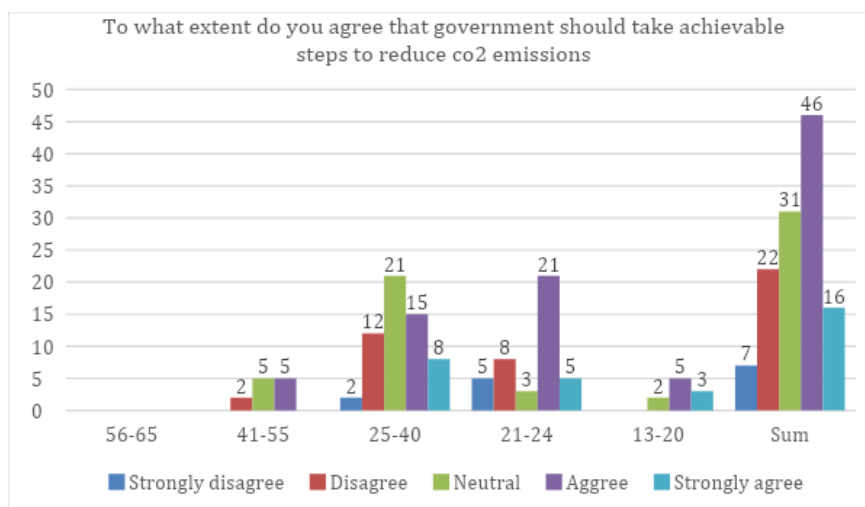
**Figure 11:** Statistics of what attendants think about electric cars effect on global warming

From this diagram, it could be seen that most of the attendants are aware of problems related to CO2 emissions. The interesting is that with the age increase, the number of concerned people is reducing.



**Figure 12:** Statistics of how attendants are concerned in CO2 emission problems

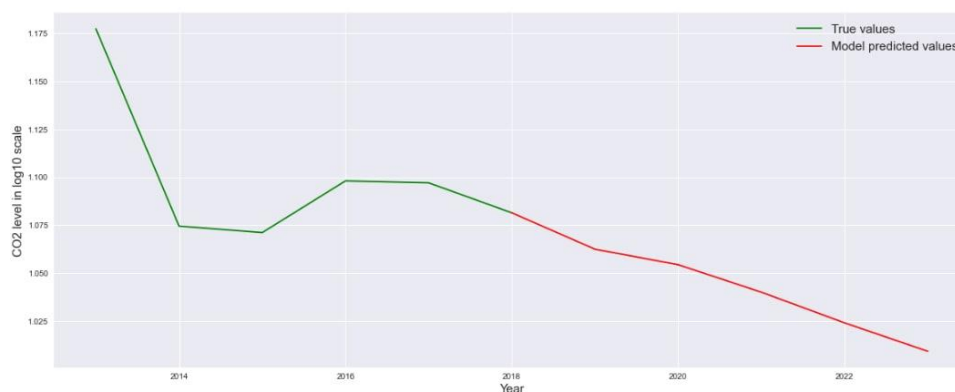
The last exciting question was related to the aid from the government side. Most of the attendants agreed that the government should take achievable steps to reduce CO2 emissions.



**Figure 13:** Statistics of attendants' thought about aid from the government side to reduce CO2 emissions

## 4.2. Simulation results on prediction of CO2 emissions in Kazakhstan

The simulation results took place until 2023. As can be seen from figure 13, the green line describes data taken from the archives, and the red line represents the predicted values. According to the model we used, the whole picture will follow the decreasing trend until 2023. As discussed before, data after 2018 was not available, and prediction values were done only after 2019. The reason for the decreasing numbers could be an increase in the efficiency of the technologies despite the rise in the population and transportation number. The five-year model predicted that the decrease would be about 6%, which is relatively high considering the people of Kazakhstan is increasing by 1.3-1.5% annually.



**Figure 14:** Predicted results of CO2 in Kazakhstan between 2019 and 2023

## 5. Conclusion

This paper was dedicated to forecasting CO2 emissions with the help of AI models. Increasing carbon dioxide concentrations in the atmosphere were noticed to have negative impact not only to the environment but also on social, political, and economic sectors. Since all of the sectors are interconnected, prediction of CO2 emissions is critical in making future decisions.

The results section contained two parts: first, a survey of the population in Kazakhstan, where 122 people attended, and second, a DL model to predict the CO2 emissions in Kazakhstan until 2023. The survey results showed that young-aged attendants look at the future with more optimism, and the



prediction model indicated that CO<sub>2</sub> emissions would decrease. Despite optimistic results, it should be noted that forecasting tools are not perfect, they only serve as a help in seeing future trends and making decisions based on them.

## 6. References

- [1] M. Ahmadi, H. Jashnani, K. Chau, R. Kumar, M. Rosen, "Carbon dioxide emissions forecast in five Middle Eastern nations using artificial neural networks," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, pp. 1-13, 2019. 10.1080/15567036.2019.1679914.
- [2] Ahmadi, M. H., Jokar, M. A., Ming, T., Feidt, M., Pourfayaz, F., & Astarai, F. R. (2018). Multi-objective performance optimization of irreversible molten carbonate fuel cell Braysson heat engine and thermodynamic analysis is an objective ecological technique. 707-722 in *Energy*.
- [3] L. Amarpuri, N. Yadav, G. Kumar, and S. Agrawal (2019, August). A case study in the Indian context for predicting CO<sub>2</sub> emissions using a deep learning hybrid technique. The Twelfth International Conference on Contemporary Computing (IC3) will be held in 2019. (pp. 1-6). IEEE.
- [4] L. F. W. Anthony, B. Kanding, and R. Selvan (2020). Carbontracker is a tool for tracking and estimating the carbon impact of deep learning model training. preprint arXiv:2007.03051, arXiv:2007.03051, arXiv:2007.03051, arXiv:2007.0305B.
- [5] Baghban A, Ahmadi M A, and Hashemi Shahraki B 2015 Prediction carbon dioxide solubility in the presence of various ionic liquids using computational intelligence approaches *J. Supercrit. Fluids* 98 50–64.
- [6] Chithrananda, S., Grand, G., & Ramsundar, B. (2020). Chamber: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.
- [7] De Paz J F, Pérez B, González A, Corchado E and Corchado J M 2010 A Support Vector Regression Approach to Predict Carbon Dioxide Exchange Distributed Computing and Artificial Intelligence (Springer) pp 157-64.
- [8] Ekwurzel, B., Boneham, J., Dalton, M. W., Heede, R., Mera, R. J., Allen, M. R., & Frumhoff P. C. (2017). The rise in global atmospheric CO<sub>2</sub>, surface temperature, and sea level from emissions is traced to major carbon producers. *Climatic Change*, 144(4), 579-590.
- [9] Fan, W., & Hao, Y. (2020). Empirical research on renewable energy consumption economic growth, and foreign direct investment in China. *Renewable energy*, 146, 598-609.
- [10] Friedlingstein, P., Jones, M. W., O'sullivan, M, Andrew, R. M., Hauck, J., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quere, C. and Bakker, D.C., 2019. Global carbon budget 2019. *Earth System Science Data*, 11(4), pp 1783-1838.
- [11] Frumhoff, P. C., Heede, R., & Oreskes, N. (2015). The climate responsibilities of industrial carbon producers. *Climatic Change*, 132(2), 157-171.
- [12] Gacek, L. (2017). The External dimension of china's energy policy towards Kazakhstan: perspective of cooperation within the «One Belt, One Road» initiative. *Roczniki Humanistyczne*, 65(9), 87-108.
- [13] Gorus, M. S., & Aydin, M. (2019). The relationship between energy consumption, economic growth, and CO<sub>2</sub> emission in MENA countries: Causality analysis in the frequency domain. *Energy*, 168, 815-822.
- [14] Hasanov, F. J., Mikayilov, J. I., Mukhtarov, S., & Suleymanov, E. (2019) Do CO<sub>2</sub> emissions-economic growth relationship reveal EKC in developing countries? Evidence from Kazakhstan. *Environmental Science and Pollution Research*, 26(29), 30229-30241.
- [15] Hitchon, B., Gunter, W.D., Gentzis, T., & Bailey, R. T. (1999). Sedimentary basins and greenhouse gases: a serendipitous association. *Energy Conversion and Management*, 40(8), 825-843.
- [16] Houser, T. (2010). Copenhagen, the accord, and the way forward (No. PB10-5). Washington, DC: Peterson Institute for International Economics.
- [17] Hubacek, K., Guan, D., & Barua, A. (2007). We are changing lifestyles and consumption patterns in developing countries: A scenario analysis for China and India. *Futures*, 39(9), 1084-1096.
- [18] International Trade Administration. Renewable Energy-Kazakhstan, viewed Nov. 17,2021, <https://www.trade.gov/energy-resource-guide-kazakhstan-renewable-energy>.

[19] Jiang, Z., Xiao, T., Kuznetsov, V. A., & Edwards, P. A. (2010). Turning carbon dioxide into fuel. *Philosophical Transactions of the Royal Society A; Mathematical, Physical and Engineering Sciences*, 368(1923), 3343-3364.

[20] Kim, Y., Tanaka, K., & Matsuoka, S. (2020). Environmental and economic effects of the Kyoto Protocol. *PloS one*, 15(7), e0236299.

[21] Klimenko, V. V., Klimenko, A. V., & Tereshin, A. G. (2019). From Rio to Paris via Kyoto: How the efforts to protect the global climate affect the world energy development. *Thermal Engineering*, 66(11), 769-778.

[22] Lacis, A. A., Schmidt, G. A., Rind, D., & Ruedy, R. A. (2010). Atmospheric CO<sub>2</sub>: Principal control knob governing Earth's temperature. *Science*, 330(6002), 356-359.

[23] Lin, X., Zhu, X., Feng, M., Han, Y., & Geng, Z. (2021). Economic and carbon emissions optimization of different countries or areas globally uses an improved attention mechanism based on the short-term memory neural network-the science of *The Total Environment*, 148444.

[24] Malla, S. (2009). CO<sub>2</sub> emissions from electricity generation in seven Asia-Pacific and North American countries: a decomposition analysis. *Energy Policy*, 37(1), 1-9.

[25] Mardani, A., Liao, H., Nilashi, M., Alrasheedi, M., & Cavallaro, F. (2020). A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. *Journal of Cleaner Production*, 275, 122924.

[26] Pinto T., Sousa T., I. Praca, Vale Z., and Morais H., «Support Vector Machines for decision support in electricity markets' strategic bidding», *Neurocomputing*, vol. 172, pp. 438-445, 2016. Available: 10.1016/j.neucom.2015.03.102.

[27] Protic M, Shamshirband S, Petkovic D, Abbasi A, Mat Kiah ML, Ulnar J A, Zivkovic L and Rao M 2015 Forecasting of consumers heat load in district heating systems using the support vector machine with a discrete wavelet transform algorithm *Energ* 87 343-51.

[28] Rahman, Z. U., Khattak, S. I., Ahmad, M., & Khan, A. (2020). A disaggregated-level analysis of the relationship among energy production, energy consumption, and economic growth: Evidence from China. *Energ*, 194, 116836.

[29] Saleh C., Dzakiyullah N., and Nugroho J., «Carbon dioxide emission prediction using support vector machine,» *IOP Conference Series: Materials Science and Engineering*, vol. 114, p.012148, 2016. Available: 10.1088/1757-899x/114/1/012148 [Accessed 30 September 2021].

[30] Sansyzybaeva, G., Temerbulatova, Z., Zhidibekkyzy, A., & Ashibekova, L. (2020). Evaluating the transition to a green economy in Kazakhstan: A synthetic control approach. *Journal of International Studies*, 13(1), 324-341. DOI:10.14254/2071-8330.2020/13-1/21.

[31] Savaresi, A. (2016) The Paris Agreement: a new beginning? *Journal of Energy & Natural Resources Law*, 34(1), 16-26.

[32] Wang, X., Zheng, H., Wang, Z., Shan, Y., Meng, J., Liang, X., Feng K., & Guan, D. (2019) Kazakhstan's CO<sub>2</sub> emissions in the post-Kyoto Protocol era: Production-and consumption-based analysis. *Journal of environmental management*, 249, 109393.

[33] Zhekey, A. (2016). Coal, power, and Kyoto protocol: regulating greenhouse gas emissions in Poland and Kazakhstan.

# Voice Control Technology for Drones Based on Intelligent Command Recognition

Khuralay Moldamurat<sup>1</sup>, Niyaz A. Belgibekov<sup>2</sup>, Abzal Kyzyrkanov<sup>3</sup>, Saule Brimzhanova<sup>4</sup>,  
Olga Bizhanova<sup>5</sup>, and Aidos M. Kalkabek<sup>6</sup>

<sup>1</sup> *L.N. Gumilyov Eurasian National University, Astana, Kazakhstan*

<sup>2</sup> *NIOCR Department of military-technicians Project R&D centers Kazakhstan engineering, Astana, Kazakhstan*

<sup>3</sup> *Astana IT University, Astana, Kazakhstan*

<sup>4</sup> *Kostanay Academy of the MIA of the RK named after Sh. Kabyldaev, Kostanay, Kazakhstan*

<sup>5</sup> *NPLC A. Baitursynov KRU, Kostanay, Kazakhstan*

<sup>6</sup> *LLC «R&D CENTER «Kazakhstan Engineering», Astana, Kazakhstan*

## Abstract

This article is about teaching remote sonic control of an unmanned aerial vehicle (hereafter UAV). The UAV in sound control uses Sentence words. Sound order words are written in the Latin alphabet. The implementation of sound (word) control of the UAV is done through a microcontroller system. There are also sensors that receive sound and implementation devices. The remote implementation of the UAV is done through a satellite control system and is available for control through an Internet system where an Internet network is available. C and C++ programming languages are written in a computer program environment. Training methods for sonic control of the UAV and protection against unauthorized persons in sonic control are implemented. A Foundation of order words most commonly used in UAV control was created. It was compiled from a set of words most commonly used in emergency command and control of UAVs. The relevance of controlling objects by means of sound command was given.

## Keywords

Latin alphabet, Kazakh language, microcontroller system, microphone, command word stock, sensors, unmanned aerial vehicle (UAV), sound control, signaling, intelligent technology, TTL logic modeling information, algorithms in SCL network, sound control

## Abbreviated words:

UAV, ECM, SCL, BLYNK, OLED

## 1. Introduction

The transition of the Kazakh language from Cyrillic to Latin gives great opportunities for the development and management of global information technology and modern technology. Microprocessor-based technology control systems (on-board computers, microcontroller systems and processor systems, BLIS, crystal devices) include [1]. All control commands (machine code) in interfaces and processors on these devices consist of Latin letters and binary codes 0 and 1. Therefore, the role of the Latin alphabet in the technological control of modern devices is great. It is technologically expedient to replace the Kazakh words with Latin words. As a moving object, a simulation of the sound control system of an unmanned aerial vehicle (UAV) is given. The controlled sound is the words in the Kazakh language and the sound control system from a distance by creating its interface.

An unmanned aerial vehicle is one of the most widely used devices in today's world. UAV is a reconnaissance of roadblocks in the area, on the border, exploration of complex objects, reconnaissance of nuclear sites and explosion sites, as a transport for transportation of medicines during the world pandemic corona virus, control of fires, water, earthquakes, and criminal mutinies during emergencies. It is also widely used in amateur artists and the film industry-live-action, film-making, glossy radar shooting, and so on.

When creating an automated UAV sound control system, we base it on a microcontroller system. We use words in the Kazakh language with the Latin alphabet as the basic words of the «order» in the sound control. A hardware-software system for identifying the TECHNOLOGY of Internet things when

using the sound control of the UAV was created [2].

Control by sound command has many advantages for the human child. An indispensable feature for people with disabilities without arms and legs in life. Controlling a seated wheelchair with audible command. Pilot who is inexperienced must control the object of flight by sound command, etc.

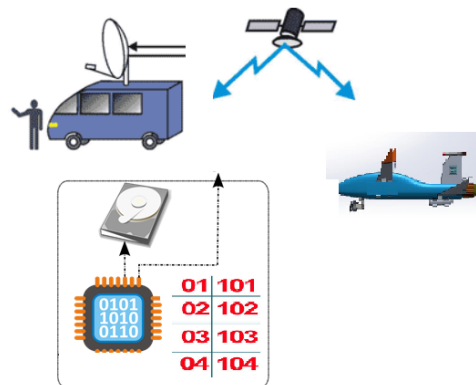
## 2. Methods and technologies

Sound control of the UAV facilitates the work in the field of bodies in emergency situations - police, firefighting, ambulance. It saves time of special team units. In addition, the center monitors the level of the object that is in an emergency state, transmits information to the control center. The system of automation of UAV sound control is implemented using Blynk, installed through special cell phones and central control points (Android or IOS and ECM) [3].

Blynk is a program that is used to remotely control electronic devices. Most often the Internet is used with the technology of things (IoT). In addition, for simple tasks such as receiving notifications and controlling the device IOS / Android apps are used without code. App Store app and used on Google Play. Blynk can create its own prototypes with the app, allowing the app to add a logo, colors and icons [4].

The main development areas of Internet of Things technology are widely used in the fields of public transportation, medicine and the military, the film industry, and education. The creation of sound control of UAVs based on global technology is relevant to the state of Kazakhstan. In areas provided with the Internet and radio communication, it is possible to receive audio information through the UAV in a matter of seconds. And also by issuing a special «order», we determine the level of the object with the help of sensors [5].

The software and hardware system of the UAV sends a signal to the Rescue Service within 1-2 seconds, transmitting the alarm to the special rescue authorities 01, 02, 03, 04.



**Figure 1:** Classic digital control of the UAV

Figure 1 control of the UAV by a special program is carried out by a pilot, for which it is trained. It is implemented with the help of special programmers. The alarm was transmitted by digital signals through the program [6].

The UAV is created in a special microcontroller sound control system. Combined devices are used for the implementation. Devices connected to the microcontroller-microphone, recognition sensors, Arduino devices, modems, video surveillance devices, antennas, servo drives, power supply devices and more. The UAV blocks the signal over the Internet and radio communication. The main board of the system is an Arduino Mega 2560, the sound recognition module V3 is a GSM module SIM800L, which is connected to the Internet, OLED display is connected to the message display SH1106 and Micro Servo SG90 micro. The creation of the sound initialization program is written in Arduino IDE in a high-level C programming language. Arduino IDE is an Arduino-special development environment used in C and C++ to write and download programs to compatible boards [7].



**Figure2:** MT Technology Co. Module V3 integrated board

The sound control is implemented on this board. MT Technology Co., Ltd. V3 speech recognition module is compact and easy to operate. There is a board for recognizing words (sounds / voices) in speech. This product has a word recognition module linked to the speaker. It can support up to 80 command words to sound command. Can master 7 voices / sounds at the same time. Any sound can be taught as a «command». This board is based on an intelligent neural system. The user must first fully train the module before determining the voice commands [8].

The board has two ways of control: the serial port (full function), the universal input is controlled via a pin (part function).

The common target output connections on the board can generate multiple signals when the corresponding voice commands are detected. The device can execute 7 audio commands at a time [9].

**Parameter:**

Pressure: 4.4-5.5 V

Current: <40mA

Board integrated with V3 module

Digital Port: UART port and GPIO 5V TTL level

Port simulation: Suitable for 3.5mm microphone jack and microphone contact interface

Size: 31mm \* 50mm

Accurate recognition: 99% (in an ideal environment)

Volume detects 80 sound orders, 1500ms of each sound (one and two words)

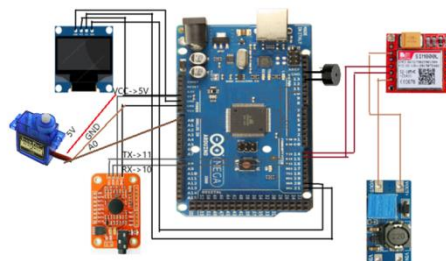
Let's prepare a list of the most necessary words of the orders in the Kazakh language and put it into the library of the Arduino device. It is written to the memory of a special microcontroller.

On the Arduino device a library collection is created, presented as a database. The library stock contains a list of command words in Latin in Kazakh. It is easy to control: The UART / GPIO Pinout is a general-purpose user control application.

**Optional device:**

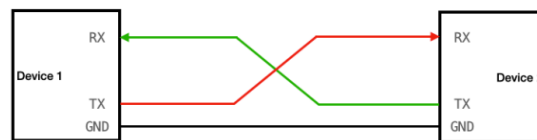
Types of sensors temperature sensor, used distance sensor, rotation sensors, etc. and connect 1\* sound recognition module, 1\* microphone.

Integrate all devices into the microcontroller system. No need to create sound control software system. Because the above MT Technology Co. V3 module integrates with the board that processes sounds. Another advantage is to work with the provided Arduino library.



**Figure 3:** Microcontroller sound control system for UAVs

In Figure 3 you can see the sets of devices that Arduino Mega has registered in the microcontroller. The Arduino Mega microcontroller inputs an analog signal via a pin, the incoming signal forms a set of binary codes 0 and 1 in digital processing [10]. There are two main data interfaces-UART and I2C. UART (Universal Asynchronous Receiver-Transmitter) is an asynchronous TX and RX pin transceiver, which means that it is connected via modules and converted to (TX-RX, RX-TX) (Figure 4).



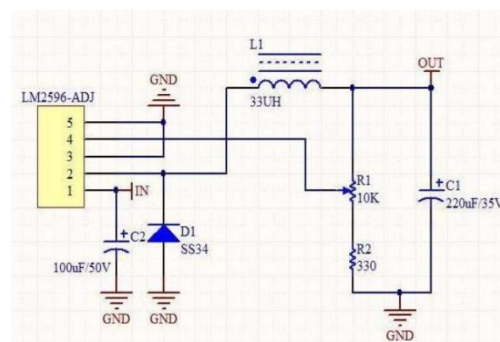
**Figure 4:** Adding UART

The network is kept high +5V, Arduino Mega special TTL logic is executed before the information is transmitted. After a certain time interval, depending on the set speed, a byte transmission starts in the form of a series of zeros and/or ones according to the set time intervals. After the eighth bit is a stop signal goes as a high level and the situation repeats until the necessary bytes will not be transmitted. The UART protocol is different, the main characteristics are: speed, number of bits, parity and stop signal length. I<sup>2</sup>C is a serial bus that operates to transfer 8-bit data at 100 to 400 kbps at low transfer rates. The data exchange process is done via two buses: the SDA data network and the SCL clock network. The chip with the I<sup>2</sup>C connection protocol has a hardware interference suppression algorithm, which ensures data safety even in case of strong interference. This chip makes it possible to contact each other even with different supply voltages (e.g. OLED display SH1106 with an Arduino Mega) [11].

Each of the devices connected to the bus has its own unique address. The devices on it can work as a receiver or transmitter.

When transmitting data they can be master or slave. A master is the device that initiates data transfers and generates clock signals on the SCL network. At any given time on the I2C bus there can only be one device that is the master and generates a signal to the SCL line. In addition, the master can be a master receiver or a master transmitter.

Since it is not difficult to recognize sirens, the Arduino board does a great job with the Internet of Things technology thanks to its easy connection to many applications and devices. The Arduino has more than 9 versions, and they are all used for different purposes: robotics, automation, learning in a connected environment, etc. The power output ports are welded to Ground (GND) and Power Pin (VCC) (Figure 2).



**Figure 5:** LM2596 chip

**Characteristics of the stabilizer:**

- Input voltage-4 V-35 V;
- Output voltage-1,23 V-30 upcoming output current-3 A;
- Switching frequency-150 kHz [14].

The Voice Recognition Ver module to the left of the board is connected to the Arduino Mega.3. The

third version is the latest version of the module, which can store up to 80 sounds and recognize up to 7 sounds simultaneously, the second version can only store 15 and recognize up to 5 sounds. This module connects to the Arduino through the TX, RX, VCC and GND ports and pins A8-A15. VCC (Voltage Common Collector) is the voltage output pin, 5V and GND is ground, a zero potential circuit node. Sends data to the Arduino using the UART interface. [0].

A GSM module with a SIM card to send SMS messages, calls and internet and an antenna to provide radio communication is needed to have internet. There are 3 best GSM modules for Arduino: NEOWAY M590, SIM800L / SIM900, Arduino MR GSM 1400. So, the Sim800l module is connected to the Arduino Mega [12]. The module has a UART interface, TX and RX pins, a 4V power supply, Quad-band lines. The module itself is welded to the board along with SMD resistors (special mounting resistors on the PCB). Their power rating is 0.05 W, operating voltage of 15V and the maximum allowable voltage of 50 V. Since the GSM module consumes up to 2A at 4V peak, an optional LM2596 power module [0] is used.

The display module on this board is an OLED SH1106 display. It is an electroluminescent display consisting of LEDs that emit light. The display gets text, images, video and more. The advantage of the display is that it has a higher contrast ratio and can respond faster than an LED display [0].

The same pins VDD, GND are 5V and GND, while the SCK and SDA pins are connected to pins 10 and 11 respectively for data transfer [13]. The Micro Servo SG90 is used to simulate the opening of a barrier in a servo. The main difference between its predecessor and the Micro Servo MG90 is that the gearbox is made of nylon, so the weight is significantly reduced and the operating voltage is low - only 3.5 to 5V.

80 different kazakh words «orderer» in the alternation of signals uses piezoemits. Adding the pins of all modules on the board below (Figure 5). The main module on the board will be the Arduino Mega 2560, to which the modules and sensors and antenna will be connected. The Arduino Mega 2560 is an interface development platform based on the ATmega2560 microcontroller. A large number of I/O pins are placed in the smart recognition sensor systems.

**Features:**

- 54 digital inputs/outputs
- 16 analog inputs
- 4 UART ports (hardware serial ports)
- 16 MHz processor clock frequency
- Logic level voltage: 5B
- Board input voltage: 7-12 B
- I/O port current limit: 40mA.

The module itself, connected with internal pull-out resistors, all kinds of which are 20-50 kOm, is connected via a USB cable, has a power connector, an ICSP connector and a reset pin. It is connected to the computer via USB (Figure 5) [14].

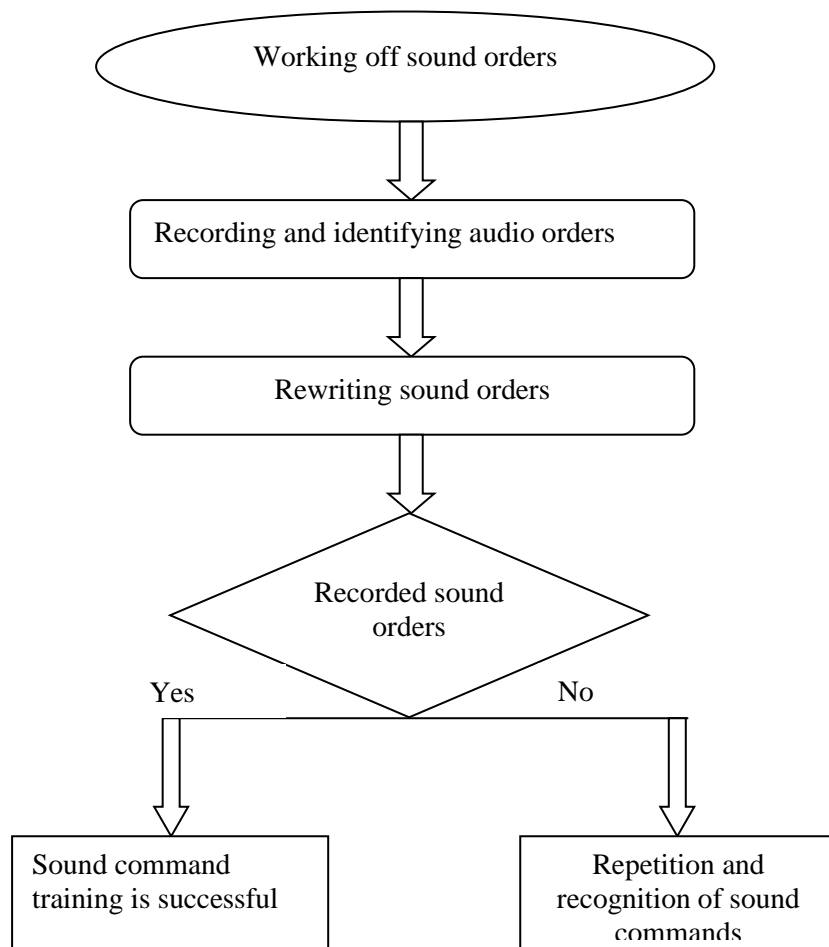
**Arduino has three types of memory:**

- Flash memory-256 KB, used to store programs loaded into the microcontroller.
- SRAM-operating memory, where variables are stored temporarily. When the power is turned off, all data in the memory is deleted, 8KB.
- EEPROM-4 Kbytes of permanent memory stores data that will not be deleted when power is turned off.

The ATmega2560, based on the RISC architecture, has 256KB of flash (of which 8KB is used for the loader), 8KB of SRAM and 4KB of EEPROM.

The power supply voltage of the Arduino Mega 2560 when connected via USB is 5V. (Figure 3)

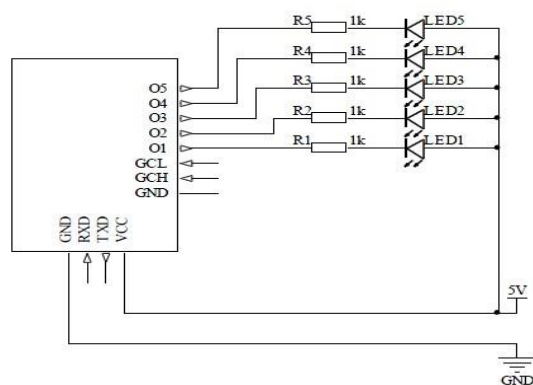
The module is connected to the Voice Recognition Module V3 board because the simplicity of the siren is easily recognized by the voice module. A specific Voice Recognition V algorithm is used to train the module 3 (block diagram 1):



**Figure 6:** Block diagram in learning audio commands modular learning algorithm Voice Recognition V3

In the second version, you can train 15 sounds in parallel, divided into groups, and load the neural recognition system. And in the third version, the voice recognition module supports up to 80 voice commands and implements a maximum of 7 voice commands simultaneously. Any sound can be taught to a command as a «command».

Before we allow any voice command to be recognized, we must first train the module. The module can recognize sound with up to 99% accuracy in an ideal environment (Figure 7).



**Figure 7:** Voice Recognition Module Schematic

This board has 2 ways of control: serial port (full function), common input pins (part function). The



common output pins on the board can create several types of waves when the corresponding voice commands are recognized.



Figure 8: Voice Recognition Module Schematic

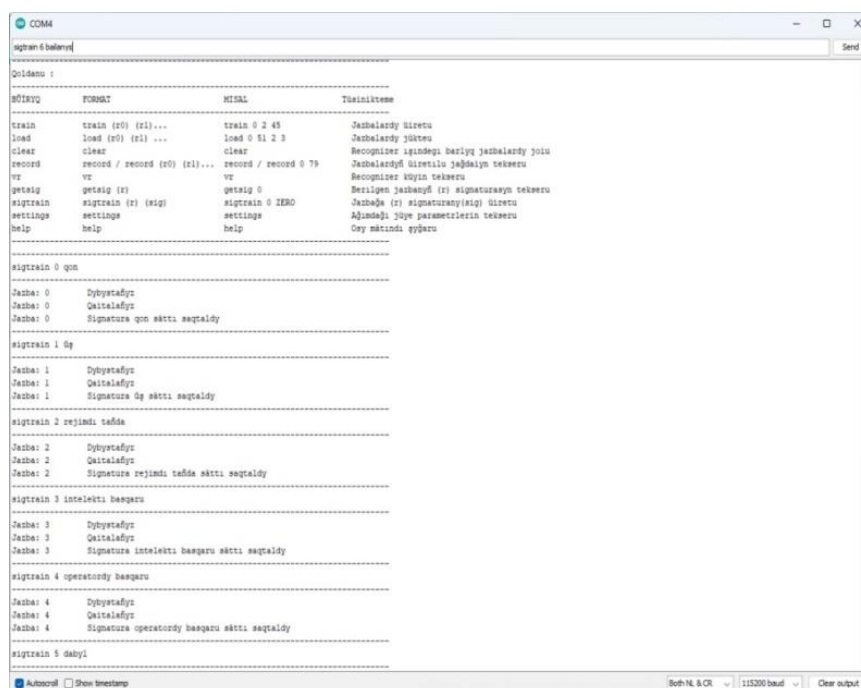


Figure 9: 5-7 operators in the program setting and recording in the recognition module by the pronunciation of Kazakh words of command in the Latin alphabet.

The recording of the sound word stock into the Arduino Mega microcontroller memory is done via the AVR Studio4 software environment.

Another advantage of having two types of module memory (memory and recognition) is that in one module it is possible to create a voice control for several people (up to 5-7 people).

Table 1

List of words included in the knowledge base of the module with intelligent sound recognition

«ush» - tip,	«qon»-guest,
«toqta» - stop,	«basta» - basta,
«zhogary» - supreme,	«túsir» - descent,
«kóter» - up,	«rejimdi tandaý» - mode selection,
«tómen tús» - down,	«jógary ush» - aerobatics,
«obektini izde» - search for an object	«obektini sýretke túsir» - take a picture of the object,
«time» - time,	«araqashyqtyq» - distance,

«buryl oǵǵa»-turn to the right,	«buryl solǵa»-turn to the left,
«buryl artqa»-turn backward,	«bailanys bar»-there is a connection,
«bailanys joq» - no connection,	«jauap»-answer,
«qaitala» - repeat,	«operator» - operator,
«órt» - fire,	«sý» - water,
«basqyn» - occupant,	«tóbeles» - fight,
«tótenshe jaǵdaı» - emergency,	«180 gradýs buryl» -180 degrees turn,
«tura ush» - get up and fly,	«dabyıl ber» - give the signal,
«kómek» - help,	«bailanys» - contacts,
«GPS núkte» - GPS point,	«biiktik orny» - altitude position,
«soltústik polús» - North Pole,	«Ontustik polus» - South pole,
«batus polus» - west pole,	«shygys polús» - east pole,
«alǵa» - forward,	«artqa» - backwards,
«ońǵa» - to the right,	«solǵa» to the left,
«tómen» - down,	«joǵary» - higher,
«túsir» - downhill,	«kóter» - ascent,
«tómen tús» - down,	«joǵary ush» - the highest,
«ash» - open,	«jap» - close,
«óshir» - off,	«Kir» - mud,
«shyq» - dew,	«qabyldandy» - accepted,
«qabyldanbady» - not accepted,	« signal joq» - no signal,
«dabyıl órt 01» - 01 emergency,	«dabyıl polisia 02» - 02 police service,
«dabyıl dáriger» - doctor on call,	« jedel járdem 03» - 03 ambulance.
«gaz jarlisi 04» - 04 fire department,	

**Table 2**

A list of the victim's frequently uttered rescue sound orders is recorded in the knowledge base of the module with intelligent sound recognition

Open authorized sound word order
« qútqar» - save,
«Polisia kerek» - need Police,
«Jol apaty» - Road accident,
« Jedel járdem kerek» - need an ambulance,
«kömektes» - help,
«su basty» - flooded,
«gaz jaryldy» - gas explosion,
«qauıptı» - dangerous,
«ört shyqty» - fire.

All of the above words are controlled by sound commands from the UAV. When controlling the UAV via dual orders, the voices of the operators involved in the control are entered into the knowledge base on the microcontrollers and trained to intelligently recognize the voices of each operator.

The order words associated with the special flight function of the UAV are listed in Table 1. The list of sound order words in Table 1 is entered into the microcontroller knowledge base. The sound order words in Table 1 are specifically trained using the words in the UAV launch sample.

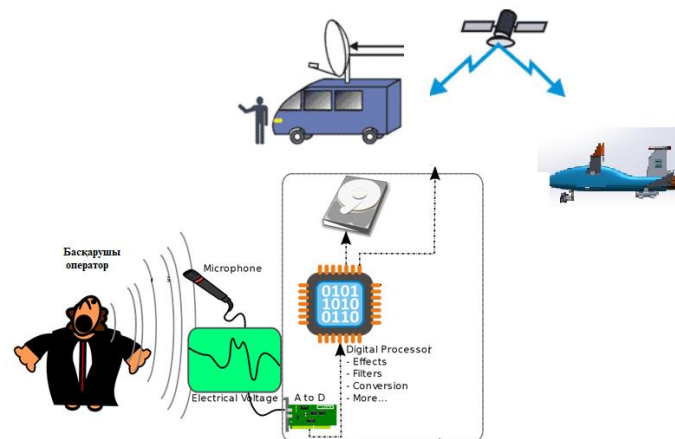
This is due to the fact that the UAV fully executes sound orders when supervised from above and remotely controlled when rescuing locations where a special situation has occurred. During reconnaissance, the UAV transmits the location of the incident to a central operator via video camera, the operator at the center determines the location of the incident and provides directional assistance. In addition, through the UAV's sensitive sensors, the center transmits to the operator the parameters of the

fire at the scene, how dangerous the gas is. The operator assigns the rescue authorities - fire department 01, police 02, ambulance 03 and gas 04 to rescue work.

Which body will come as first aid and inform the operator as an alarm.

For example, «dabyl órt 01» - fire alarm, «dabyl polisia 02» - police alarm, «jedel járdem 03» - ambulance, «gaz jarlisi 04» - gas explosion.

The voices of each operator record voices in different cells of the storage of the implementing module, saying aloud the sound «order» from Table 1. Recorded in the knowledge base of the module from the 1st to the 80th word by the sound of the operator involved in the control of the UAV. For example, «Ushu- take-off», «Honu-landing», «Tohtau-stop» and all operators involved in the process of using command words are entered and loaded into the recognition device. Operators train in the word recognition module «buryl onga»-turn to the right, «buryl solga»-turn to the left, «bastafhy orun»-the initial position, etc. by repeating the Command Words. In addition, 80 words in the Kazakh language, written in all Latin alphabets, are entered according to the cell numbers of the voice command, and it is specified who pronounces them. This method is designed to ensure that UAVs cannot be used by unauthorized people in controlling the sound. A dedicated UAV is provided with voice control security.



**Figure 10:** Control of the UAV with sound alarms

Figure 2 alarms rescue workers in critical situations using the sound signal of the UAV operator or a normal person. In this case, hearing the cries of bystanders encountered in critical situations for rescue, help, danger, he signals the alarm to the central operator. Table 1 lists the words frequently uttered by human victims where the knowledge base is accessed. The operator or a simple person who sounded the alarm does not need to know the program. If there is a large person or small child in a difficult situation, the rescue operation is carried out with the help of sound. The UAV, using the device's built-in sound receiver, transmits a real-time alarm via satellite communication to the appropriate locations.

2 types of communication are provided in the control of the UAV:

- By 1<sup>st</sup> satellite communication;
- The 2<sup>nd</sup> is realized through the Internet network.

We use a hybrid communication system to ensure UAV communication. To implement satellite communication, the UAV has special signal receiver/transmitter antennas. Internet receiver/transmitter device is used to implement Internet network. UAV special internet receiver RTK-4671-MH dual frequency receiver designed for applications requiring centimeter accuracy. The Internet receiver is compact and economical and has a high accuracy GNSS RTK board. The RTK-4671-MH provides a continuous and reliable RTK solution using multiple systems such as GPS, GLONASS, BeiDou, GALILEO, QZSS and SBAS for improvement. In regions where there is no Internet network, radio communication is automatically activated. In addition, UAVs provide information about the emergency area to vehicles that arrive to help with reconnaissance from above. UAV reconnaissance opens up dams on the road ahead of time and communicates anticipated timelines or warns of difficult situations [15]. All processes and blockages are alerted to the opening. Operators work with specialized UAV control centers by direct contact. In remote control of UAVs, the operator implements all «order words»

in the Kazakh language. The order words are made up of the composition of words frequently used in day-to-day rescue, and depending on the functional functions in UAV control.

Is extremely effective for controlling intelligent robotic systems by transmitting audio commands. Does not require writing a program, ideal for use by ordinary people. Requires only a microphone and microcontroller and sensors, as well as a library of the necessary sound stock. We implement using sound orders. The sound control algorithm can be widely used in robotic moving objects other than UAVs. There were many examples that could be made at the time when various robotic system hardware devices were being developed. Sound control is just one of these examples. Sound control is implemented through special intelligent sensors and transducers. Control by sound command gives great possibilities for use, from small children to school children to the elderly to ordinary people.

Algorithm implemented by recording sound commands and entering into the microcontroller memory the sounds of the necessary persons involved in the control.

### 3. Conclusion

In conclusion, a model of algorithms for controlling the sound control of UAVs using remote sound commands was created. Sound commands a fund of words «order» in the Kazakh language was created on the basis of education. Under the control of the ATmega2560 microcontroller, the Knowledge Base Sentence words were entered into the memory vocabulary of the intelligent cognitive module. The sound team performed training exercises for the UAV, transmitting the words to the control operators. Work was also done on identifying and recognizing operator voices and bystander voices by incorporating an intelligent sensor with sensors. This algorithm for remotely controlling the UAV with audio command is very relevant. It speeds up the alarm system in emergencies and facilitates the work of employees in emergencies. In addition, the real-time UAV of the scene directly reported observations from the ceiling and saved the time spent working to reset. The video data recorded in the UAV's memory will allow for future review and investigation of the scene.

### 4. References

- [1] Development and Implementation of Automated UAV Flight Algorithms for Inertial Navigation Systems, Yemelyev, A.K., Moldamurat, K., Seksenbaeva, R.B., SIST 2021 - 2021 IEEE International Conference on Smart Information Systems and Technologies, 2021, 9465965.
- [2] Low-fidelity design optimization and development of a VTOL swarm UAV with an open-source framework, Sethi, N., Ahlawat, S., Array, 2022, 14, 100183, 2 Delhi Technological University.
- [3] Cloud robotics: A software architecture: For heterogeneous large-scale autonomous robots, Miratabzadeh, S.A., Gallardo, N., Gamez, N., Rad, P., Jamshidi, M. World Automation Congress Proceedings, 2016, 2016-October, 7583017.
- [4] Design and Optimization of Wing Structure for a Fixed-Wing Unmanned Aerial Vehicle (UAV). Jiawen Yu., Modern Mechanical Engineering > Vol.8 No.4, November 2018. DOI: 10.4236/mme.2018.84017.
- [5] Path Planning for Autonomous Mobile Robots: A Review, José Ricardo Sánchez-Ibáñez, Carlos J. Pérez-del-Pulgar and Alfonso García-Cerezo, <https://doi.org/10.3390/s21237898>.
- [6] Teleoperation by Using Nonisomorphic Mechanisms in the Master-Slave Configuration for Speed Control, Shukla, A., Karki, H., Behera, L., Jamshidi, M.M. IEEE Systems Journal, 2018, 12(2), p. 1369–1380.
- [7] An Intelligent Decision Support System for Aircraft Landing Based on the Runway Surface, Seitbattalov, Z.Y., Atanov, S.K., Moldabayeva, Z.S. SIST 2021 - 2021 IEEE International Conference on Smart Information Systems and Technologies, 2021, 9466000.
- [8] Development an Intelligent Task Offloading System for Edge-Cloud Computing Paradigm, Atanov, S.K., Seitbattalov, Z.Y., Moldabayeva, Z.S. Proceedings - 2021 16th International Conference on Electronics Computer and Computation, ICECCO 2021, 2021.
- [9] Feature Fusion Models for Deep Autoencoders: Application to Traffic Flow Prediction, Moussavi-Khalkhali, A., Jamshidi, M., Applied Artificial Intelligence, 2019, 33(13), стр. 1179–1198.

[10] The Usage of Extended Kalman Filter to Increase Navigation Accuracy of Mobile Units in Closed Spaces, Adilzhan, K.K., Sabyrzhan, A.K., Timur, T.Z. SIST 2021 - 2021 IEEE International Conference on Smart Information Systems and Technologies, 2021, 9465903.

[11] Utegen, A. S., Moldamurat, K., Ainur, M., Talgat, A., Amandykuly, A. G., & Brimzhanova, S. S. (2021). Development and modeling of intelligent control system of cruise missile based on fuzzy logic. Paper presented at the *Proceedings - 2021 16th International Conference on Electronics Computer and Computation, ICECCO 2021*, doi:10.1109/ICECCO53203.2021.9663808.

[12] Moldamurat, K., Utegen, A. S., Brimzhanova, S. S., Kalmanova, D. M., & Yrskeldi, N. G. (2021). Development of a software simulator for small satellite swarm control. Paper presented at the *Proceedings - 2021 16th International Conference on Electronics Computer and Computation, ICECCO 2021*, doi:10.1109/ICECCO53203.2021.9663828.

[13] Brimzhanova, S., Atanov, S., Moldamurat, K., Baymuhambetova, B., Brimzhanova, K., & Seitmetova, A. (2022). An intelligent testing system development based on the shingle algorithm for assessing humanities students' academic achievements. *Education and Information Technologies*, doi:10.1007/s10639-022-11057-w.

[14] Moldamurat Khuralai, Brimzhanova Saule, Bibianar Baizhumanova, Bizhanova Olga, Akhmetov Kairat, Moldamurat Aibek Computer simulation of the path and control of an intelligent mobile robot in Python.

[15] Viktorovna, U. N., Kulmukhambetovna, I. G., Serikovna, B. K., Pavlovna, R. L., & Kydyrbaevna, K. A. (2017). Formation of communicative competence as a condition of development of social orientation of the future teacher. *Man in India*, 97(16), 407-414. Retrieved from: [https://www.scopus.com/record/display.uri?eid=2-s2.0-85074669298&origin=AuthorNamesList&txGid=bf117a5b3cca9648f346045f3e0853b5&isValidNewDocSearchRedirection=false&featureToggles=FEATURE\\_NEW\\_DOC\\_DETAILS\\_EXPORT:1](https://www.scopus.com/record/display.uri?eid=2-s2.0-85074669298&origin=AuthorNamesList&txGid=bf117a5b3cca9648f346045f3e0853b5&isValidNewDocSearchRedirection=false&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1).

# Using of Machine Learning Algorithms to Determine the Content of Toxic Elements in the Environment

Lyazat Naizabayeva<sup>1</sup>, Nurgul Ainabek<sup>1</sup>, and Assem Berkimbayeva<sup>1</sup>

<sup>1</sup> International University of Information Technologies, Almaty, Kazakhstan

## Abstract

Computer (information) technologies have opened up enormous opportunities for studying the processes occurring in nature. Among the tasks that are successfully modeled on a computer, a special place is occupied by planning, forecasting, analysis and modeling of environmental processes in order to prevent adverse consequences of economic and other activities for systems of natural ecology.

Currently, the content of toxins in the air contributes to the deterioration of the well-being of urban residents, as evidenced by the data from the disease risk map of Kazakhstan.

The purpose of the article is to conduct an intellectual analysis of data and factors that have a negative impact on human health and propose an adaptive research algorithm for assessing toxic elements of the air.

Four machine learning algorithms were tested: Logistic Regression, Decision Tree, Random Forest, and K Nearest Neighbor. As a result, the Random Forests and K Nearest Neighbor algorithms gave the best performance in predicting the content of toxicity.

The main result of this study is the application of the exported model to predict new data. As a result, data were obtained on the limited dispersion of emissions, from which the subsequent thermal inversion causes severe air pollution in Almaty, regardless of the time of year.

## Keywords

Computer technology, intellectual analysis, climate change, air pollution, machine learning

## 1. Introduction

One of the biggest threats to human health is climate change. Environmental factors are responsible for 23% of all deaths worldwide, with an additional 250,000 climate-related deaths per year expected by 2030, according to the World Health Organization. geographical distribution and seasonality of infectious diseases. These changes disproportionately affect vulnerable populations, especially people with limited resources such as children, the elderly and those suffering from disease or health inequalities [1,2].

The relevance of the development of forecasting environmental risks is associated with the accumulation of numerous experimental materials on the content of heavy metals and other toxic elements of the environment in the air. In this regard, the improvement and implementation of information technologies in the environment, first of all, allows for the rapid exchange of information on a global and local scale with minimal costs of financial, labor resources and with maximum indicators of reliability, accuracy and objectivity. aimed at effectively preventing negative changes. in the natural environment [3-4].

As a result of data collection, the following goal of the study was identified - to conduct an intellectual analysis of data and factors that have a negative impact on human health, and improve the adaptive algorithm for assessing air quality.

### Tasks to achieve the goal:

- Analysis of the volume of emissions of pollutants into the atmosphere;
- Make a preliminary data analysis with the calculation of basic statistics;
- Suggest the best algorithm for assessing air quality.

Scientific novelty of the work: an adaptive model of regression equations was built to calculate the factors affecting air pollution, based on the processing of experimental data.

The ability of public health and safety systems to adapt to or prepare for these changing hazards, as well as individual behavior, age, gender and economic status, will determine the severity of these health

problems. The impact will vary depending on where a person lives, how vulnerable they are to health problems, how vulnerable they are to climate change, and how well they and their communities can adapt [5].

## 2. Review of related works

In recent years, climate change has become a more popular topic, attracting the attention of a growing number of people, from science to politics. Despite the scientific evidence for climate change, there are many points of view and methods for mitigating it. The European Commission Climate and Energy Package 2020 aims to reduce greenhouse gas emissions (compared to 1990 levels), increase energy efficiency by 20% and reduce greenhouse gas emissions (compared to 1990 levels). The report examined the strength of the evidence supporting its findings and determined relative levels of reliability. It is used by NASA and other organizations studying climate change over time, and new approaches have been introduced [6].

Air pollution refers to chemicals or particles in the air from anthropogenic or natural sources that pose a risk to the health of living beings. Increasing emissions from rapidly growing modern industry, urbanization and traffic, in addition to traditional biomass fuels, are affecting air quality in both developed and developing countries. According to the World Health Organization (WHO), more than 90% of the world's population breathes air that does not meet WHO standards, and seven million people die each year as a result of the adverse health effects of air pollution [7].

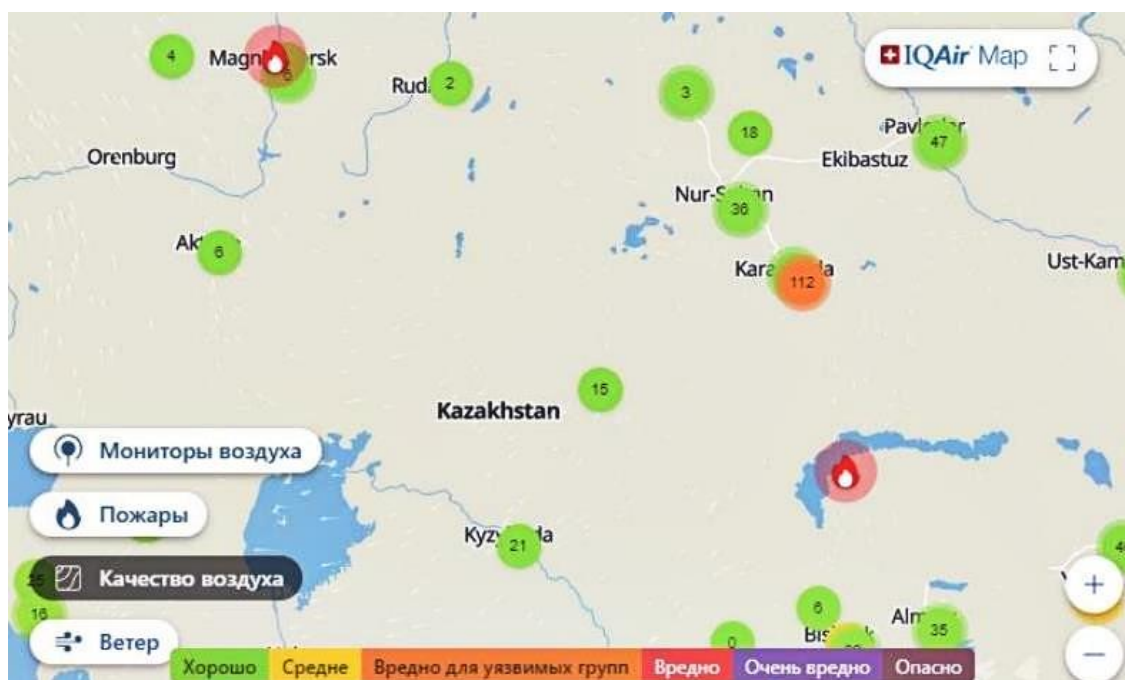
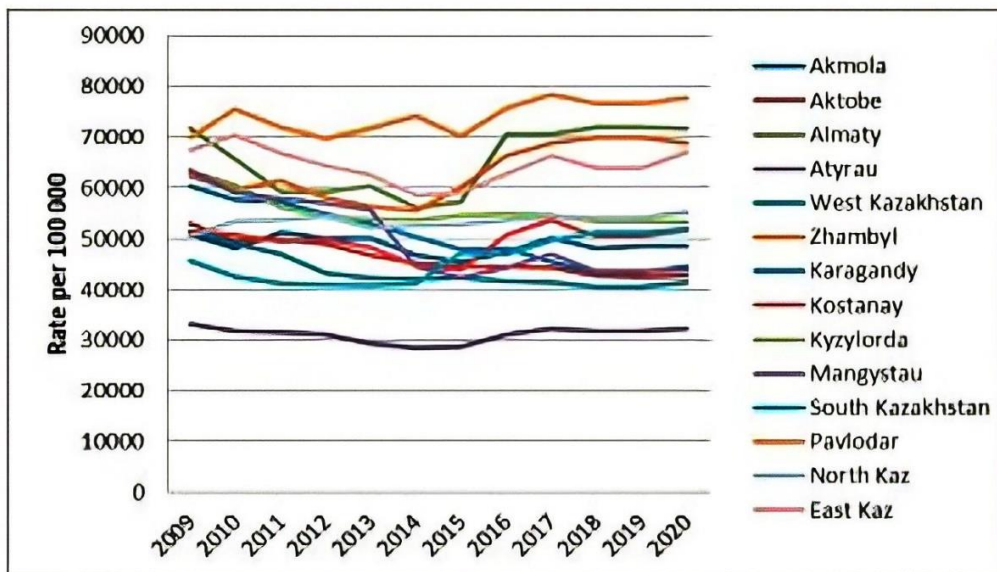


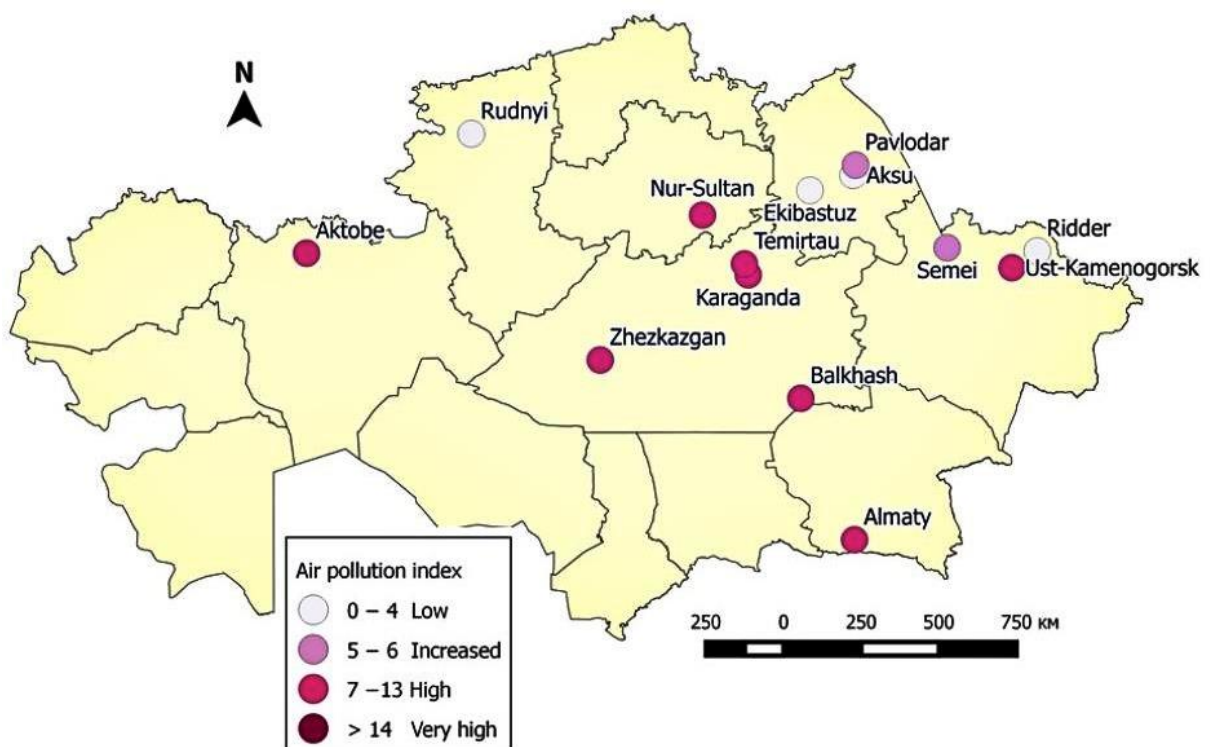
Figure 1: Air quality index (AQI) and PM2.5 air pollution in Kazakhstan

For this study, three diseases were taken into account, since the analysis of the data showed that these three diseases are pronounced among the population of Kazakhstan. There are: blood diseases, neoplasms and endocrine diseases. According to reports on statistical data collected by the Ministry of Health of the Republic of Kazakhstan from 2009 to 2021, namely the “Statistical compendium “Health of the population of the Republic of Kazakhstan and healthcare activities”, there were about 18 million people [8]. This compilation presents statistical materials on the activities of healthcare organizations and indicators of the state Health of the population of the Republic of Kazakhstan. The data are given per 100,000 populations per year [9]. The line graph below shows how the data for the 10-year regions of the Republic of Kazakhstan changed (Figure 2). The graph shows the rate of population growth in

cities and thus shows how many people are exposed to air pollution problems.



**Figure 2:** According to World Health Organization guidelines, air quality in Kazakhstan is considered moderately unsafe



**Figure 3:** Emissions of pollutants into the atmosphere, 2021

The most recent data show that the average annual concentration of PM<sub>2.5</sub> in the country is 14 µg/m<sup>3</sup>, which exceeds the recommended maximum of 10 µg/m<sup>3</sup>. Air quality in Kazakhstan can be affected by extractive industries such as oil, coal, iron, lead, agriculture and vehicle emissions. Available data indicate that in Pavlodar, Almaty region, there is a consistently high level of air pollution [10].

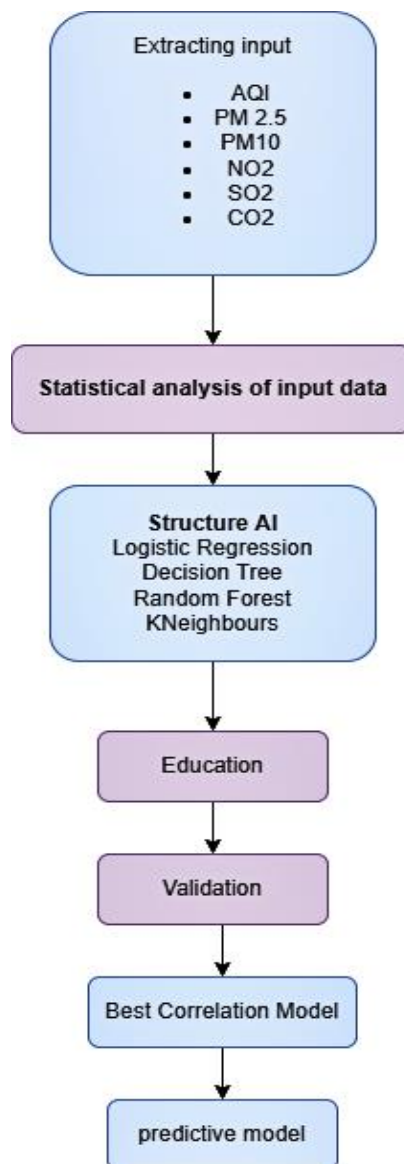


### 3. Materials and methods

#### 3.1. Logistic regression

Machine learning (ML) is becoming more and more popular in enterprise data analysis systems to extract useful information from data. Model development in machine learning involves collecting data from various trusted sources, processing the data to make it fit for model building, choosing a model building algorithm, calculating performance metrics, and choosing the most efficient model [11].

From the search for initial data to the results and predictive models, Figure 4 shows the steps used in this study.



**Figure 4:** Research stages

The analysis period was chosen from 2020 to 2022, during which all the necessary data were available. Table 1 presents a preliminary analysis of the data, with the calculation of key statistics useful for scaling.

Four machine learning algorithms were tested: Logistic Regression, Decision Tree, Random Forest,

and K Nearest Neighbor. In general, the Random Forests and K Nearest Neighbor algorithms gave the best results in content prediction.

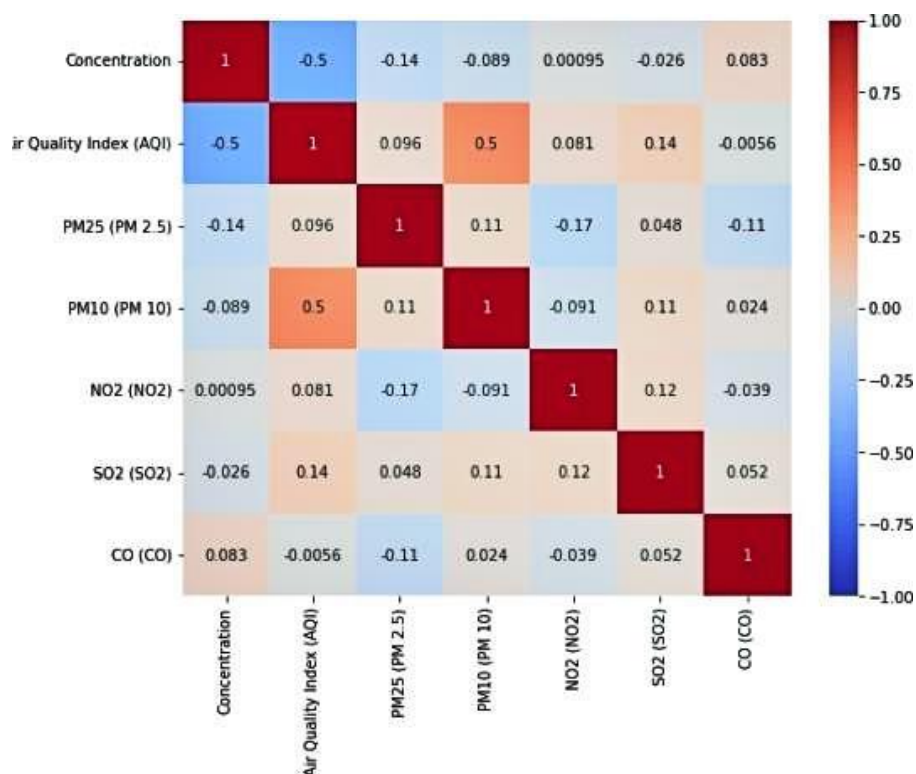
**Table 1**  
 Data analysis with calculation of key statistics for scaling.

	Concentration	Air quality Index(AQI)	Pm25 (PM 2,5)	PM10(PM 1-)	NO2(NO2)	SO2(SO2)	CO(CO)
Count	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000
Mean	0.673913	62.451085	26.256658	26.594781	33.239798	68.379645	634.057971
Std	0.470487	16.308412	11.184869	7.502475	7.437223	43.584301	69.644170
Min	0.000000	31.600000	14.6000000	12.300000	11.400000	15.699997	500.000000
25%	0.000000	51.824999	19.677875	22.967000	28.700001	52.350001	614.000000
50%	1.000000	60.950001	23.280000	26.164001	32.299999	61.100001	632.000000
75%	1.000000	69.782499	27.352525	29.175001	35.890000	70.199997	658.000000
max	1.000000	120.620000	98.699997	74.599998	71.099998	420.20000	931.000000

Logistic regression is a supervised learning technique that is one of the most commonly used machine learning algorithms. It is a method for predicting a categorical dependent variable from a set of independent variables. The output of the categorical dependent variable is predicted using logistic regression [12]. As a result, the result must be a discrete or categorical value. It can be yes or no, 0 or 1, True or False, etc., but instead of specifying the exact value as 0 or 1, it provides probabilistic values that range from 0 to 1. Logistic regression equation:

$$h_0(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x)}} \quad (1)$$

$h_0(x)$  is the output of the logistic function where  $0 \leq h_0(x) \leq 1$ ,  
 $\beta_1$  - slope,  
 $\beta_0$  intersection point with y coordinate,  
 $x$  is an independent variable,  $(\beta_0 + \beta_1 * x)$  line  $y$  obtained from the equation  $(predicted) = (\beta_0 + \beta_1 * x) * error$  value. The exported model can be used to predict new data (Figure 5).



**Figure 5:** Data visualization, to see how the value changes depending on the values of two other variables

### 3.2. Using the decision tree method classifier

The decision tree method was used, which is an inductive data mining technique that uses a greedy algorithm to recursively split a set of data records either "depth" or "breadth" until all data components belong to a particular class.

Relevant objects in the proposed set of objects are determined using tree induction. The method of examining a tree classifier by developing a tree structure in which each internal node (without a leaf node) is a feature test is known as decision tree induction. Each branch represents a result, and each other node (class prediction) represents a class prediction. At each node, the method finds the best section information for the given classes. The highest split score is selected by selecting an attribute with a set of facts. The attribute is of the highest possible quality and collects the maximum amount of data possible. Figure 6 shows the decision tree classification model in action.

```
from sklearn.tree import DecisionTreeClassifier
model_dt = DecisionTreeClassifier(max_depth=4, random_state=42)
```

**Figure 6:** Implementation of the decision tree classification model

The K-NN method assumes that the latest incident/data contains similarities along with existing incidents, and also ranks it in a systematization that is fully comparable with existing classes. The K-NN method preserves all easily accessible information without exception and also shows a different place of information in the approximation base. This means that the presence of the latest data, the K-NN calculation is usually able to instantly sort it according to the required class.

The K-NN calculation is also capable of being used for the purpose of regression in a ranking relationship, but as a rule one is used for the purpose of ranking questions [13,14].

When using the random forest algorithm to solve regression problems, you use the meansquare error (MSE):

$$MSE = \frac{1}{N} \sum_{i=1}^N (f_i - y_i)^2 \quad (2)$$

where  $N$  - is the number of data points,  
 $f - i$  is the value returned by the model, and  $y_i$  - is the actual value for data point  $i$ .

## 4. Results of the experiment

The results of the analysis are presented below (Table 2), where the Logistic Regression and KNeighbours classifiers achieved the highest accuracy compared to other models.

**Table 2**  
 Accuracy scores for each model classifier

	Model	Accuracy_score
0	Logistic Regression	0.826087
3	KNeighbours	0.782609
1	Decision Tree	0.760870
2	Random Forest	0.760870

As a result of the above, there is only limited dispersion of emissions, and the subsequent thermal inversion causes severe pollution in Almaty, regardless of the time of year. In Almaty, for example, heavy fogs observed on average 40-50 days a year, which usually coincides with critical levels of air pollution over the city [15,16].

## 5. Conclusion

The overall purpose of the research was to explore the mental exploration of information and conditions that have a negative impact on health and how exactly it affects the decision-making process.

Four machine learning algorithms were tested: Logistic Regression, Decision Tree, Random Forest, and K Nearest Neighbor. In general, the Random Forests and K Nearest Neighbor algorithms gave the best results in content prediction.

As a result, data were obtained on the limited dispersion of emissions from which further thermal inversion generates very significant air pollution in Almaty, regardless of the time of year.

The most recent data show that the average annual concentration of PM<sub>2.5</sub> in the country is 14 µg/m<sup>3</sup>, which exceeds the recommended maximum of 10 µg/m<sup>3</sup>. Air quality in Kazakhstan can be affected by extractive industries such as oil, coal, iron, lead, agriculture and vehicle emissions. Available data indicate that in Pavlodar, Almaty region there is a consistently high level of air pollution. Therefore, it is necessary to take immediate action in areas with a level of risk, such as a prevention and control strategy, rather than a response to an emerging epidemic.

Data mining by region of disease from environmental exposure may be sufficient to inform interventions in some cases for decision making. An analysis of the distribution of risk factors helped prioritize preventive measures. By understanding avoidable risk factors, disease mapping can be useful for health care delivery and individual interventions.

## 6. References:

- [1] K.R. Prabhat, J.S. Singh. "Invasive alien plant species: Their impact on environment,

ecosystem services and human health.” *Ecological Indicators*, Volume 111, 2020.

[2] G.L. Nicole, C. Claudia, B. Iñigo. Residential sidewalk gardens and biological conservation in the cities: Motivations and preferences that guide the floristic composition of a little-explored space, *Urban Forestry & Urban Greening*, Volume 63, 2021.

[3] D.H. Bogachan, P.M.J. Fisher. “How are nature based solutions helping in the greening of cities in the context of crises such as climate change and pandemics? A comprehensive review.” *Journal of Cleaner Production*, Volume 288, 2021.

[4] G.D. Kaziyeva, A.E. Abzhanova, S.K. Sagnayeva, G.K. Sembina, T.K. Ermagambetov. “Features of the implementation of olap-systems in environmental monitoring of the marine environment” *CEUR Workshop Proceedings*, 2020, 2570. <http://ceur-ws.org/Vol-2570/paper20.pdf>.

[5] Y. Kaluarachchi. “Potential advantages in combining smart and green infrastructure over silo approaches for future cities.” *Front. Eng. Manag.* 8, 98–108, 2021.

[6] H. Zhu. “Research on Agricultural Ecological Factors Information Technology based on Internet+” *Advances in Computer Science Research. 3<sup>rd</sup> International Workshop on Materials Engineering and Computer Sciences (IWMECS 2018).* – 2019. – Vol.78. – P.19-22.

[7] R.B.K. Singh, S. Hales, N. De Wet, R. Raj, M. Hearnden, P. Weinstein. “The Influence of Climate Variation and Change on Diarrheal Disease in the Pacific Islands” *Environ. Heal. Perspect.* 2021, 109,155–159.

[8] F.J. Doblas-Reyes, R. Hagedorn, T.N. Palmer. “The rationale behind the success of multi-model ensembles in seasonal forecasting— II.” *Calibration and combination. Tellus* 2015, 57, 234–252.

[9] Partners, Whiteshield, 2018. “Sustainable development goals and capability- based development in regions of Kazakhstan.” Available: *Natl. Hum. Develop.* 2018.

[10] D. Kalyanmoy. “Multi-Objective Optimization using Evolutionary Algorithms.” John Wiley & Sons, Ltd., Chichester, England, 2021.

[11] V.N. Dymnikov, E.M. Lykosov, V.P. Volodin. *Bulletin of the Russian Academy of Sciences climat modeling and its changes: modern problems*, 2016, vol. 82, no. 3, p. 227–236.

[12] V.E. Fodor, J. Pájer. “Application of Environmental Information Systems in Environmental Impact Assessment.” *Acta Silv. Lign. Hung.* – 2017. – Vol. 13, N. 1. – P.55–67.

[13] R.J. Lempert, M.E. Schlesinger. “Robust Strategies for Abating Climate Change.” *Climatic Change* 45 (3–4), 2019: 387–401.

[14] S. Patel, I.U. Sayyed. “Impact of information technology in agriculture sector” *JFAV.* – 2018. – Vol.4,Is.2. – P.17-22.

[15] L. Naizabayeva, Ch. Nurzhanov, J. Orzabekov, G. Tleuberdiyeva. “Corporate environmental information system data storage development and management.” *Central European Journal Open Computer Science*; Online ISSN 2299-1093; 7:29-35 pp 24-30. Web of Science Score Collection and Scopus, 2018.

[16] L. Naizabayeva, M.S. Arinova. “Intellectual analysis and prediction of toxic elements in soil.” *International Journal of Information and Communication Technologies*, Volume 2, Issue 1. March 2021, pp.40-47.

# Machine Learning Algorithms for Biometric Face Identification

Timur Shormanov<sup>1</sup>, Aigerim Mazakova<sup>1</sup>, Sholpan Jomartova<sup>1</sup>, Talgat Mazakov<sup>1</sup>, and  
Majit Orynbay<sup>1</sup>

<sup>1</sup> Al-Farabi Kazakh National University, Almaty, Kazakhstan

## Abstract

The article is devoted to the study of the problem of biometric identification of a person by face. The article uses various linear classifiers: 1) support vector machine, 2) neural networks and 3) histograms of directional gradients.

In this study, a database of photographs obtained from open sources was used, and the result of the work was the name and photograph of a person who had already been previously identified in the database.

The analysis of algorithms for searching and identifying faces in images showed that it is effective to use artificial neural networks to solve this problem, due to the fact that they provide the possibility of obtaining a face classifier with a high degree of accuracy, which well models the complex function of determining face images. An experimental study of biometric face identification, created on the basis of the proposed convolutional neural network on 128 face measurements, showed that the developed software system is invariant to face image rotations. Able to work in a wide range of lighting changes from 50 to 100% of the lighting level in the image, and also has invariance to scale changes and other distortions.

## Keywords

Machine learning, neural networks, biometrics, identification

## 1. Introduction

Machine learning methods are widely used in various fields of activity from the creation of speech recognition algorithms and automatic translation to growing vegetables. The first developments in the field of machine learning were made in the 40s of the last century. In the early 2000s, due to the growth of computing power of computers and the boom of information systems, new, more advanced machine learning algorithms began to appear. Now machine learning algorithms are the main tool for improving performance in various industries such as information retrieval systems, recommendation services, diagnostic medicine, finance, and many others. The exponential growth of data arrays, as well as the increasing complexity of the tasks being solved, have led to the need to develop new algorithms in many areas that are somehow related to data collection and analysis. For example, on average, one offshore rig generates 50 TB of data per year, and less than 1% of it is of practical value.

One of the topical areas of application of machine learning algorithms is image processing and recognition. Image recognition is used in a wide variety of applications – it can be access control, personal identification, search in an image file, and so on.

As for the use of machine learning algorithms for biometric face identification, this area has received a new development since the early 2000s, when various face recognition algorithms appeared. Thanks to various methods of digital processing and image analysis, such as the Viola-Johnson method [1], [2], directional gradient histograms [3], [4], etc., it became possible to quickly obtain unique characteristics for each person. However, none of the existing methods of face identification is universal, and therefore the search for solutions based on new methods of analysis is an urgent task.

As mentioned, due to the variety of analysis methods, it is possible to extract a sufficient number of unique face characteristics for each person, and therefore the task of optimizing the interpretation of the data obtained becomes relevant. Machine learning is perfect for this purpose.

The main idea of machine learning is to speed up the process of processing data and identifying patterns that will be useful in identification, and based on this information, a face in the image is identified.

The use of machine learning algorithms for biometric facial identification can be broken down into several interrelated tasks:

- Image analysis and face search;
- Face recognition, while the image may not be complete, poor lighting, the person turned his head, changed his hairstyle, and so on;
- Identification of unique facial features that distinguish one person from another, such as eye size, face shape, and so on;
- Classify and compare identified unique face features with all the people the system already knows to understand who is in the photo.

The objects of analysis are photographs of faces obtained from open sources, their unique features are a set of characteristics, and the output is the name and photographs of a person who has already been previously identified in the database. Training takes place on a sufficient volume of precedents.

## 2. Research methods

As mentioned above, the task of biometric identification of faces is one of the tasks solved using machine learning. To solve this problem, hybrid systems are best suited, including both supervised learning and artificial neural networks.

In this study, a database of photographs obtained from open sources was used, and the result of the work was the name and photograph of a person who had already been previously identified in the database. There are certain dependencies between the personal characteristics of a person that need to be established. For this, so-called precedents are used, that is, such sets of images of people that have already been identified using this algorithm. Such precedents are called training samples. Based on them, the classification algorithm is trained. In this study, we considered and used both linear classifiers – the support vector method and neural networks and histograms of directional gradients.

### A. Histogram of directional gradients

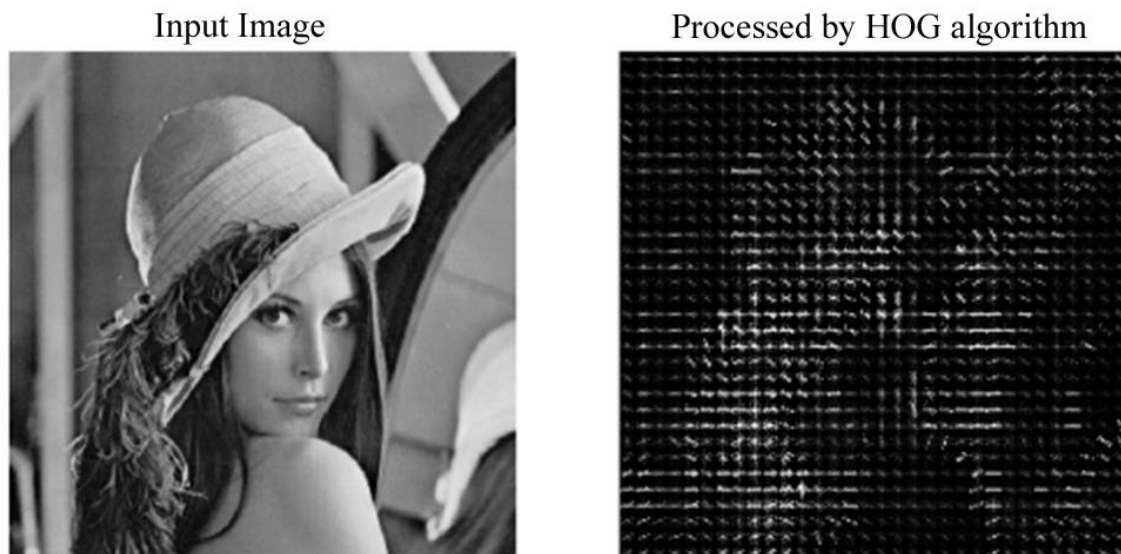
The Histogram of Oriented Gradients method analyzes an image and splits it into pixels, in which it finds out how dark the current pixel is compared to neighboring ones. Then an arrow is drawn showing in which direction the image becomes darker, by performing this procedure for each individual pixel of the image, the pixels are replaced by the direction arrows. These arrows are called gradients and they describe the direction from light to dark pixels throughout the image and can be described by the distribution of intensity gradients or edge direction. The combination of gradients is called a descriptor. In order to increase the accuracy, the processed image is usually made black and white, and local histograms are normalized by contrast relative to the measure of intensity calculated on a larger image fragment (Figure 1). Contrast normalization makes it possible to achieve greater invariance of descriptors to illumination [1].

A minor disadvantage of the system is that when constructing a histogram by cells of a constant size, it is necessary that all images lead to a resolution common to the sample.

The final step in object recognition through a directional gradient histogram is the classification of the resulting descriptors with a trained support vector method (SVM) classifier.

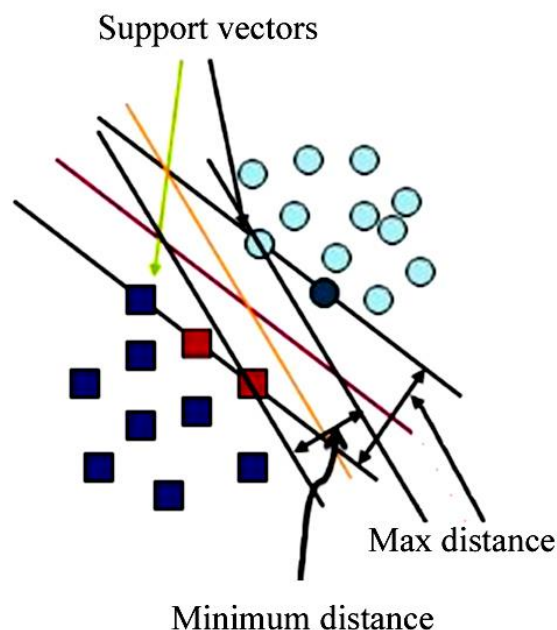
### B. Support Vector Method

The support vector method belongs to the group of linear classifiers. The purpose of the method is to find a plane (in the case of a multiclass classification, a hyperplane) that separates sets of objects. The separating hyperplane will be a hyperplane that maximizes the distance to two parallel hyperplanes (Figure 2). The algorithm works under the assumption that the greater the difference or distance between these parallel hyperplanes, the smaller the average classifier error will be.



**Figure 1:** The result of image processing by the HOG algorithm. This figure was created by [Tanase](#)

The support vector method is unstable with respect to noise in the original data and data standardization. If the training sample contains noise emissions, this method is inappropriate to use [2].



**Figure 2:** An example of a separating hyperplane This figure was created by [Milind Paradkar](#)

Advantages of the method: this is the fastest method for finding the decisive functions. It reduces to solving a quadratic programming problem in a convex domain, which always has a unique solution. The method finds a separating strip of maximum width, which allows further more confident classification.

Appendix 1 shows a program that analyzes an image using HOG transformation algorithms, followed by SVM classification, to recognize faces in the image, a frame is drawn around each recognized face. The result of the program is indicated in graphical form (Figure 3).

#### *C. Method for assessing facial landmarks*

The next task after searching and classifying a face in an image is the task of finding a face when

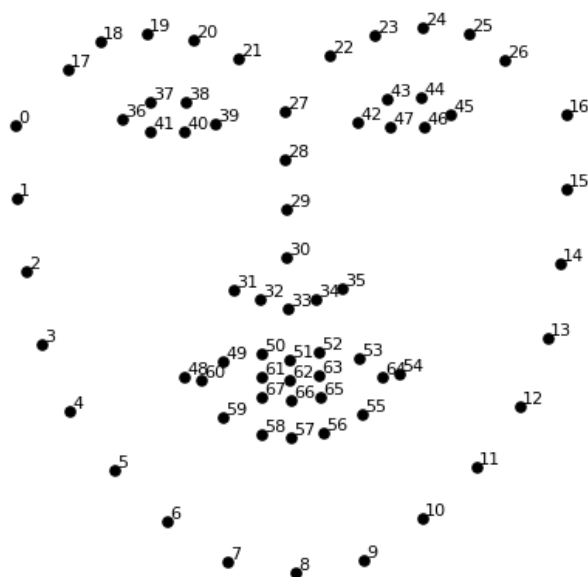


the face is turned or tilted in different directions. To solve this problem, the face landmark estimation method is used. This algorithm was proposed in 2014 by Vahid Kazemi and Josephine Sullivan [5]. The algorithm involves transforming each image so that the eyes and lips are always in a particular place. To do this, it is necessary to mark 68 special points (landmarks) that exist on each face – the upper part of the chin, the outer point of each eye, the inner point of each eyebrow, and so on (Figure 4). Then we need to train the machine learning algorithm to find these 68 feature points on any face.



**Figure 3:** Result of applying HOG and SVM analysis

After determining the main unique facial parameters such as eyes, mouth, nose, eyebrows. It is possible to scale, rotate or shift the image so that the eyes and mouth are as well centered as possible.



**Figure 4:** 68 anthropometric points on each face. This figure was created for [OpenFace](#)

Benefits of the method: all image transformations will only use basic transformations such as rotation and scaling that preserve parallel lines. No matter how the face is rotated, we can center the eyes and mouth at roughly the same position in the image. Below (Figures 5 and 6) is an example of anthropometric points assessment for the photo of Nurlan Koyanbaev.

#### *D. Convolutional Neural Networks*

A convolutional neural network (CNN) is a special architecture of neural networks proposed in 1989 and designed for image recognition [7]. Architecture copies the features of the cerebral cortex. Simple cells react when they perceive straight lines at various angles, the reaction of complex cells is associated with a certain set of simple cells. Convolutional neural networks use three kinds of layers: convolution,

pooling (also called subsampling or subsampling layer), and fully connected layer.



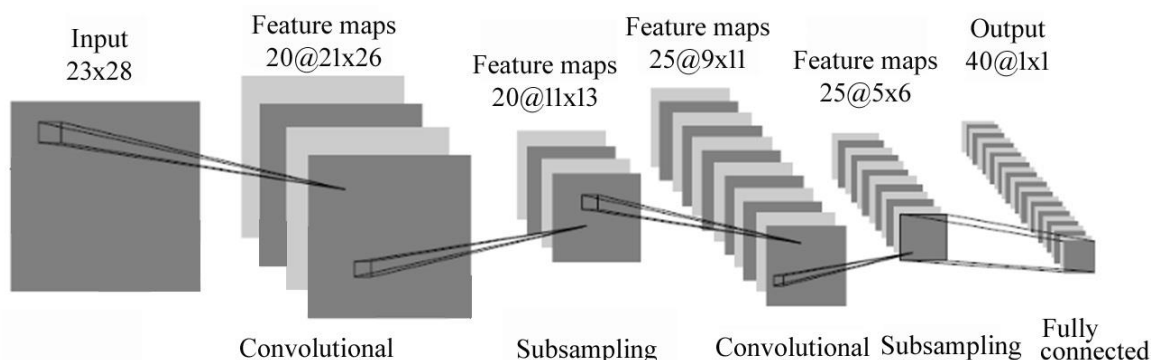
**Figure 5:** Analysis of 68 anthropometric points



**Figure 6:** Analysis of anthropometric points in case of head rotation

The structure of the network is unidirectional, multilayer, for training, as a rule, the method of error back propagation is used, the activation function of neurons is determined by the researcher. This architecture got its name due to the use of the convolution operation, which consists in element-by-element multiplication of each image fragment by the convolution kernel and then writing the result to the corresponding position of the output image [8]. (Figure 7)

The convolution operator forms the basis of the convolutional layer of the network. The layer consists of a set of kernels and calculates the convolution of the output image from the previous layer using this set, adding an offset corresponding to the kernel at each iteration. The result of this operation is the addition and scaling of the input pixels, the kernels can be obtained from the training set using gradient descent, similar to the calculation of weights in fully connected networks, which can also perform these operations, but require much more time and data for training. However, in comparison with fully connected networks, convolutional networks use a larger number of hyperparameters – parameters set before the start of training, such as: depth (number of cores and bias coefficients in the layer), height and width of each core, step (kernel shift at each step when calculating next pixel). The pooling layer takes individual image fragments as output (usually 2x2) and combines them into a single value. There are various aggregation methods, usually the largest value is selected from the resulting fragment.



**Figure 7:** Convolutional Neural Network Architecture. This figure was created by Steve Lawrence et. al

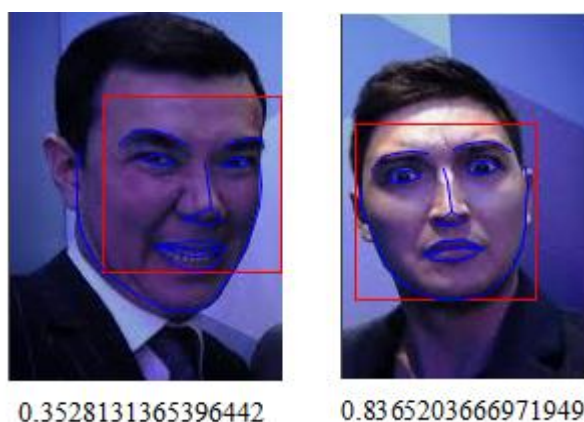
The use of a neural network for biometric identification involves the analysis of network training on 128 dimensions (descriptors) simultaneously for three persons:

1. Teaching face image;
2. Another photo of that person;
3. Image of a completely different person.

The algorithm then looks at the measurements it makes for each of those three images. It then tweaks the neural network a bit to make sure that the measurements generated for images 1 and 2 are more similar and the measurements for image 2 and 3 are less similar. By repeating this step for a statistically large number of images of thousands of different people, the neural network learns to create 128 measurements for each person. The resulting 128 measurements of each face are called a map or descriptor. The map obtained as a result of the algorithm is compared with the maps (descriptors) of other people in the database, the maps of the same people must match with a high probability.

The disadvantage of using convolutional neural networks is their exactingness to the computational hardware. Until recently, training a large neural network was too slow, and only recently, with the advent of video cards with 3D graphics, has it become technically possible to effectively apply neural networks in order to analyze an image.

Appendix 2 shows a program that receives face descriptors from two images, followed by their comparison, as well as the results of comparing different images using the example of a photograph of Nurlan Koyanbaev taken in different poses and different ages and different sizes. For analysis, a similarity index from 0.0 (identical images are analyzed) to 0.6 means that this is the same face, while the higher the value, the less similar the faces are (Figure 8).



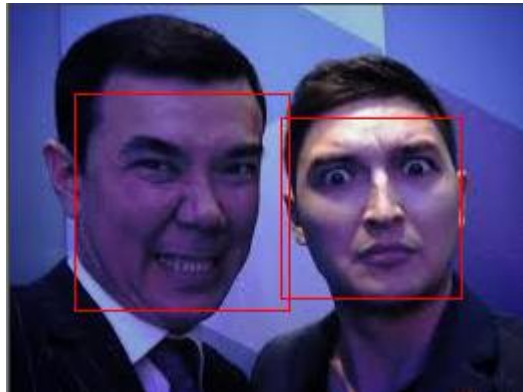
**Figure 8:** Analysis of face descriptors indicating the degree of similarity

### 3. Results of the study

#### Image analysis by HOG and SVM methods

```
from skimage import io
file_name = "images1.jpg"
# HOG image analysis and face search
face_detector = dlib.get_frontal_face_detector()
win = dlib.image_window()
image = io.imread(file_name)
detected_faces = face_detector(image, 1)
print("Found a face in the image".format(len(detected_faces), file_name))
win.set_image(image)
for i, face_rect in enumerate(detected_faces):
    print("- Face #{} found at Left: {} Top: {} Right: {} Bottom: {}".format(i,
    face_rect.left(), face_rect.top(),
    face_rect.right(), face_rect.bottom()))
win.add_overlay(face_rect)
dlib.hit_enter_to_continue()
```

#### Image analysis result



#### Found a face in the image

- Face #0 found at Left: 32 Top: 44 Right: 139 Bottom: 152
- Face #1 found at Left: 135 Top: 56 Right: 225 Bottom: 146

#### Image analysis by HOG and SVM methods

```
import dlib
from skimage import io
from scipy.spatial import distance
sp = dlib.shape_predictor('shape_predictor_68_face_landmarks.dat')
facerec = dlib.face_recognition_model_v1('dlib_face_recognition_resnet_model_v1.dat')
detector = dlib.get_frontal_face_detector()
# Loading and analyzing the first image
img = io.imread('images2.jpg')
win1 = dlib.image_window()
win1.clear_overlay()
win1.set_image(img)
dets = detector(img, 1)
for k, d in enumerate(dets):
    print("Detection {}: Left: {} Top: {} Right: {} Bottom: {}".format(
        k, d.left(), d.top(), d.right(), d.bottom()))
    shape = sp(img, d)
    win1.clear_overlay()
    win1.add_overlay(d)
    win1.add_overlay(shape)
Detection 0: Left: 278 Top: 30 Right: 353 Bottom: 105
face_descriptor1 = facerec.compute_face_descriptor(img, shape)
# Loading and analyzing the second image
```

```
img = io.imread('images5.jpg')
win2 = dlib.image_window()
win2.clear_overlay()
win2.set_image(img)
dets_webcam = detector(img, 1)
for k, d in enumerate(dets_webcam):
    print("Detection {}: Left: {} Top: {} Right: {} Bottom: {}".format(
        k, d.left(), d.top(), d.right(), d.bottom()))
    shape = sp(img, d)
    win2.clear_overlay()
    win2.add_overlay(d)
    win2.add_overlay(shape)
Detection 0: Left: 270 Top: 98 Right: 425 Bottom: 253
# Comparison of descriptors obtained from two images
face_descriptor2 = facerec.compute_face_descriptor(img, shape)
a = distance.euclidean(face_descriptor1, face_descriptor2)
print(a)
0.4177814853297572
```

Image analysis result:

The first image against which the others will be compared.



Second image



Similarity index 0.4177814853297572

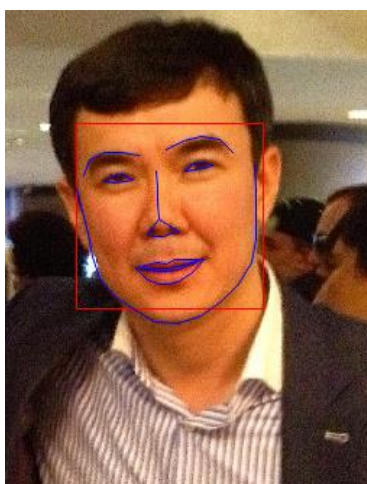
Results of analysis of other images:



Similarity index 0.3528131365396442



Similarity index 0.37658814616551456



Similarity index 0.25094970508809716



Similarity index 0.37845888198332295



Similarity index 0.8365203666971949

The analysis of algorithms for searching and identifying faces in images showed that it is effective to use artificial neural networks to solve this problem, due to the fact that they provide the ability to obtain a classifier (map) of a face with a high degree of accuracy, which well models the complex function of determining face images. An experimental study of biometric face identification, created on the basis of the proposed convolutional neural network on 128 face measurements, showed that the developed software system is invariant to face image rotations. It is able to work in a wide range of lighting changes from 50 to 100% of the lighting level in the image, and also has invariance to zoom changes and other distortions.

#### 4. Conclusion

Based on the results of the analysis of the effectiveness and speed of methods and algorithms for biometric identification of persons, the following conclusions can be drawn.

The use of a hybrid approach for carrying out biometric identification of faces allows you to create on its basis a software system for quick search and subsequent face recognition in various images, where there are different backgrounds, styles and a different number of people, followed by biometric identification, these techniques are more suitable for streaming video analysis.

Algorithms using convolutional neural networks have better classifying abilities when solving the problem of biometric identification, but require much more computing power from the hardware and are more suitable for tasks that require greater accuracy. An example of passport control, where several portrait images are analyzed, one from a surveillance camera, the second from a document, as well as both images are identified and compared with a database on which markup was carried out in advance.

In the course of the research work, the following tasks were performed, both hybrid methods for biometric identification of faces in images and the use of convolutional neural networks for biometric identification of faces were investigated and analyzed. A promising area of application of algorithms is video surveillance and access control systems to prevent unauthorized access.

We consider the use of interval computations to study the tasks of biometric identification of a person to be a promising direction. [14, 15].

The work was carried out at the expense of program-targeted funding of scientific research for 2021-2022 under the IRN project OR11465437 «Development of a national electronic data bank on the scientific zoological collection of the Republic of Kazakhstan, ensuring their effective use in science and education».

#### 5. References

- [1] D. Navneet, B. Triggs. “Histograms of oriented gradients for human detection.” Computer Vision and Pattern Recognition (CVPR) (2005), doi: [10.1109/CVPR.2005.177](https://doi.org/10.1109/CVPR.2005.177).
- [2] P. Viola and M.J. Jones, «Rapid Object Detection using a Boosted Cascade of Simple Features», Conference: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on Volume: 2001, doi: [10.1109/CVPR.2001.990517](https://doi.org/10.1109/CVPR.2001.990517).
- [3] P. Viola and M.J. Jones, «Robust real-time face detection», International Journal of Computer Vision, vol. 57, no. 2, 2004., pp.137–154, doi: [10.1023/B%3AVISI.0000013087.49260.fb](https://doi.org/10.1023/B%3AVISI.0000013087.49260.fb).
- [4] T. Hastie, R. Tibshirani, J. H. Friedman, The elements of statistical learning : data mining, inference, and prediction, Second. New York, NY, USA: Springer, 2009. ISBN: 978-0387848570. doi: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7).
- [5] V. Kazemi, J. Sullivan, “One millisecond face alignment with an ensemble of regression trees” Computer Vision and Pattern Recognition (CVPR), 2014. doi: [10.13140/2.1.1212.2243](https://doi.org/10.13140/2.1.1212.2243).
- [6] F. Schroff, D. Kalenichenko, J. Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering” Computer Vision and Pattern Recognition (CVPR), 2015. doi: [10.1109/CVPR.2015.7298682](https://doi.org/10.1109/CVPR.2015.7298682).
- [7] Y. LeCun, B. Boser, J. Denker et al., “Handwritten Digit Recognition with a Back-Propagation Network” Neural Information Processing Systems Conference. - 1990. - No.2.-P.396-404.
- [8] S. Lawrence, C. Lee Giles, A.Ch. Tsoi, A.D. Back (1997). "Face Recognition: A Convolutional Neural Network Approach". IEEE Transactions on Neural Networks. 8 (1): 98– 113 doi: [10.1109/72.554195](https://doi.org/10.1109/72.554195).
- [9] V.A. Dolgov. Overview of image recognition methods // Modern trends in technical sciences: materials of the VI International Scientific Conference. (Kazan, May 2018). – Kazan: Young scientist, 2018. – P. 7-9.
- [10] Vincent Dumoulin and Francesco Visin. “A guide to convolution arithmetic for deep learning.” arXiv preprint arXiv:1603.07285 (2016).
- [11] Y. LeCun. Convolutional networks for images, speech, and timeseries / Y. LeCun, Y. Bengio // The Handbook of Brain Theory and Neural Networks. 1995. P. 255-258.
- [12] J. Howse, J. Minichino. Learning OpenCV 3 Computer Vision with Python – Second Edition, Packt Publishing, September 2015, Packt Publishing, ISBN: 9781785289774.
- [13] K. He, X. Zhang, Sh. Ren, J. Sun. “Deep Residual Learning for Image Recognition” 2016 IEEE Conference On Computer Vision And Pattern Recognition (CVPR)2016. doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [14] M. Masakazu, K. Mori, Y. Mitari, Y. Kaneda. “Subject independent facial expression recognition with robust face detection using a convolutional neural network” 2003 Neural Networks 16 pages 555–559, doi: [10.1016/S0893-6080\(03\)00115-1](https://doi.org/10.1016/S0893-6080(03)00115-1).
- [15] G. Bairbekova, T. Mazakov, Sh. Jomartova, S. Nugmanova. “Interval arithmetic in calculations” Open Engineering Formerly Central European Journal of Engineering Editor-in-Chief: Noor, Ahmed, Open Engineering. Volume 6, Issue 1. – P.259-263, DOI: <https://doi.org/10.1515/eng-2016-0036>, October 2016.
- [16] Sh.A. Jomartova, T.Zh. Mazakov, N.T. Karymsakova, A.M. Zhaydarova. “Comparison of Two Interval Arithmetic”. // Applied Mathematical Sciences, Vol. 8, 2014, no. 72. – P.3593 – 3598. doi: <https://dx.doi.org/10.12988/ams.2014.44301>.



# Information Search Algorithms

Bakhtygerey Sinchev<sup>1</sup>, Anel S. Auyezova<sup>1</sup>, Tolkynai S. Sadykova<sup>1</sup>, and  
Aksulu M. Mukhanova<sup>2</sup>

<sup>1</sup> International Information Technology University, Almaty, Kazakhstan

<sup>2</sup> Kainar Academy, Almaty, Kazakhstan

## Abstract

The paper considers search algorithms for unstructured (text) data. Documents are searched for by natural language keywords used in search engines. The proposed search algorithms differ from existing methods in terms of time and memory used, as well as the ease of implementation of software products. Algorithms for sampling a subset satisfying the sum (certificate)  $S$ , problems about the sum of subsets, lemmas and algorithms for solving the problem of finding unstructured data based on a search query with several (two or three) keywords are given. The time and memory required for a search query with two keywords is proportional to  $O(n)$ . The task of information search with three keywords is reduced to the task of information search with two keywords.

## Keywords

Search, method, algorithm, unstructured information

## 1. Introduction

One of the first fundamental reviews of information retrieval problems and search engines was presented in [1]. To perform a search, many search engines build logical and physical data structures based on the initial information, which are a search index that allows you to implement some given model of information retrieval. Let's define the main types of search queries. Boolean search is one of the most important parts in information search that we encounter everywhere. The whole Boolean search is based on a combination of AND, OR and NOT. Relevance search is how much the search result meets a person's expectations. More detailed information about the types of search is given in [2]. Currently, several methods of sequential and binary searches are used. Sequential search involves sequentially viewing all the items from the list in the order of their location until the desired item is found. The systems using these keyword search methods include most of the widely used search engines, among them such Web search systems as Yandex, Google, AstaVista, Yahoo, etc.

## 2. Basic concepts of information search

The task of information retrieval is to find one or more elements in the set, and the desired elements must have a certain property. This property can be absolute or relative. A relative property characterizes in relation to others: for example, the minimum element in a set of numbers.

*Definition 1.* Let's call an alphabet a finite set of characters  $A = \{\tau, \alpha_1, \dots, \alpha_k\}$ , where  $\tau$  is a space character,  $|A| = k$ , and  $k > 0$  is the number of characters in the alphabet.

*Definition 2.* We will call a word a finite sequence of characters from the alphabet that does not include the space character  $\tau$ . We will assume that the set of words  $W$  is always finite.

*Definition 3.* A string of length  $n$ , we will call a sequence of words  $D = \{d_1, \dots, d_n, \$\}$ , where  $\forall i$ ,  $d_i \in W$  and  $\$$  are a special character that does not belong to the alphabet and denotes the end of the string.

*Definition 4.* The search query  $P = \{p_1, \dots, p_m\}$  is a string consisting of a finite set of words separated by a space character  $\tau$ . In this case,  $|P| = m$  - is the length of the query in words,  $p_i \in W$ ,

where  $i$  - is the number of the word in the template, and  $W$  - is the set of all words. Words in search queries and documents will be called terms (keywords).

*Note.* The length of the search query  $P$  will always be denoted by the symbol  $m$ , and the total length of the source data  $D$ , for which the search problem will be solved, will be denoted as  $n$ .

There is a  $2 \times n$  table and a given number  $S$ . You need to find 2 numbers from different rows (one from each row) that add up to  $S$ . The big notation describes the complexity of running time using algebraic terms.

Algorithm A. Brute force. Running time  $-O(n^2)$ .

Algorithm B. Search with sorting. Sort the first row, for each element in the second row, subtract it from  $S$ , and look for that difference in the first row. Running time  $-O(n \log n)$ .  $O(n)$  memory requirement.

There is a  $3 \times n$  table and a given number  $S$ . You need to find 3 numbers from different rows that add up to  $S$ .

Algorithm A. Brute force. Running time  $-O(n^3)$ .

Algorithm B. For one row, find all the differences with  $S$ , and for the other two, iterate through all the options. Running time  $-O(n^2)$ .

There is a  $4 \times n$  table and a given number  $S$ . You need to find 4 numbers from different rows that add up to  $S$ .

Algorithm 1. There is a complex Schreppel-Shamir algorithm with running time  $O(n^2 \log n)$ .

In the future, this material will allow a comparative analysis with the proposed scientific results below.

### 3. Open problems of unstructured text information search

Works [3,4] are devoted to solving the problem (subset sum problem), in which the search time  $T=O(2^{n/2})$  and the required memory  $M=O(2^{n/4})$  do not allow applying the results obtained in practice. The main disadvantage of table methods is the construction of each row of the table according to the property defined by each keyword. This means that we are obliged to carry out preliminary work on some structuring of the input data. In turn, there is an additional problem of splitting the vector space into subspaces according to each keyword.

Therefore, we will change the formulation of the problems of tabular sums in a more generalized form, suitable for the practical search for any information, using the subset sum problem. In the future, algorithms for solving these problems can be directly applied to search for arbitrary unstructured information based on a vector-spatial model and a search digital index.

Let us reformulate the statement of the main problem directly related to the length  $m$  of the search query.

*The main practical problem.* Given a set of  $n$  numbers and a number  $S$ . It is required to find out if there are one or more subsets, each of which consists of  $m$  elements, and the sum of these elements is equal to  $S$ .

1. Given a set of  $n$  numbers and a number  $S$ . It is required to find out if there are one or more subsets of two numbers whose sum of elements is equal to  $S$  and with running time less than  $O(n \log n)$ .

2. Given a set of  $n$  numbers and a number  $S$ . It is required to find out if there are one or more subsets of three numbers whose sum of elements is equal to  $S$  and with running time less than  $O(n^2)$ .

The solution of the following search problems is considered in the second part of the work.

3. Given a set of  $n$  numbers and a number  $S$ . It is required to find out if there are one or more subsets of four numbers whose sum of elements is equal to  $S$  and with significantly less running time than  $O(n^3)$ .

4. Given a set of  $n$  numbers and a number  $S$ . It is required to find out if there is one or more subsets of  $m$  numbers whose sum of elements is equal to  $S$  and with significantly less running time than  $O(n^{m-1})$  ( $m$  is greater than or equal to 5).

Now we can move on to mathematical formulations of search problems and their solutions.

*The main practical problem.* Given a set of integer (natural) numbers  $(x_1, x_2, \dots, x_n) \in X^n$  dimensions  $n$ . It is required to find out whether there exists a subset  $X_m$  of dimension  $m$  such that the following conditions are satisfied:

$$X_m = \{x_i + x_j + \dots + x_g + x_h = S, i \neq j \neq \dots \neq g \neq h, x_i, x_j, \dots, x_g, x_h \in X^n, (i, j, \dots, g, h) \in N = (1, 2, \dots, n), m \leq n\} \quad (1)$$

Here  $x_i, x_j, \dots, x_g, x_h \in X_m$ , and the number of elements  $x_i, x_j, \dots, x_g, x_h$  equal  $m$ .

Let us introduce the notation:  $C_n^m$  - combination,  $S_n^m$  - sum of elements of one subset from the set of subsets  $X_m$  of the set  $X^n$ . In this case, the variable  $m$  can vary from 0, 1, 2, ...,  $n$ . The totality (set) of these subsets  $X_m$  is determined based on the combination

$$C_n^m = \frac{n!}{m!(n-m)!} \quad (2)$$

Let's sort the given vector  $x$  from the set  $X^n$  in descending order and get the sorted set  $Z^n$  - this set of vectors  $x$ , the values of the elements of which are sorted in descending order, and find the values

$$S_{min}^m = \sum_0^m z_{n-m} \quad (3)$$

$$S_{max}^m = \sum_0^m z_m \quad (4)$$

It should be noted that

$$S_{min}^0 = S_{max}^0 = 0, S_{min}^n = S_{max}^n = \sum_1^n x_i = \sum_1^n z_i \quad (5)$$

Compose the possible ranges of ownership of the certificate  $S$  corresponding to a subset from this set of subsets  $X_m$ ,

$$S \in [S_{min}^n, S_{max}^n] \quad (6)$$

The solution of the subset sum problem is based on the following lemmas.

It is not difficult to find the running time of the algorithm based on Theorem 1 using a sorted vector  $x$  and the merge method:

$$T = O(C_n^m) < O(2^n) \quad (7)$$

The required memory  $M = O(n)$  is needed to save the vector  $e$ . Vectors  $e$  can be generated based on the Gray code.

*Example 1* from [4]. Consider the vector  $x = (7, 3, 9, 6, 2)$ ,  $S = 11$ ,  $C_5^2 = C_5^3 = 10$ ,  $S \in [S_{min}^2, S_{max}^2] = [5, 16]$  or  $S \in [S_{min}^3, S_{max}^3] = [11, 22]$ . Then the solutions of the subset sum problem are the vectors  $e = (00101)$ ,  $\bar{e} = (01011)$  and  $sum S = (e, x) = 11$  or  $S = (\bar{e}, x) = 11$ , for  $S = 10$ ,  $e = (11000)$ .

Let's move on to solving practical problems.

Task1. It is required to find out if there is a subset

$$X_2 = \{x_i + x_j = S; i \neq j; x_i, x_j \in X^n; i, j \in N\} \quad (8)$$

where  $X^n = (x_1, x_2, \dots, x_n)$  - set of integer (or natural) numbers,  $N = (1, 2, \dots, n)$  - set of natural numbers.

Task 2. It is required to find out if there is a subset

$$X_3 = \{x_i + x_j + x_k = S; i \neq j \neq k; x_i, x_j, x_k \in X^n; i, j, k \in N\} \quad (9)$$

To solve the problems posed, we introduce a mapping of the set  $X^n$  into the set  $Y^n$ :

$$y = r(S, x) = (S - x)x, \quad \forall x \in X^n. \quad (10)$$

Based on mapping (10), we have that

$$Y^n = \{y_1, y_2, \dots, y_n \leftrightarrow r(S, x_i) = y_i, x_i \in X^n, i = 1, 2, \dots, n\} \quad (11)$$

Let there be elements among the set  $Y^n$  such that the identity holds:

$$y_i = y_j, i \neq j; i, j \in N. \quad (12)$$

Certificate  $S$  allows you to find a set of  $X_2 = \{x_i, x_j\}$ , consisting of pairs of elements of the original set  $X^n$ , based on formulas (3) and (4). Then.

**Lemma 1.** Let the certificate  $S$  belong to the range  $[S_{min}^2, S_{max}^2]$  and identity (12) holds for set (11). Then problem 1 is solved.

Proof. The first condition shows the existence of a subset  $X_2$  from Theorem 1 that satisfies the certificate  $S$ . To construct vectors  $e$  from identity (12), we have that  $y_i = r(S, x_i) = (S - x_i)x_i = x_jx_i$ , assuming that  $x_j = S - x_i$ . On the other hand,  $y_j = r(S, x_j) = (S - x_j)x_j = x_ix_j$ , similarly assuming that  $x_i = S - x_j$ . In fact, the values  $x_i, x_j$  are the roots of the quadratic equation  $x^2 - Sx + c = 0$ . According to the Vieta theorem,  $c = x_ix_j$ . Thus, we get  $y_i = y_j = x_ix_j$ . The latter means that the fulfillment of identity (12). Then there are elements  $x_i$  and  $x_j$  such that  $x_i + x_j = S$ , as was to be shown.

We introduce the quantity

$$S(x_k) = S - x_k, \quad \forall x_k \in X^n. \quad (13)$$

**Lemma 2a.** Let the certificate  $S$  belong to the range  $[S_{min}^3, S_{max}^3]$  and for some element  $x_k \in X^n$  and considering formula (13), identity (12) holds for  $i \neq j \neq k; i, j, k \in N$ . Then problem 2 is solved.

Proof. The first condition shows the existence of a subset  $X_3$  from Theorem 1 that satisfies the certificate  $S$ . To construct vectors  $e$  from identity (12), taking into account formula (10), we have  $r(S(x_k), x_i) = (S(x_k) - x_i)x_i = x_jx_i$ , assuming that  $S(x_k) - x_i = x_j$ . On the other hand,  $r(S(x_k), x_j) = (S(x_k) - x_j)x_j = x_ix_j$ , similarly assuming that  $S(x_k) - x_j = x_i$ . The latter means that the conditions of Lemma 1 are satisfied when formula (13) is considered. Then we have that  $x_i + x_j + x_k = S$ , as was to be shown.

The next lemma is based on computational geometry and is of independent scientific interest. It is a well-known fact that problem 2 was reduced to the equivalent problem of belonging of three points to one straight line on the plane. However, problem 2 in this formulation has not been solved so far.

In this case, mapping (10) for a search query with three keywords will be rewritten as:

$$y = r(S, x) = (S - x)xx, \quad \forall x \in X^n \quad (14)$$

and introduce a 3x3 matrix

$$H = \begin{pmatrix} x_i & y_i & 1 \\ x_j & y_j & 1 \\ x_k & y_k & 1 \end{pmatrix} \quad (15)$$

The coordinates  $(x_k, y_k)$  on the plane are calculated by the formulas

$$x_k = (S - (x_i + x_j)), y_k = (S - (x_i + x_j))^2(x_i + x_j), \quad x_i, x_j \in X^n. \quad (16)$$

**Lemma 2b.** Let the certificate  $S$  belong to the range  $[S_{min}^3, S_{max}^3]$  and the determinant  $\Delta$  of the matrix (15) is equal to zero when formulas (16) are taken into account, and some element  $x_k$ , defined by the first formula (16) belongs to the set  $X^n$ ,  $i \neq j \neq k$ ,  $i, j, k \in N$ . Then problem 2 is solvable.

Proof. The first condition shows the existence of a subset  $X_3$  from Theorem 1 that satisfies the certificate  $S$ . The second condition ensures the construction of vectors  $e$  based on the well-known result [5] on the membership of three points  $(x_i, y_i)$ ,  $(x_j, y_j)$ ,  $(x_k, y_k)$  one straight line lying on the plane  $(x, y)$ . Replacing the third coordinate  $(x_k, y_k)$  with variables (16) completes the proof of the lemma.

## 4. Search algorithms

Based on these lemmas, we formulate search algorithms. Search algorithm 1 for the first problem.

Step 1. Input of initial data: set  $X^n$ ,  $n$ ,  $S$ .

Step 2. Formation of the set  $Y^n$  based on the mapping(4).

Step 3. Checking the identity (5) and forming a subset  $X_2 = \{ r(S, x_i) - r(S, x_j) = 0 : i \neq j; x_i, x_j \in X^n ; i, j \in N \}$ .

Step 4. Output a subset  $X_2$ .

Working time of the search algorithm  $T = O(n)$ , required memory  $M = O(n)$  to form a set  $Y^n$ .

For a tabular 2-sum, the full search time is  $T = O(n^2)$ , and in the case of sorting  $T = O(n \log n)$ .

Remark 1. The maximum number of pairs in the set  $Y^n$  will be  $m = \lfloor n/2 \rfloor$ . Thus, the number of pairs in  $Y^n$  it can vary from 1 to  $n/2$ .

Especially note that this algorithm allows you to find all subsets of  $X_2$  with a small modification.

*Example 2.* The set is given  $X^7 = \{2, 1, 6, 4, 3, 5, 3\}$ . dimension  $n=7$ . It is required to find out whether a subset exists  $X_2 = \{x_i, x_j\}$ , the sum of these elements from the set  $X^7$  equal to  $S=6$ . Here  $S \in [S_{min}^2, S_{max}^2] = [3, 11]$ . Initially, based on the mapping  $r(S, x)$  (mapping(8)) we will transform the set  $X^7$  into the set  $Y^7 = \{8, 5, 0, 8, 9, 5, 9\}$ . Next, to find a subset of  $X_2$  we use the identity(12):  $y_i = y_j$ ,  $y_i, y_j \in Y^7$ ,  $i, j \in N = \{1, 2, 3, 4, 5, 6, 7\}$ . Then we get  $X_2 = (2, 4)$ ,  $X_2 = (1, 5)$ ,  $X_2 = (3, 3)$ .

The search algorithm for the second problem.

Step 1. Input of initial data: set  $X^n$ ,  $n$ ,  $S$ .

Step 2. Calculating the value of  $S(x_k) = S - x_k$  for some element  $x_k \in X^n$ .

Step 3. Formation of the set  $Y^n$  based on the mapping (4) taking into account  $S(x_k)$

Step 4. Forming a subset  $X_3 = \{ r(S(x_k), x_i) - r(S(x_k), x_j) = 0 \text{ in case of } i \neq j \neq k; x_i, x_j, x_k \in X^n ; i, j, k \in N \}$ .

Step 5. Output a subset  $X_3$ .

The remark given in the search algorithm 1 allows you to determine the search time  $T = O((n - 2m)2m)$ . Here  $m$  is the number of pairs in the set  $Y^n$ ,  $n - 2m$  – the number of remaining indexes without taking into account the used index  $k$ . It is not difficult to show that with the tendency of  $m \rightarrow \lfloor \frac{n-1}{2} \rfloor$  the search time varies within  $O(n) \leq T \leq O((n - 2m)2m)$ . So, the running time of the

algorithm is  $T = O(n^2/2)$ .

Search time for tabular 3-sum  $T=O(n^2)$ .

*Example 3.* Given a set  $X^9 = \{17,43,38,14,20,10,36,47\}$  dimension  $n=9$ . It is required to find out whether there is a subset  $X_3 = \{x_i, x_j, x_k\}$ , the sum of these elements from the set  $X^9$  equal to  $S=100$ . Here  $S \in [S_{min}^3, S_{max}^3] = [51,126]$ . First we choose an arbitrary element  $x_k = x_6=10$ . Find  $S(x_k)$  based on the formula (13)  $S(x_6)= S - x_6 = 100 - 10 = 90$ . Now let's use the mapping (10) ( $r(S, x)$ ) from the first part of the work, we define the set  $Y^9$  for the value  $S(x_k)$ . Next, we apply the identity (12):  $y_2=y_9, S=x_2 + x_9 + x_6 = 43 + 47 + 10 = 100$ .

The search algorithm 2b for the second problem.

Step 1. Input of initial data: set  $X^n, n, S$ .

Step 2. Forming the matrix H.

Step 3. Checking the condition  $\Delta = |H| = 0$ .

Step 4. Checking the membership of the calculated element  $x_k = (S - (x_i + x_j))$  to the set  $X^n$ .

Step 5. Output a subset  $X_3$ .

The operating time of the algorithm varies within  $O(n) \leq T < O(n^2)$ , required memory  $M = O(n)$ .

Note 2. From all combinations of  $n^2$  determinants  $|H| \neq 0$  and  $|H| = 0$  those for which  $\Delta=0$  are selected and among them all elements  $x_k$ , belonging to the original set  $X^n$ , are selected, the number of such elements will be no more than  $n-2$ .

*Example 4.* Given a set  $X^9 = \{2,1,6,4,3,5,3,9,7\}$  of dimension  $n=9$ . It is required to find out whether there is a subset  $X_3 = \{x_i, x_j, x_k\}$ , the sum of these elements from the set  $X^9$  equal to  $S=15$ . Here  $S \in [S_{min}^3, S_{max}^3] = [6,22]$ . If  $x_2 = 1, x_6 = 5$ , and  $x_k$  is determined by the formula (16),  $x_8 = 9$ . Then for these elements it is satisfied that the determinant  $\Delta = |H| = 0, x_8 \in X^9$ , subset  $X_3 = \{x_2, x_6, x_8\}$ . It is not difficult to obtain other subsets, in particular  $X_3 = \{x_1, x_3, x_9\}$ .

## 5. Discussion of the results and Conclusion

There are a lot of search algorithms in the scientific literature based on exponential algorithms [3,4]. The search time and required memory are  $O(2^{n/2})$  и  $O(2^{n/4})$  respectively. The use of these algorithms is difficult due to finding  $2^{n/2}$  subsets. Tabular search methods are based on the construction of tables. The proposed theorems actually do not depend on the length of the search query and require finding only one subset of the sum of subsets problem. Lemmas and examples 2-4 show the solution of the tasks regardless of the combination (7). The developed search algorithms with two and three keywords are the most effective compared to tabular methods. In particular, for  $m=1$  and  $m=n-1$ , the traditional search methods follow from the theorems: sequential search and pattern search (mask search). Lemmas 1 and 2 allow us to construct a whole family of algorithms for sampling unstructured data for a "short" search query with  $m$  keywords and a "long" search query with  $n-m$  keywords.

The analysis shows that search algorithms significantly reduce the search time for unstructured data, and also reduce the hardware requirements for the power of computers, servers and other computing devices used. The developed algorithms for searching unstructured data are completely different from the well-known algorithms based on arrays, trees, index arrays, index trees.

## 6. References

- [1] C. J. van Rijsbergen."Information Retrieval". Dept. of Computer Science. University of Glasgow,1979.
- [2] A. Adamanskij. Obzor metodov i algoritmov polnotekstovogo poiska. -Novosibirsk,

Novosibirskij gosudarstvennyj universitet, 2018, p.26.

[3] E. Horowitz, S. Sanni. Computing Partitions with Application to the Knapsack Problem //Journal of the ACM(JACM), 1974, T21, pp.277-292.

[4] R. Schroepel, A. Shamir A  $T=O(2^{n/2})$ ,  $S=O(2^{n/4})$  Algorithm for Certain NP-Complete Problem // SIAM Journal on Computing, 1981, Vol.10, № 3, pp.456-464.

[5] A. Korn, M. Korn. Mathematical Handbook.-New York, McGraw-Hill Company, 1968.-832p.

[6] Y. Lifshiz Tochnye algoritmy i otkrytye problemy. //yura@logic.pdmi.ras.ru.

[7] S.S. Ahtanova Algoritmy poiska dannyh //Sovremennye tekhnologii, 2007, №3, pp.11-17.

[8] V.S. Simakov, D.M. Tolkachev Metody i algoritmy poiska informacii v Internete. –M.: Globus,2017, p.332.

[9] D. Knut Iskusstvo programirovaniya. Sortirovka i poisk. T.3-M.: Vil'yams, 2000, p.844.

[10] Dzh. Makkonell Analiz algoritmov –M.: Tekhnosfera, 2002, p.304.

[11] V.A. Urvacheva Obzor metodov informacionngo poiska. // Vestnik TI im. A.P. CHEkhova, 2016,p.1-7

[12] I.A. Selivanova, Zykina M.E. Obzor algoritmov resheniya zadachi o nahozhdenii summy elementov podmnozhestva //Vestnik Ural'skogo instituta ekonomiki, upravleniya i prava, №4, 2016, p.86-92.

# Algorithm for Solving Pronominal Anaphora in the Kazakh Language

Gulzhamal Kalman<sup>1</sup>, Madina A. Sambetbayeva<sup>1</sup>, and Yerzhan S. Zhumabay<sup>2</sup>

<sup>1</sup> L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

<sup>2</sup> Astana International University, Astana, Kazakhstan

## Abstract

One of the urgent tasks facing computer linguistics is to extract information about various objects in textual documents: people, organizations, events, places, etc., as well as the relationships between them. Each information object (entity) corresponds to a specific concept/relationship of the subject domain and has a specific structure. This issue in natural language processing will be related to the notion of referentiality. In this article, we will consider the ways of solving referential relations in the Kazakh language in the course of the study, find the pair "antecedent-anaphora" using algorithms for the classification of morphological, syntactic, and semantic features of pronoun anaphora, Support Vector Machine method, and decision tree.

As a set of training and test data, we find the pair "antecedent-anaphora" in different types of text, we used annotated corpora: National Corpus of Kazakh language (NCL). and calculate the distance between words. We also assess how semantic features, in particular semantic roles, affect the performance of anaphora decisions in Kazakh. Considering the peculiarities of Kazakh nouns, classification algorithms, support vector machines, and the decision tree method were used, using the method of formal analysis of various texts. The most common classification, reference names, and proper names among anaphoric names in the Kazakh language, the actual indicators of anaphoric connection were calculated, knowing the number of "antecedent-anaphora" pairs from the texts collected during the study, the number of "antecedent-anaphora" pairs was calculated.

## Keywords

Anaphora, machine learning, support vector, tree of solutions, semantic roles

## 1. Introduction

Solving anaphora is one of the major problems of natural language processing. Anaphora and conference methods include machine translation, information retrieval, information retrieval, and more. systems. The anaphora-solving problem has been widely studied in English and other European languages.

Anaphora is a means of referring to a particular object (or objects) in discourse, the reference is called anaphora, and the object (or objects) referred to is its REFERENCE or ANTECEDENTS.

The problem of solving the anaphora for the Kazakh language has just begun, (Zhumabay et al. 2022) the first research work on solving the reference in the Kazakh language, where the authors (Kibrik 1999) tried to integrate it into the proposed method (Garanina et al 2018a,2018b) and proposed a model for solving reference relations in a multilingual system.

In this paper, we conduct the following research in solving the anaphora of pronouns in the Kazakh language. First: research based on machine learning using the annotated case.

Second: to study how semantic features, in particular semantic roles, affect the solution of anaphora in the Kazakh language.

We solve the pronoun anaphora only by comparing statistical and inductive methods of the solution tree based on a vector machine for the classification of noun reference nouns and proper nouns (see Table 1 for the complete set of pronouns treated). As the training and testing database, we used annotated corpora: National Corpus of Kazakh language (NCL).



**Table 1**

Types of anaphora treated

Category of pronouns	Anaphoric item
Personal pronouns	Ол, олар
Demonstrative pronouns	Ана, анау, әне, бұл, осымына, міне, сол, сонау
Reflexive pronouns	өзім, өзің, өзіңіз, өзі

## 2. Literature review

Numerous studies have been conducted on the solution of pronoun anaphora. Important works include Hobbes (1978,1976), Lappin and Liss (1994), Kennedy and Boguraev (1996), Mitkov (1998,2001,2007) Tetreo (2003,2004) and Trouille (2002) applies.

In the next work Yu-Hsiang Lin and Tyne Liang (2004), a method was used to solve the pronoun anaphora using UMLS ontology and SA / AO (subject-action / action-object) models, where an F-value of 92% was obtained.

Saoussen Mathlouthi Bouzid et al 2021, proposed an SVM method of self-learning based on a set of patterns and linguistic criteria to solve an ellipse with a pronoun in the Arabic language. For the solution step, they developed a new hybrid method that combines an enhanced learning method with Word implementation templates. The learning enhancement method used an adapted version of the Q-learning algorithm to find the optimal combination of capabilities. It uses a set of morphological and syntactic features. The Word-based approach uses word display templates to test the semantic validity of candidates. Evaluation of the identification method gives an accuracy of 99.23% for pronominal anaphoras and 94.33% for ellipses.

In the next study, (Senapati et al 2020) Support Vector Machine (SVM) was the most effective of several methods for solving the pronoun anaphora using the machine learning method for the first time for the Nepali language.

(Abaci et al 2022) In this work the solution of the pronoun anaphora in the Turkish language is provided, the algorithms used here are J48, Voted Perceptron, SVM (support vector machine).

Naive Bayes and k-nearest neighbors.

The next study describes a quantitative analysis performed to compare two different methods on the task of pronoun resolution for Swedish. They are Mitkov's algorithm and SVM (support vector machine), a result of SVM-based methods significantly outperformed the implementation of Mitkov's algorithm (Ahlenius. 2020).

(Ardiyani et al, 2020) They are using the SVM (support vector machine) method, and solved the pronoun anaphora in the Holy Quran, based on the evaluation results, the system can find the front row of an anaphor with the best accuracy value of 88.08%.

(Pak, A. et al, 2019) in this work is to identify the target of a pronoun within a text passage. In the present work, they are considered the classical problem of resolving anaphoric connections as exemplified by the gender-balanced corpus of texts. Using the BERT Model, they tested the distance between noun groups and candidate pronouns in lexemes and the sign of a linguistic determinant.

(Lata, K. et al 2021) In this work, an overview of the methods of solving the anaphoric relationship is presented, and it is possible to get answers to the problems that arise in the research. The methods used for different languages and their features are described. Also, the features used in the AR system and the features of their use in different languages are mentioned.

(Kim, Y., et al 2021) the following paper presents approaches to solving zero anaphora in Korean. Taking into account the linguistic features of the zero anaphora, he effectively used the bidirectional encoder approach. BERT can encode deep bidirectional by word and sentence level; can encode phrase-level information on the lower layers; encode a rich hierarchy of language information such as syntactic signs on the middle layers and semantic signs on the upper layers; as a result, the solution of zero anaphora in Korean was solved with an accuracy of 79.6%.

The above works [1-23]: methods of solving anaphoric relations in English, Russian, Korean, and

Turkish are shown. the lack of research in the Kazakh language in solving anaphoric relations allows us to update the development of a model for solving reference relations in this area. Therefore, such a study in the Kazakh language helps to solve the problem associated with the need for automatic text processing.

### 3. Materials and Methods

#### 3.1. The pronoun anaphora

The most common type of anaphora is the pronoun anaphora. This type of anaphora includes the third type of pronouns: personal pronouns, demonstrative pronouns, reflexive pronouns. From the following examples, we can see the anaphoric function of the reference classificatory noun.

For example: Труба түбіндегі жапырық тас үй – **мехцех**. **Бұл** - әншейін келешегіне қарай қойылған ат, әйтпесе нобайы түзу бір механизм жоқ.

This syntactically complex unit that connects the first sentence and the second sentence is an anaphoric relationship, where the word **мехцех** in the first sentence is repeated with the pronoun **бұл** in the second sentence.

**Елжас** өткен айда **Бразилияда** болды. **Ол сол** елден саған сыйлық алды.

In the first sentence “**Елжас**” is repeated in the third person form of the personal pronoun “**ол**” in the second sentence, and “**Бразилияда**” is repeated with the “**сол**” demonstrative pronoun they are called anaphoric relation.

Figure 1 shows the application of the decision tree method to the second example.

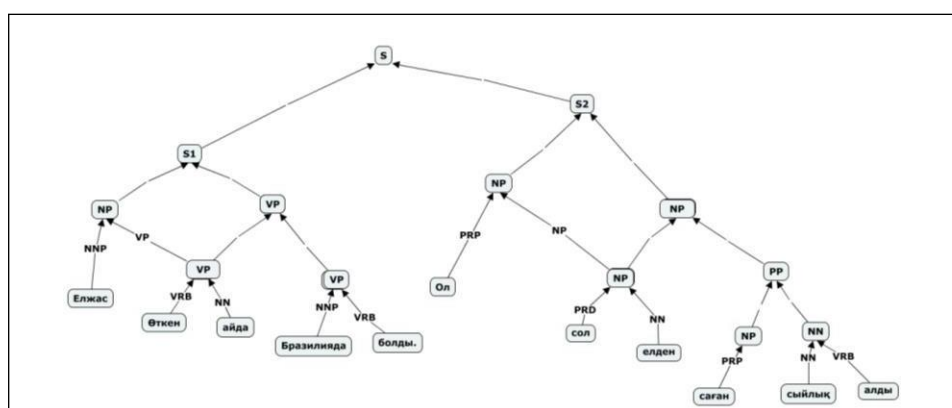


Figure 1: Parse trees for the sentence “Елжас өткен айда **Бразилияда** болды. **Ол сол** елден саған сыйлық алды”

#### 3.2. Solving the pronoun anaphora

Solving anaphora - the problem of determining the correct pair of "antecedent-anaphora" in our study we consider only the classificatory, reference, and personal pronouns.

The antecedent must match the anaphora in number and spelling. The distance between the anaphora and the antecedent in a word must not exceed a given value, depending on the text. We consider the anaphora problem as a classification problem and solve it using machine learning techniques. The following symbols were used for classification:

Morphological and syntactic features:

1. name, number, and classification of anaphora
2. number, distribution of the antecedent;
3. juxtaposition of an anaphora and an antecedent between an animate and inanimate (if there are nouns);
4. the number of sentences between the anaphora and the antecedent;

5. the number of words between the anaphora and the antecedent;
6. the number of nouns between the anaphora and the antecedent; Semantic features:
7. semantic roles of anaphora;
8. semantic roles of the antecedent;

In the morphological analysis, anaphora corresponds to the antecedent, i.e. the correspondence of number, surname, classification, animate and inanimate (if nouns) in classifications 1-3, signs in classifications 4-6 give information about the scale of distance between anaphora and antecedent.

Spacing in words. For each candidate, the distance to the pronoun in words is calculated. Depending on this distance the vector is filled in units. There are three gaps to attach them to the vector:

- count from 10 words; vector [1, 0, 0];
- 10 to 30 words; vector [0, 1, 0];
- more than 30 words; vector [0, 0, 1].

Only one vector with a description in vector form can correspond to a candidate.

### 3.3. Algorithm of creating a training database

1. Find the pair "antecedent-anaphora" in the set of texts.
2. Find all nouns and nouns between the anaphora and the antecedent. The anaphora must match their number and quantity. The search area is limited to a predetermined number of words.
3. All nouns and pronouns found in step 2 are incorrect hypothetical antecedents.
4. If the correct antecedent is not in the search area, it will not be included in the textbook.
5. Follow steps 1-4 for each processed example.

We used the vector machine REPTree (SVM) method [Chang and Lin, 2014] and the decision tree method [Waikato University, 2014] to train and classify correct/incorrect pairs. Below is an algorithm for finding antecedent-anaphora.

```

Input: P training examples( $a_i, x_j$ )
          number of iterations T
init:  $\vec{q} \leftarrow \vec{0}$ ;
fort  $\leftarrow 1$  to T,  $i \leftarrow 1$  to P do
 $\hat{x}_i \leftarrow \operatorname{argmax}_{c \in \operatorname{ant}(a_i)} F(c, a_i) \cdot \vec{q}$ 
If  $\hat{x}_i \neq x_i$  then
 $\vec{q} = \vec{q} + f(a_i, x_j)$ 
End
End
Output: pair "antecedent-anaphora" ->  $\vec{q}$ 
    
```

**Algorithm 1:** algorithm for finding the "antecedent anaphora". F is the feature extraction function,  $a_i$  is the anaphoric expression,  $x_j$  is the true antecedent

### 3.4. Anaphora solution algorithm

1. Find the first anaphora without an antecedent. If no anaphora is found, the algorithm terminates.
2. Find all nouns or pronouns that are anaphors between the anaphora and the antecedent. The anaphora must match their number and quantity. The search area is limited to a predetermined number of words.
3. Add them to the set of hypothetical antecedents.
4. Match the semantic roles of the antecedent to each noun in the set of hypothetical antecedents.
5. Calculate the probability that each hypothetical antecedent will be the correct antecedent using the classification method.
6. Choose the most likely antecedent and connect it to the appropriate anaphora. Skip stepping

7. The search area for the hypothetical antecedent is limited to step 2 since the anaphora usually represents the nearest hypothetical antecedent. This value was calculated in our experiments.

#### 4. Corpus analysis and results

A large corpus was created from the Kazakh language, where personal pronouns are tagged with their antecedents. Over 23,563 pronouns are tagged with their antecedent information. Also, the antecedents are maintained as an ontological list of concepts. NKL is designed to be a large- scale corpus containing over 12 million words. <https://qazcorpus.kz>. Table 2 gives a quantitative measure of this corpus. Note that the personal pronouns are tagged, and relative and demonstrative pronouns represent only about 12% of the total number of pronouns and most of them are non- anaphoric. Although there are some cataphoric cases in the corpus, they are only 55 (0.24%) of the antecedents. Among the total number of pronouns, there are 56.6% of pronouns have antecedents.

**Table 2**  
 Quantitative measure from the NKL corpus

Measure	Count	%
Personal pronouns	23,563	80.11%
Relative pronouns	3753	14.20%
Demonstrative pronouns	1116	3.84%
Total	28432	
Anaphor antecedent with existing	15317	51.60%
Anaphor antecedent with hidden	12765	49.10%
Cataphors	45	0.36%
Total	28127	

Table 3 gives a statistic of this corpus and describes the distribution of the pronouns.

**Table 3**  
 Statistics of the various types of pronouns in the NKL corpus

Pronoun types	Count	%
Person		
1st person	4560	17.85%
2nd person	5645	24.72%
3rd person	12766	60.38%
Gender		
Masculine	13876	56.62%
Feminine	2533	10.62%
Others	3145	14.76%
Number		
Singular	7045	22.48%
Dual	458	3.51%
Plural	13424	63.01%

Table 4 shows the distances between the anaphor and its antecedent in terms of verses, words, and segments. Among the pronouns which have antecedents, most of them are within the same verse as anaphor 52.6%. In the second place, 17.5% of the antecedents are found within 9 verses from the anaphor. In terms of word distance, 45.16% of the antecedents are found within 12 words away from its anaphor, and only about 1% of antecedents are found in the last preceding word of the anaphor.

**Table 4**  
Distances between anaphor and antecedent

Distance	Occurrence		
	Verse	Word	Segment
0	6793	120	63
1 to 9	5725	6855	5327
10 to 19	378	2506	2356
20 to 29	56	1302	989
30 to 49	18	1023	678
50 to 100	8	958	549
100 to 200	0	235	456
200 to the end (1060)	0	154	312

#### 4.1. Discussion of the results of the study of the algorithm for solving the pronoun anaphora based on machine learning

During the study, all types of Pronominal anaphora relations were analyzed and a theoretical study of methods for their solution was carried out. Works [1–23] consider the referential choice only between full name groups and anaphoric pronouns, in addition, studies were conducted on Pronominal anaphora relations (the English, and Russian languages).

In our study, we used the vector machine REPTree (SVM) method [Chang and Lin, 2014] and the decision tree method [Waikato University, 2014] to train and classify correct/incorrect pairs.

This makes it possible to design a solution for pronominal anaphora research in the field of computational linguistics and automatic text processing.

The peculiarity of the Algorithm for solving pronominal anaphoric communication in the Kazakh language we calculated the optimal distance for each data set covering 90%. An experimental study of the resolution of Pronominal anaphora relations was carried out. As part of the research work, we analyzed the texts of more than 1000 different topics from the corpora, in the quantitative evaluation of the results of the study we obtained the following data. In the first set of texts, 170 texts were examined, and 367 pairs of "ancient anaphora" were found, in the second set of texts 200 texts were examined, and 467 pairs of "ancient anaphora" were found.

Advancing this study could lead to the development of an intelligent information resource on modern methods of automatic text processing. The information resource to be designed would provide convenient meaningful access to information about a given method of automatic text processing, pre-trained models, test data, marked text corpora, and other information resources on this topic.

## 5. Conclusions

In this work, we tried to find pronominal anaphora in the Kazakh language using machine learning, the three most frequently used groups of pronominal anaphora in the Kazakh language were compared with examples, and we used data from the national corpus of the Kazakh language as research data.

The vector machine learning method and the decision tree method were used in the work. In both methods, a very good result was obtained in solving anaphora in the Kazakh language. The results of the study conducted in the study data are shown in tables 2, 3, and 4.

We believe that this research will make a great contribution to the study of the Kazakh language. In the future, it is planned to solve coreference, anaphora, and cataphora relationships in the Kazakh language.

## 6. Acknowledgements

The work was supported by a grant for financing scientific, scientific and technical projects for 2022-2024. Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (AP14972834).

## 7. References

- [1] Zhumabay, Y., Kalman, G., Sambetbayeva, M., Yerimbetova, A., Ayapbergenova, A., & Bizhanova, A. Building a model for resolving referential relations in a multilingual system. *Eastern-European Journal of Enterprise Technologies*, (2022). 2(2 (116), 27–35. <https://doi.org/10.15587/1729-4061.2022.255786>.
- [2] Garanina, N. O., Sidorova, E. A., Seryi, A. S. Multiagent Approach to Coreference Resolution Based on the Multifactor Similarity in Ontology Population. *Programming and Computer Software*, 2018. 44 (1), 23–34. doi: <https://doi.org/10.1134/s0361768818010036>.
- [3] Sidorova, E. A., Garanina, N. O., Kononenko, I. S. Mnogomestnye ontologicheskie otnosheniya v zadache razresheniya koreferentsii. *Shestnadsataya natsional'naya konferentsiya poiskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2018*.
- [4] Kibrik, A. A. Reference and Working Memory. *Current Issues in Linguistic Theory*, 1999, 29. doi: <https://doi.org/10.1075/cilt.176.04kib>.
- [5] Hobbs JR, Resolving pronoun references. *Lingua* 44: 1978, pp 339-352.
- [6] Hobbs JR Pronoun resolution. Technical Report, Department of Computer Science, City College, City University of New York 76-1, 1976.
- [7] Lappin S, Leass HJ) An Algorithm for Pronominal Anaphoric Resolution. *Computational linguistics* 20: 1994, pp 535-561.
- [8] Kennedy C, Boguarev B Anaphora for everyone: pronominal anaphora resolution without a parser. 16th International Conference on Computational Linguistics (COLING'96) Denmark, 1996pp 113-118.
- [9] Mitkov R, Robust Pronoun Resolution with Limited Knowledge. 18th International Conference on Computational Linguistics, Canada, 1998.
- [10] Mitkov R Outstanding Issues in Anaphora Resolution. *Lecture Notes in Computer Science* 2004: 2001, pp 110-125.
- [11] Mitkov R, Anaphora resolution: to what extent does it help NLP applications? *Anaphora: Analysis, Algorithms and Applications*. SpringerVerlag, Berlin Heidelberg, 2007, pp,179-190.
- [12] Tetreault J, Allen J, An Empirical Evaluation of Pronoun Resolution and Clausal Structure. 2003 International Symposium on Reference Resolution, 2003.
- [13] Tetreault J, Allen J, Dialogue Structure and Pronoun Resolution. *Lecture Notes in Computer Science* 1793: 2004, pp 515-525.
- [14] Trouilleux F, Insertions et interprétations des expressions pronominales. *Actes de l'Atelier Chaînes de référence et résolveurs d'anaphores*. TALN 2002 Nancy, 2002.
- [15] Yu-Hsiang Lin and Tyne Liang. Pronominal and Sortal Anaphora Resolution for Biomedical Literature. In *Proceedings of the 16th Conference on Computational Linguistics and Speech Processing*, 2004, pp 101–109.
- [16] Saoussen Mathlouthi Bouzid, Chiraz Ben Othmane. Zribi.2021, Efficient Learning Approach for Pronominal Anaphora and Ellipsis Identification and Resolution in Arabic Texts *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- [17] Senapati, A., Poudyal, A., Adhikary, P., Kaushar, S., Mahajan, A., & Nath Saha, B. A Machine Learning Approach to Anaphora Resolution in Nepali Language. 2020 International Conference on Computational Performance Evaluation (ComPE), 2020.
- [18] H. Abaci, M. Eminagaoglu, I. Gor, and Y. Kilicaslan, "Pronoun Resolution for Tweets in Turkish", *Int J Intell Syst Appl Eng*, 2022, vol. 10, no. 2, pp. 207–215.
- [19] C Ahlenius. Automatic Pronoun Resolution for Swedish. - 2020 - [diva-portal.org](https://diva-portal.org).
- [20] Ardiyani, A. D., Bijaksana, M. A., & Huda, A. F. Anaphora Resolution on Al-Quran with Indonesian Translation. *Indonesia Journal on Computing (Indo-JC)*, 5(2), 2020, pp 99- 106. <https://doi.org/10.34818/INDOJC.2020.5.2.496>.

[21] Pak, A. A., Amirzhan, S., & Ziyaden, A. A. The Resolution of Gender Anaphora Reference with the Help of Kernel Trick Mechanisms. In IOP Conference Series: Materials Science and Engineering 2019, Vol. 630, No. 1, pp 12-29. IOP Publishing. DOI: 10.1088/1757-899X/630/1/012029.

[22] Lata, K., Singh, P., & Dutta, K. A comprehensive review on feature set used for anaphora resolution. Artificial Intelligence Review, 2021, 54(4), pp 2917-3006. <https://doi.org/10.1007/s10462-020-09917-3>.

[23] Kim, Y., Ra, D., & Lim, S. Zero-anaphora resolution in Korean based on deep language representation model: BERT. ETRI Journal, 2021, 43(2), pp 299-312. <https://doi.org/10.4218/etrij.2019-0441>.

# Development of a Career Guidance System Using Machine Learning Algorithms

Dana Zhussupova<sup>1</sup>, and Aidos Sarsembayev<sup>2</sup>

<sup>1</sup> International Information Technology University, Almaty, Kazakhstan

## Abstract

The determining of personality traits is significant in career guidance. The purpose of this study is to optimize the process of determining the type of personality by using natural language processing methods. Various models were tested, among which the model with the best quality metrics was identified. This work demonstrates the possibilities of artificial intelligence methods for determining the type of personality and developing career guidance.

## Keywords

Career guidance, MBTI, personality types, machine learning, NLP

## 1. Introduction

According to a study by the American Gallup Institute, about 10 years ago there were almost twice as many people who are dissatisfied with their work as those who like it [1]. To be more precise, only 13% of people in the world enjoy their work. They are satisfied with the working conditions, strive for career growth in their company and make every effort to ensure that it flourishes. Most respondents to the annual survey (63%) feel dissatisfied with what they do. The working day for them is a routine, and they treat their duties carelessly. Another 24% of employees, according to researchers, actually hate their job and actively demonstrate it on occasion. In sum, these two categories (dissatisfied and extremely dissatisfied employees) account for 87%. In other words, work is more or less a source of frustration for the vast majority of the world's working population.

But how does it happen that people do not do what they would like to do? Why do they go to unloved specialties? Everything is simple. They do not understand themselves; they cannot see and recognize their own abilities. How to fix it? Try to help the majority understand the answer to the question "Who am I and what do I want to do in this life?".

The exploration of a set of different parameters, such as learning ability, character, sociability, logic, preferences are important in order to more accurately determine the work area where the needs and abilities of a person will best suit him.

### 1. MBTI Topology

The Myers Briggs Type Indicator (MBTI) is a system of personality types that is popular in Europe and America [2]. The MBTI is based on the idea of four fundamental pairs of preferences that determine the psychological type of a person. This complex typology classifies every person into 16 distinct personality types across 4 axes:

- Introversion (I) - Extroversion (E);
- Intuition (N) - Sensing (S);
- Thinking (T) - Feeling (F);
- Judging (J) - Perceiving (P);



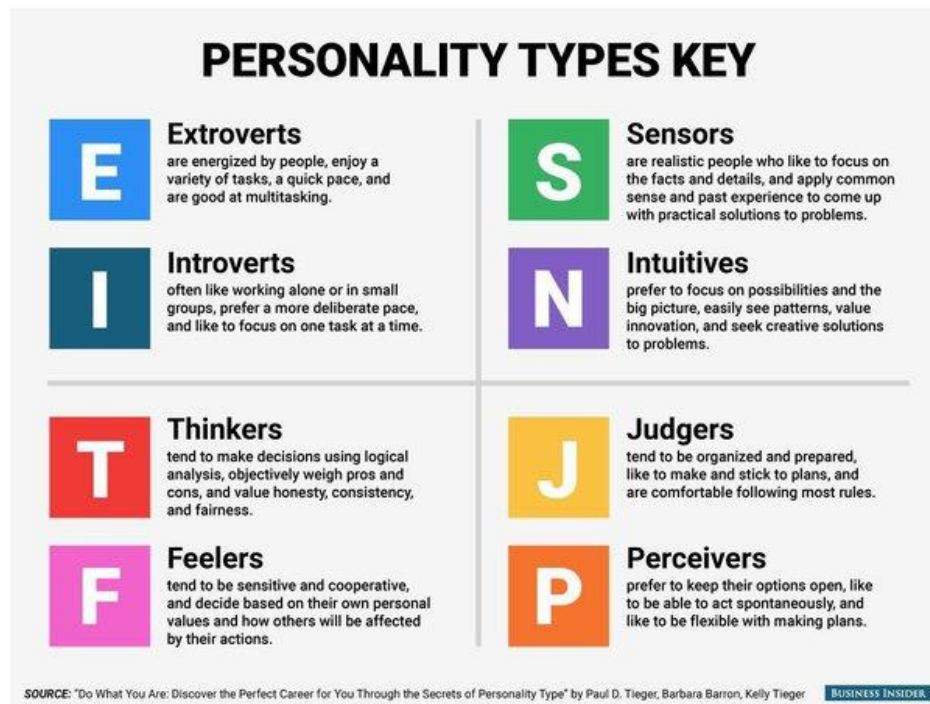


Figure 1: Personality types keys [2]

## 2. Practical significance

Consider a small example from real life about a story of a tax lawyer who, despite her best efforts, couldn't climb the career ladder. "I'm a good specialist, but others are always being promoted, not me. And I go to advanced training courses, and I read developing literature, but someone is always ahead of me". She is trying to understand why she is not able to become the best in her field despite all her efforts. Why haven't you been promoted for so many years?

It is known that her sociotype is Dreiser (ISFJ) and the answer becomes obvious.

The fact is that knowing the sociotype of a person, we automatically understand which functions of his psyche are strong (that is, which it makes sense to rely on both in work and in his personal life) and which are weak.

A function called structural logic is responsible for the ability to jurisprudence. If a person has a strong structural logic by nature, he can perfectly analyze, find cause-and-effect relationships, make diagrams, calculations, operate with arguments and facts. That is, everything that makes a person a good lawyer.

So - Dreiser's structural logic is naturally weak. Many others are strong. But the woman chose the job because of her weak function. And what happens when a person chooses a profession based on a weak function?

In order to be able to fulfill his duties, he artificially develops this weak function in himself. And thanks to this, he copes with his duties quite tolerably. That is, if you are a Dreiser, you can still work as a lawyer.

But you will always be second or third. Not the first. The first in the profession will always be a specialist who has chosen a job according to his strong function. So, Robespierre (INTJ) has a strong structural logic by nature. In the world of structures, facts and analytics, he feels like a fish in water. After all, what Dreiser intensively develops in himself, he has from childhood. And Robespierre will always be a better lawyer than Dreiser. He'll beat him in court, he'll get a promotion, he'll make more money. And he will get more pleasure from his work, because in fact he is doing what he was created for.

Therefore, the best way to become a sought-after and highly paid specialist is to choose a job according to the strong functions of your psyche.

## 2. Literature Review

According to a British psychologist Hans Jürgen Eysenck, “personality is the sum-total of the actual or potential behavior-patterns of the organism, as determined by heredity and environment. It originates and develops through the functional interaction of the four main sectors into which these behavior-patterns are organized” [3]. Using factor analysis, he developed a theory of personality traits, from which many other studies later originated. More recent research has already applied machine learning or deep learning to solve the problem.

The authors of [4] in their work describe personality classification experiment by applying k-means clustering machine learning algorithms. Several previous studies have been attempted to predict personality types of human beings automatically by using various machine learning algorithms. However, only few of them have obtained good accuracy results. To classify a person into personality types, they used Jungian Type Inventory. The method consists of three parts: data collection, data preparation, and hyper-parameter tuning. Testing results showed that the k-means model has 107 inertia value, which is a good number for an unsupervised learning model as an interim result. With the result, they divided the data into 16 clusters, which can be considered as personality types.

Kiselev et al. in earlier research have shown that personality traits can be predicted by mining social network data [5]. Their next work describes the social constructivism grounds of machine learning methods in career guidance and broadens understanding role of social networks in psychological researches. The theoretical grounds are empirically confirmed by AUC-ROC measure calculation in career guidance modeling. Implications for career guidance practice will also be presented.

The research of Brandon Cui and Calvin Qi studied various natural language processing techniques in conjunction with machine learning techniques and evaluated their results on classifying someone’s Myers-Briggs personality type based on one of their social media posts [6].

**Table 1**  
 Research on personality type prediction methods

#	Study	Aims & objectives	Techniques	Results	Future Work
1	Mohammad Hossein Amirhosseini, Hassan Kazemian (2020)	Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator	NLTK, XGBoost, Gradient Boosting	86.06% (Accuracy for I-S)	The methodology improved the accuracy of recognising the Intuition (I)–Sensing (S) and Introversion (I)–Extroversion (E) personality categories
2	Brandon Cui, Calvin Qi (2017)	Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction	NLTK, Naive Bayes, SVM, Deep Learning	89.85% (Accuracy for S-N)	Take a look at other mechanisms of representing word embeddings, including char and k-char word embeddings and pretrained GloVe vectors
3	Kiselev et al. (2020)	Career guidance based on machine learning: social networks in professional identity construction	CatBoost	0.68-0.85 (AUC-ROC)	This study finds that nusing machine learning and data of social network profile will guide students to a major can be useful and enjoyable for students.
4	Talasbek et al. (2020)	Personality Classification Experiment by Applying k-Means Clustering	k-Mean Clustering	107 (Overall inertia)	This study finds that using machine learning and data of social network profile will guide students to a major can be useful and enjoyable for students.

While developing a machine learning method for predicting personality type based on mathematical programs, Mohammad Hossein Amirhosseini, Hassan Kazemian used the MBTI type indicator, justifying this by the fact that it is considered one of the most popular and reliable methods at present [7].

### 3. Model algorithms description

Machine Learning is the branch of artificial intelligence that allows computers to improve their performance and learn new approaches without explicit instructions - enables companies to identify patterns in their data and incorporate predictive analytics into their decision making [8]. Data science is mainly about reducing business problems to data problems, the tasks of collecting, understanding, cleaning and formatting data, after which machine learning comes as an add-on [9]. There are three types of machine learning: supervised learning, unsupervised learning (unsupervised or spontaneous), and reinforcement learning:

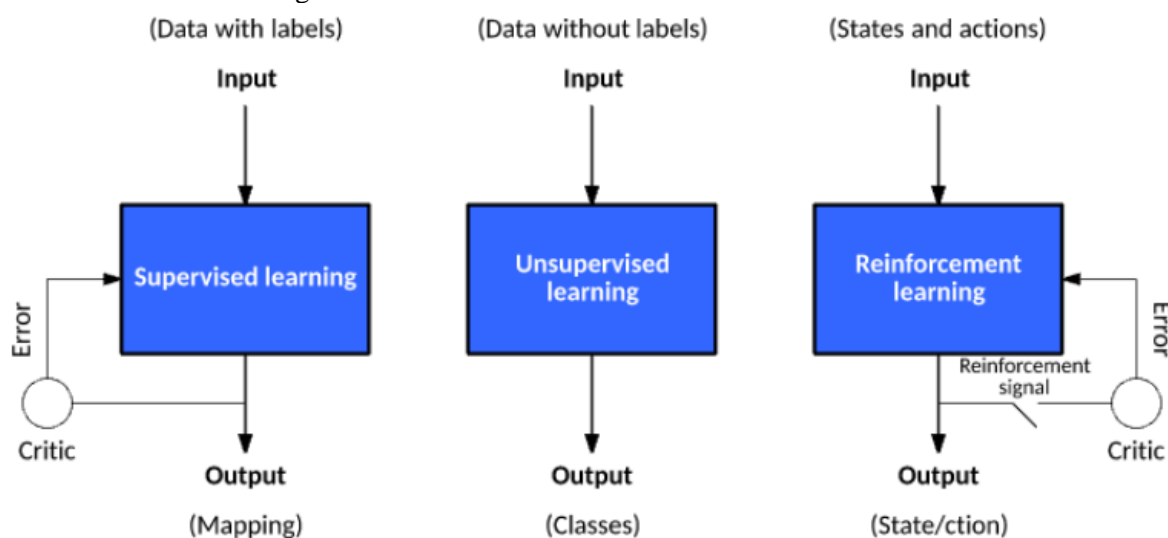


Figure 2: Types of machine learning [8]

1. Supervised learning: The main task of supervised learning is to extract a model from labelled training data that allows predictions of previously unseen or future data to be made. Here, the term "supervised" refers to a subset of patterns in which the desired output signals (labels) are already known [10]. Supervised learning methods when there are discrete marks of class membership are called classification methods. The classification task is a subcategory of supervised machine learning methods, the essence of which is to identify categorical class labels for new instances based on previous observations. A class label is a discrete, unordered value that can be understood as belonging to a group of instances;

2. Unsupervised learning: in reinforcement learning, defining a measure of reward for the individual actions performed by the agent [11]. On the other hand, in unsupervised learning, dealing with untagged data or data with unknown structure. Using unsupervised learning methods includes scouting the data structure to extract meaningful information without the control of a known outcome variable or reward function;

3. Reinforcement learning: The challenge of reinforcement learning is to develop a system (agent) that improves its quality based on interactions with the environment [12]. Since information about the current state of the environment, as a rule, also contains a so-called reward signal, reinforcement learning can be represented as an area related to supervised learning. However, in reinforcement learning, this reverse relationship is not a label or meaning determined once and for all by direct observation, but a measure of how well an action has been judged by the reward function. As the agent interacts with the environment, an exploratory approach, through trial and error or deliberative

planning, can use reinforcement learning to isolate a series of actions that maximize this reward. The several machine learning algorithms to achieve accuracy with normal values in the MBTI dataset: Random Forest, XGBoost, Gradient Descent, Logistic Regression, kNN, SVM.

- Random Forest — the basic theoretical principle of this algorithm is a decision tree [13]. The main task is to make a decision based on the available information. In a simple case, there is only one feature (metric, predictor, regressor) with clearly distinguishable boundaries between classes. The algorithm combines two main ideas: the Bagging Breiman method, and the random subspace method proposed by Tin Kam Ho. The algorithm is used for classification, regression and clustering problems.

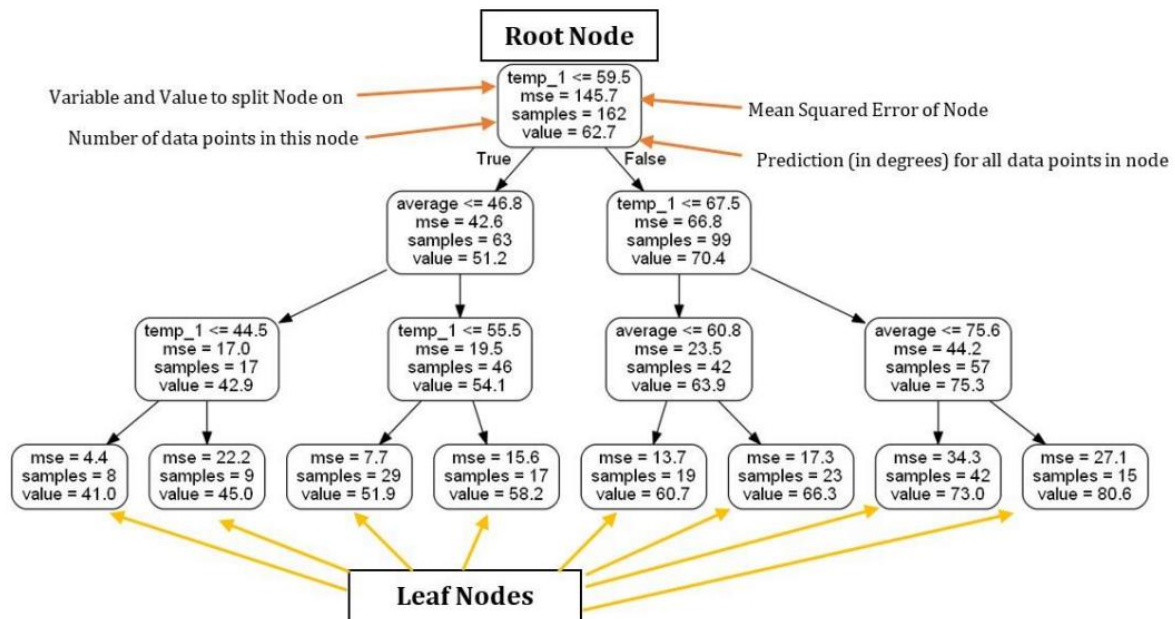


Figure 3: Random Forest algorithm [13]

- XGBoost — is a tree-based machine learning algorithm using a gradient boosting framework [14].

In prediction problems that use unstructured data (like images or text), an artificial neural network is superior to all other algorithms or frameworks.

$$\delta(\omega_1 a_1 + \omega_2 a_2 + \omega_3 a_3 + \dots + \omega_n a_n + b), \tag{1}$$

where  $\delta$  - activation function;  $\omega$  - weights of the neural network;  $a$  - values of the nodes;  $b$  – bias.

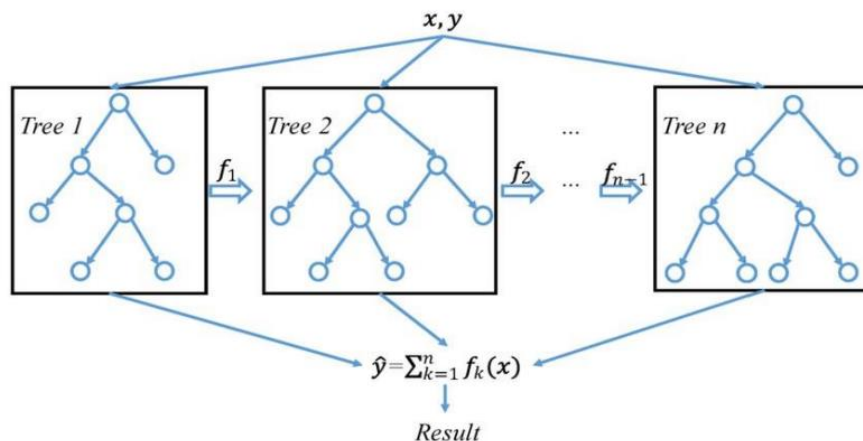
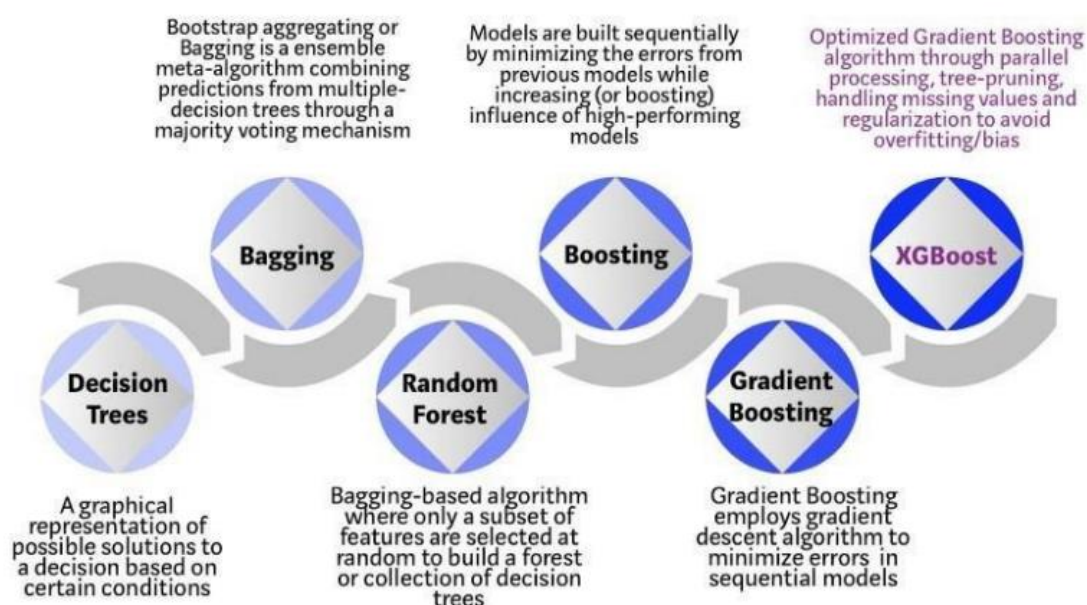
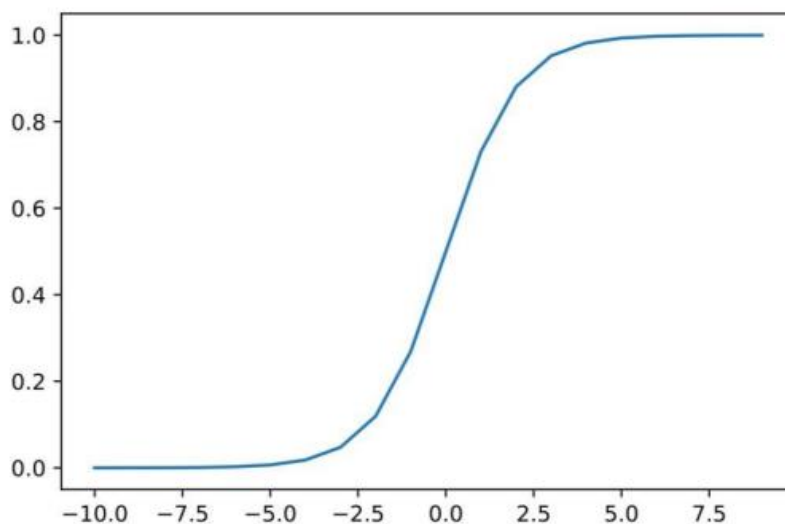


Figure 4: A general architecture of XGBoost algorithm [14]



**Figure 5:** Evolution of XGBoost Algorithm from Decision Trees [15]

- Stochastic gradient descent — is an optimization algorithm used in machine learning applications for the model parameters that correspond to the best fit between predicted and actual output values [16].
- Logistic Regression — is a powerful ML model that makes it possible to split existing data into several classes to predict unseen data with the help of the classification [17]. The heart of this algorithm is the sigmoid function which has S-shapes when plotted.



**Figure 6:** Logistic Regression graph [17]

- K Nearest Neighbors — metric algorithm for automatic object classification or regression. In the case of using the method for classification, the object is assigned to the class that is most common among the  $k$  neighbours of the given element, the classes of which are already known [18]. In the case of using the method for regression, the object is assigned the average value of the  $k$  objects closest to it, the values of which are already known. kNN is a fast and interpretable ML algorithm that can adapt classification and regression too.

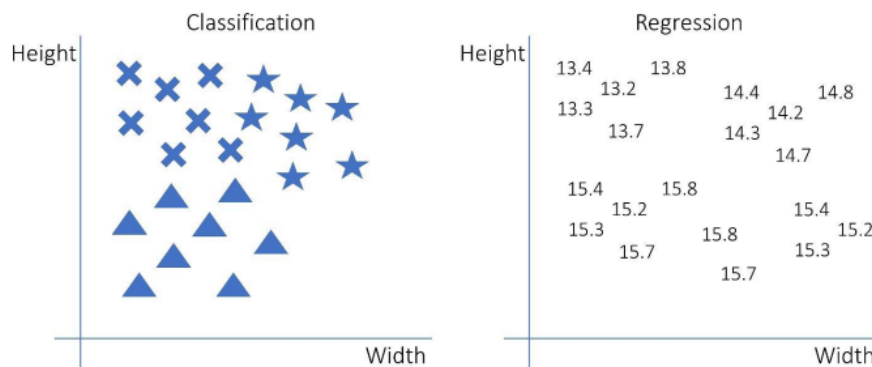


Figure 7: KNN algorithm [18]

- SVM — stands for support vector machine, the main task of the algorithm is to find the most correct line, or hyperplane, dividing the data into two classes. It is an algorithm that takes data as input and returns such a dividing line [19].

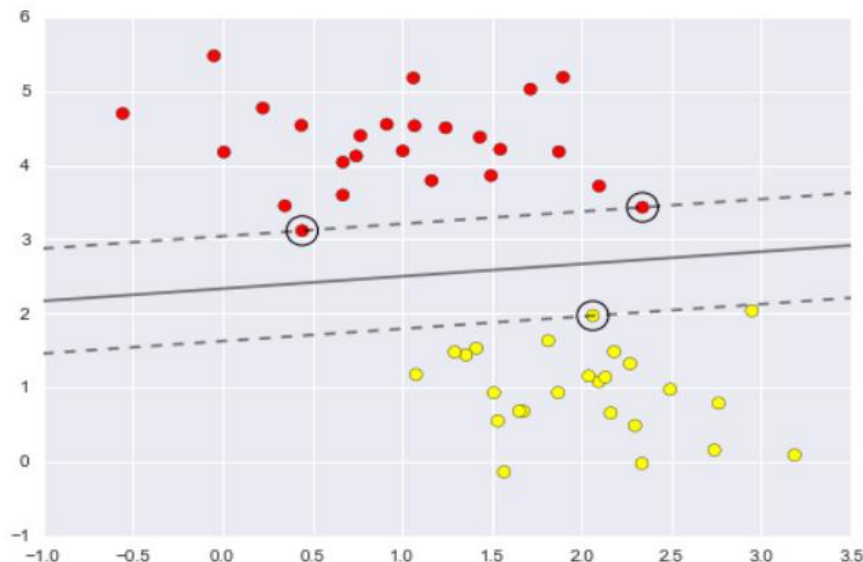


Figure 8: SVM algorithm [19]

Now that the mechanisms of operation of the various models used in this work have been described, we can directly start working on text data from the dataset.

## 4. Methodology

Classification of personality types by text that based on a personal information is an interesting and difficult task because of the highly complex nature of stylistic characteristics, including content, punctuation, syntax, comments and etc. In this case, it is necessary to resort to automatic computational processing of human languages, called Natural Language Processing. It is possible to classify personality types according to characteristics without NLP, referring only to topologies in which types are already divided according to a set of certain characteristics, but this makes it a little more difficult to provide the required information to the end user.

Also, when the types in the dataset are not initially classified, then clustering is performed, where they are combined according to the similarity of features.

In this work, Exploratory Data Analysis, Data Visualization and Pre-Processing, Feature Engineering, Training and Model Evaluation were performed. Now each of the stages will be considered in more detail.

## 4.1. Exploratory Data Analysis

The exploratory analysis begins with examining the data. The MBTI dataset is an open-source document, so it wasn't difficult to obtain data for training the model [20].

Table 2 shows that the dataset consists of 8675 rows and 2 columns containing information about personality types and posts. Through the work with the dataset, libraries were used for data analysis, data visualization, word processing, metrics, for building, training and evaluating models, as well as various machine learning packages.

**Table 2**  
 Data overview

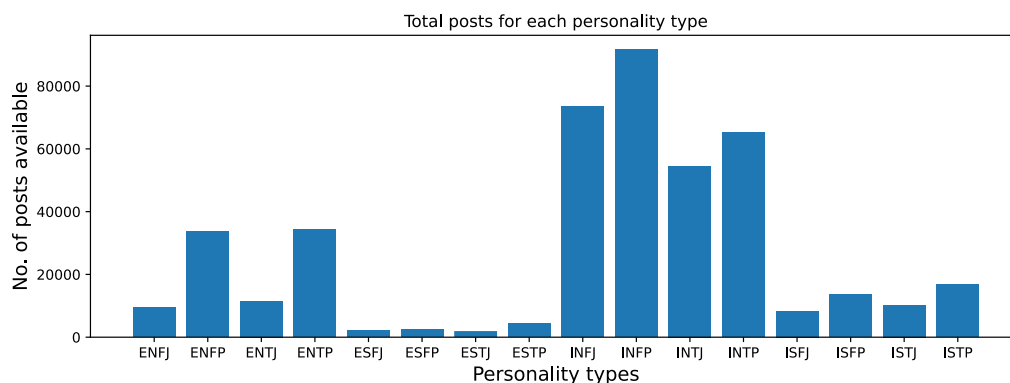
	Type	Posts
Count	8675	8675
Unique	16	8675
Top	INFP	"I'm being blackmailed in...
Freq	1832	1

It follows from the table that there are only 16 unique values of personality types, where INFP (Introvert Intuition Feeling Perceiving) is the most common. No duplicate posts were found. Since the dataset is complete and clean, the process of visualizing the data can be started.

There are no zero values, and all values were of type object. It is worth noting that all values are textual, accordingly, for training the ML model, it is required to convert them into numeric values.

## 4.2. Data Visualization

The distribution plot of posts for each personality type helps by looking at the number of data. It's known that the normal (or Gaussian) distribution represents data by the probability distribution of each value in the data. Most of the values stay around the mean, which makes the arrangement symmetrical.



**Figure 9:** The distribution plot of posts for each personality type

The built plot on Figure 9 clearly shows that the dataset is unbalanced. The distribution skewed to the right, centered around 7900 with most lengths from 4000 to 10000, and, obviously, the outliers present on the left. It is interesting to note that the largest number of posts were written by users with an introverted type, and the least with an extroverted type, who are more socializing in real life. The most common personality is INFP (Introvert Intuition Feeling Perceiving). Based on the graph, it can be concluded that users with a large number of comments are introverted, perceptive and emotional.

For a deeper understanding of plotting the distribution, below is a mathematical justification with an explanation of the formulas and values of the variables.

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \times e^{-\frac{(x-\alpha)^2}{2\sigma^2}}, -\infty < x < \infty \quad (*), \quad (2)$$

In this formula (3), the parameters  $\alpha$  and  $\sigma$  are fixed,  $\alpha$  - mean,  $\sigma$  - standard deviation.

$$f(t) = \int e^{itx} p(x) dx = \exp(i\alpha t - \frac{\sigma^2 t^2}{2}), \quad (3)$$

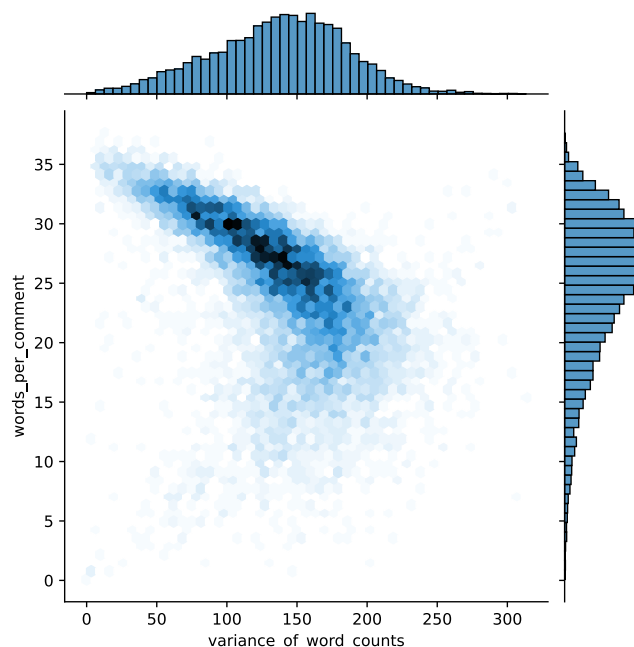
Differentiating the characteristic function and setting  $t = 0$ , the result is moments of any order. The normal distribution density curve is symmetric with respect to  $\alpha$  and has a single maximum at this point, equal to formula (4).

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \quad (4)$$

The standard deviation parameter  $\sigma$  varies from 0 to  $\infty$ . Average  $\alpha$  ranges from  $-\infty$  to  $+\infty$ .

New features “words\_per\_comment”, which counts the number of words per post out of the total 50 posts in the whole row, and “variance\_of\_word\_counts” helped to do some data exploration and to see how the raw data looks and to see how important new features were for distinguishing types across the MBTI personalities. Below are plots further showing the type imbalances in new data.

The joint plot illustrated on Figure 10 demonstrates the relationship between the two variables. The top distribution plot for the x-axis column (“variance\_of\_word\_counts”), the distribution plot on the right for the y-axis column (“words\_per\_comment”). These histogram plots represent Gaussian distribution. The scatter plot in the middle, which shows the distribution of data for these columns. The area under the histogram is calculated using a probability density function (PDF), where the highest peak of the curve is the mean of the distribution. It allows us to consider the relationship between “variance\_of\_word\_counts”, “words\_per\_comment” for each type separately as well.



**Figure 10:** The joint plot of post parameters

On a hexagon plot, the density of data points is shown in color. The hexagon with the most dots become darker. So, the graph shows that in most posts from 100 to 150 words, and in most words per user comment is from 25 to 30.



There's a large correlation between the number of words in a comment and variations in the number of words. Collaborative rafts have also been built based on more interesting parameters like the number of words in comments and the number of dots that the user enters comments.

The most common words in all posts found with the most\_common() function: 'I', 'to', 'the', 'a', 'and', 'of', 'is', 'you', 'that', 'in', 'my', 'it', 'for', 'have', 'with', 'but', 'be', 'are', 'like', 'not', 'an', 'I'm', 'on', 'was', 'me', 'as', 'this', 'just', 'about', 'think', 'or', "don't", 'so', 'your', 'do', 'what', 'at', 'can', 'if' and 'people'. This can also be seen on the Figure 11, which illustrates the top 15 common words in posts.

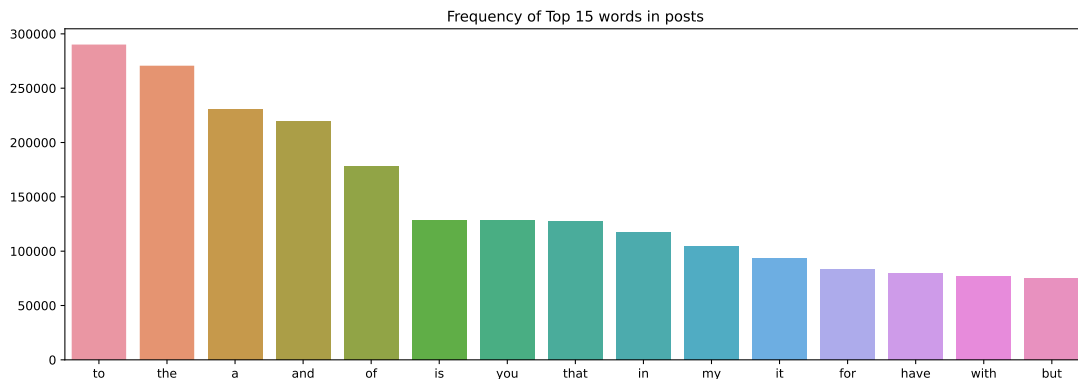


Figure 11: The most common words in all posts



Figure 12: WordCloud plot for each personality

At the stage of data preprocessing, it will be necessary to take this fact into account. Word Cloud

(or Tag cloud) is a graph that shows words of different sizes. This is a visualization of the frequency of words in a post, which is directly proportional to their sizes on the plot. The Figure 12 presents WorldCloud plots for each personality type.

Only one example of WordClouds from each general type is presented here, integrated into the form of characteristic images of 16 personality types. Based on the graphs, the following steps can be determined:

1. removal of irrelevant words like “ha”, “ar”, “Ti”, etc.;
2. the names of the MBTI types themselves in each post must also be removed since it can affect the training of the model;
3. other operations with raw data for converting it to the relevant form.

### 4.3. Pre-Processing Stage

This section covers data preprocessing, which involves cleaning up raw data in the most important text articles. The data is cleaned up using the functions below:

- `re.sub()` returns a new string resulting from a substitution with the specified pattern;
- `re.findall()` searches a string for non-overlapping occurrences of a pattern;
- `re.compile()` compiles a regex object for later use;
- `x.lower()` converts the object to lowercase.

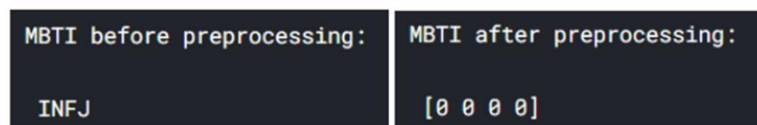
The functions use regular expressions, list comprehension, and `nltk` packages (tokenization and stopwords) to remove text that is considered irrelevant for sentiment analysis. After clearing the raw data, feature engineering begins. Each of the functions clears the data for the following notes from the previous sections:

- Removing all link data from posts;
- Keeping the end of sentence characters;
- Strip Punctuation;
- Removing multiple full stops;
- Removing Non-words;
- Converting posts to lowercase;
- Removing multiple letters repeating words;
- Removing very short or long words;
- Removing MBTI Personality labels in posts;
- Removing the posts with less than 15 words in each;

One of the main steps is dividing all personality types into basic 4 groups (I-E, N-S, F-T, J-P) in MBTI and binarizing it. To do this, the simplest solution is to use dictionaries and lists.

```
pers = {'I':0, 'E':1, 'N':0, 'S':1, 'F':0, 'T':1, 'J':0, 'P':1}
pers_list = [{0:'I', 1:'E'}, {0:'N', 1:'S'}, {0:'F', 1:'T'}, {0:'J', 1:'P'}]
```

Figure 13 demonstrates how each personality type is converted to binary using the `pre_process_type()` function.



MBTI before preprocessing:	MBTI after preprocessing:
INFJ	[0 0 0 0]

Figure 13: Binary conversion of each personality type

After data preprocessing the number of posts changed from 8675 to 8466. Not too many so begin the next part of the analysis.

### 4.4. Feature Engineering

As mentioned earlier, all data in the dataset is of type object and is presented in text format. To train

the model, it becomes necessary to convert string values to numeric values so that the data can be fitted into the model. That's why before splitting into X and Y features primarily the data should be converted into a numerical form using the LabelEncoder approach that is a part of the SciKit Learn library.

To get the attribute for the first column with categorical values, a class from the sklearn library is imported, then the column is processed by the fit\_transform() function and the text data is replaced with new encoded values.

The choice of label encoding over one-hot encoding was made to reduce the pre-processing time, and majority, because there are predefined 16 values under MBTI and assigning unique integers based on alphabetical ordering, seems like a viable option. In natural language processing, useless words are referred to as stop words.

Almost all of these were the most occurring words in the word cloud above. CountVectorizer is used to convert a collection of text documents to a vector of term/token counts and build a vocabulary of known words, but also to encode new documents using that vocabulary [21]. It also enables the pre-processing of text data before generating the vector representation.

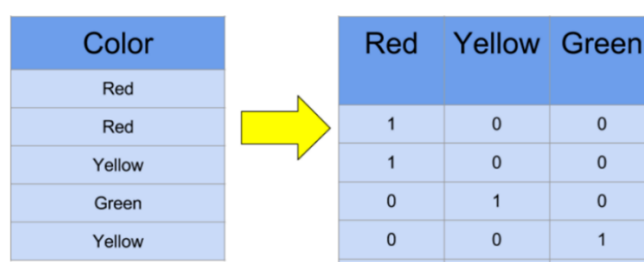


Figure 14: CountVectorizer work principle [21]

Here, stop\_words="english" is used with CountVectorizer since this just counts the occurrences of each word in its vocabulary, extremely common words like "the", "and", etc. will become very important features while they add little meaning to the text. Also, in the work, the ranking of texts by similarity to reference texts was made using the TF-IDF model. By itself, the abbreviation TF-IDF stands for TF - term frequency, IDF - inverse document frequency, that is, the ratio of the frequency of use of a word in a single text to the frequency of use of a word in all documents.

The model built on the basis of such a measure is perfect for searching for similar texts, since it allows comparing the aggregate measures of texts with each other, building a similarity matrix. This is an important step in pre-processing as the future model can often be improved if those words wouldn't be taken into account.

Now the model can be trained in multiple ML algorithms namely: Random Forest, XGBoost, kNN, Logistic Regression, Gradient Descent, Support Vector Machine, to choose the classifier which shows the best accuracy results. Additionally, the dataset will be split into testing and training in multiple ratios to find out which gives the best results.

## 4.5. Training model and Evaluation

Learning and evaluating are the most important steps in building a predictive model. The first step is to divide the data into training and test data. For this set, divisions will be used in the proportions of 60 by 40 and 70 by 30.

Varying the proportions of splitting the training and test data occurs due to the test\_size argument in the train\_test\_split() function. By setting the argument value equal to 0.4, the ratio is 60 by 40, when, accordingly, the value of 0.33 will be equated to the ratio 70 by 30. After splitting the data, models are built, fit in the data and predict the results. For convenience, predictions are rounded. Later, based on the accuracy of the prediction of each of the models, the choice of the most correctly trained one will be made.

It often happens that in the course of solving an applied data analysis problem, it is not immediately obvious or not at all obvious which training model is best suited. The most optimized solution might be

to choose the most popular or intuitively appropriate model based on the nature of the data available or just compare the results. For the future model were selected the following popular methods with already setted parameters:

- Random Forest: RandomForestClassifier(n\_estimators=100,random\_state = 1);
- XGBoost: XGBClassifier();
- Gradient Descent: SGDClassifier(max\_iter=5, tol=None);
- Logistic Regression: LogisticRegression();
- KNN: KNeighborsClassifier(n\_neighbors = 2);
- SVM: SVC(random\_state = 1).

The metric “accuracy” is used to evaluate classification models. It shows every part of correctly predicted forecasts.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (5)$$

Accuracy in binary classification is calculated in terms of negative and positive values using the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP - True Positives, TN - True Negatives, FP - False Positives and FN - False Negatives.

Despite the completeness of the data (absence of NA), the dataset is extremely unbalanced, so the accuracy values were less than the F1-measure.

## 4.6. Comparing algorithms and Results

There are a few parameters to improve the model's performance. During testing was founded that the not every model did well and the only one model predicts with reasonable accuracy. The input data were text values (small essay, story, letter, character description) of several people, whose personality type was initially determined. However, Figure 16 and Figure 17 below show comparative graphs of the results for both proportions of data splits. Hence, it is concluded that the best results are for two models: XGBoost and Logistic Regression.

In statistics and machine learning, data is usually divided into two subsets: training data and testing data (and sometimes into three: training, validation, and testing). In this way, the model is fitted to the training data to make predictions on the test data. The incorrect ratio of training and testing data has an impact on the predictability of the model - we may use a model that has lower accuracy and/or is not generalizable (i.e. it cannot be applied to other data).

The training set contains known outputs, and the model learns from that data so that it can later be generalized to other data. A test dataset (or subset) tests our model's predictions on that subset.

**Table 3**

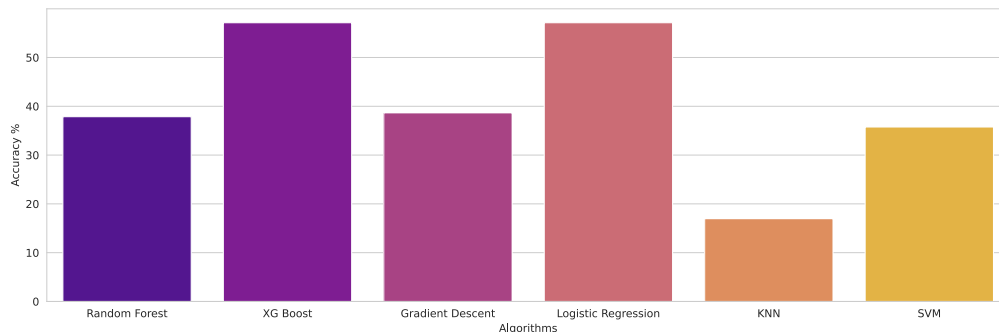
Comparison of models results in 60-40 split

	Accuracies (%)
Random Forest	39.533510
XGBoost	57.868320
Gradient Descent	37.348686
Logistic Regression	58.193091
KNN	16.445232
SVM	35.518158

According to some empirical studies, the best results are obtained if we use about 30% of the data

for testing, and the remaining 70% of the data for training.

Table 3 and Figure 15 present the accuracy values of the forecasting results of the above models in the split 60-40.



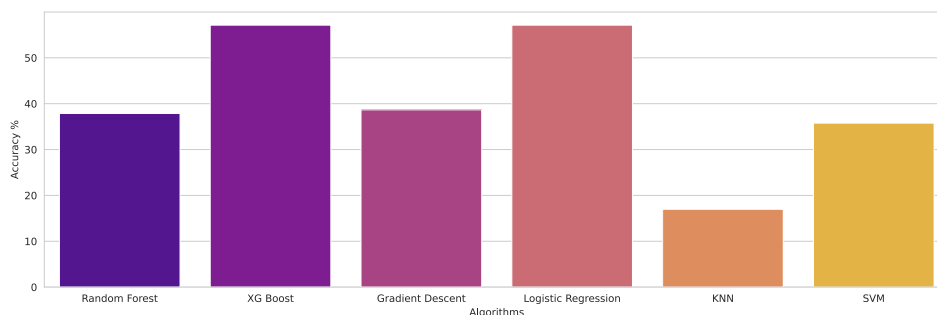
**Figure 15:** Comparison of models results in 60-40 split

Table 4 and Figure 16 present the accuracy values of the forecasting results of the above models in the split 70-30.

**Table 4**

Comparison of models results in 70-30 split

	Accuracies (%)
Random Forest	37.866858
XGBoost	57.122405
Gradient Descent	25.518969
Logistic Regression	57.122405
KNN	16.964925
SVM	35.755190



**Figure 16:** Comparison of models results in 70-30 split

Comparing all the characteristics of machine learning models, it can be concluded that Logistic Regression is the best and most efficient of all. Unfortunately, this model does not accept floating-point values, which are important for this research work.

Out of all the models, an average XG Boost gives a relatively good performance, hence it was chosen to build a personality prediction model. This will be beneficial as the XGBoost model can even be used to evaluate and report on the performance of a test set for the model during training. Some configuration heuristics have been published in the original gradient boosting papers such as:

- learning\_rate in XGBoost should be set to 0.1 or lower, and smaller values will require the addition of more trees;
- tree\_depth in XGBoost should be configured in the range of 2-to-8, where not much benefit is seen with deeper trees, and so on.

## 5. Conclusion

This study set out to determine personality traits from textual data was considered. This problem is one of the fundamental components of career guidance development. To solve it, we used natural language processing (NLP) methods, the possibilities of which revealed new approaches to this issue.

Models such as Random Forest, XGBoost, Gradient Descent, Logistic Regression, KNN and SVM were also considered and tested. It has been found that the XGBoost model shows the best accuracy metrics when the learning\_rate is set to 0.1 or lower, and the tree\_depth is between 2 and 8.

Presumably, the results of the predictions above are not as accurate due to lack of data. This prompted us to think about future work. It is assumed that the indicators will be divided among themselves into groups, or separately, each of which will represent a holistic parameter and will also be forecasted separately. In other words, in this work we found the type as a whole, for example, INFJ. In the following works, it will be necessary to break the types into subgroups, that is, N-F or NF and etc.

In addition to the work, it will be interesting to consider the influence of the presence of punctuation marks that show the emotionality of the text and, accordingly, the author of this text.

## 6. References

- [1] Steve Crabtree, "Worldwide, 13% of Employees Are Engaged at Work", Gallup, 2013, <https://news.gallup.com/poll/165269/worldwide-employees-engaged-work.aspx>.
- [2] Myers-Briggs Type Indicator (MBTI)-A Text Classification Approach, "Proceeding of International Conference on Advances in Computing", Communications and Informatics (ICACCI), 2018, 1076-1082.
- [3] H. J. Eysenck, Dimensions of Personality, 1947.
- [4] Kiselev P., Kiselev B., Matsuta V., Feshchenko A., Bogdanovskaya I., & Kosheleva A., "Career guidance based on machine learning: social networks in professional identity construction." Procedia Computer Science 169 (2020): 158-163.
- [5] Talasbek A., Serek A., Zhaparov M., Yoo S. M., Kim Y. K., & Jeong G. H., "Personality classification experiment by applying k-means clustering." International Journal of Emerging Technologies in Learning (IJET) 15.16 (2020): 162-177.
- [6] Cui, Brandon, and Calvin Qi. "Survey analysis of machine learning methods for natural language processing for MBTI Personality Type Prediction." (2017).
- [7] Amirhosseini, Mohammad Hossein, and Hassan Kazemian. "Machine learning approach to personality type prediction based on the myers-briggs type indicator®." Multimodal Technologies and Interaction 4.1 (2020): 9.
- [8] Liang, Ying Siu. End-user Robot Programming in Cobotic Environments. Diss. Université Grenoble Alpes (2019).
- [9] [https://en.wikipedia.org/wiki/Data\\_science](https://en.wikipedia.org/wiki/Data_science)
- [10] Jones, T., Models for machine learning, IBM Developer, (2017), <https://www.ibm.com/developerworks/library/cc-models-machine-learning/index.html>.
- [11] Hester T., Vecerik M., Pietquin O., Lanctot M., Schaul T., Piot B., ... & Gruslys A., "Learning from demonstrations for real world reinforcement learning." (2017).
- [12] <https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861>
- [13] Will Koehrsen, "Random Forest in Python", Towards Data Science, 2017, <https://towardsdatascience.com/random-forest-in-python-24d0893d51c0>
- [14] Wang, Yuanchao, et al. "A hybrid ensemble method for pulsar candidate classification." Astrophysics and Space Science 364.8 (2019): 1-13.
- [15] Sundar, Krishnan, Neyaz Ashar, and Liu Qinzhou. "IoT Network Attack Detection using Supervised Machine Learning." (2021).
- [16] <https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31>
- [17] [https://en.wikipedia.org/wiki/Logistic\\_regression](https://en.wikipedia.org/wiki/Logistic_regression).
- [18] <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e7336aa1>.

- [19] [https://www.sas.com/en\\_gb/insights/articles/analytics/machine-learning-algorithms.html](https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html)
- [20] <https://www.kaggle.com/datasets/datasnaek/mbti-type>.
- [21] Hunter Heidenreich, “Natural Language Processing: Count Vectorization with scikit-learn”, Towards Data Science, (2018), <https://towardsdatascience.com/natural-language-processing-count-vectorization-with-scikit-learn-e7804269bb5e>.

# Data Analysis and Development of the Cross-Platform Application Intended to Organize and Provide Services in the Field of Science-Cultural and Entertainment Events

Danel Mashakova<sup>1</sup>, and Marat Nurtas<sup>1</sup>

<sup>1</sup> International Information Technology University, Almaty, Kazakhstan

## Abstract

The intention of this research was to develop a cross-platform application for organization of a diversity of events. In the period of a global pandemic the industry of occurrence management struggled as much as never before. Thus, it was decided to simplify our life by developing a cross-platform application with well and objectively organized event management services. Descriptive analytics and exploratory data analysis of the most frequently conducted activities were performed to identify relevant service sectors and select key priorities in collaboration with partners. This application has gathered the target audience, regardless of age or social category, due to the prevalence of folk celebrations in Kazakhstan.

## Keywords

Event organization, application development, data analyzation, results prediction, features comparison

## 1. Introduction

Events create opportunities for people to connect with an area, spend time together, celebrate and experience the diversity of cultures and foster creativity and innovation. They allow a community to come alive and provide an opportunity for a destination to showcase its tourism experience and increase economic activity. Community building, lifestyle and leisure enhancement, cultural development, tourism promotion and increased visitors, volunteer participation, fundraising, and economic development are all key benefits of events.

Research initial data: Research fundamentals course knowledge, basic research data about event management, case studies from event organizers, theoretical methods of formalization.

List of tasks statements:

- Make detailed data analysis on research areas
- Train certain machine learning models
- Develop a platform for multiple clients and companies
- Arrange a survey and test demonstration

Event management system is a cross-platform tool that helps with project administration for large-scale events including festivals, conferences, ceremonies, weddings, formal parties, concerts, and conventions. It involved researching the brand, determining the target demographic, developing the event concept, and coordinating the technical aspects prior to the event's launch. The event management system enabled customers to view various packages/products about the event and make booking through the online platform. Budgeting, scheduling, venue selection, obtaining essential permits, coordinating, transportation and parking, arranging for speakers or performers, arranging decor, event security, catering, working with third-party vendors, and emergency preparations were all part of the event planning process. Because each event was unique in its character, the method of organizing and executing it varied depending on the sort of event.

The multifunctional cross-platform application designed to assist the user in searching the necessary services and organizing events in various areas. The application will include a database of partner-companies and customers providing a variety of services. Descriptive analytics and exploratory analysis of the most frequent activities will be made to identify relevant service areas and select key priorities in collaboration with partners.

The main advantages of the project are reduced time spent on searching and preparing in social



networks, search engines; guarantees of work quality and timely execution; ability to track detailed work progress and preparation with feedback. Considering all multinational holidays and popular events in Kazakhstan, this application will gather a target audience regardless of the age and social category.

## **2. Analytical review**

About ten years ago, not many people in Kazakhstan thought about requesting services from professional event organizers. Moreover, they didn't realize that event management was already considered as a whole business sphere, and it is a significant project management aspect, that people purposely study to gain main skills for professional work. But today we can find different kinds of articles [1], from news and social media to scientific journals [2] and even whole chapters in book series [3]. Some sources are mostly analytical with comparison of national celebration organization range. Others describe current trends and situations, interviews event agencies directors and main managers.

Event management improvement in Kazakhstan can help develop the domestic economy, especially tourism and image of the state since we have a very rich variety of social, cultural, political, and sporting events and activities [2]. Our country also hosted international events of both political and sports characters, so influence has been changing over the past two decades to affirm status in the eyes of the worldwide associations [3].

In the digital age, through ICT's, the ways of searching and accessing information about the event by the users have been developed and transformed. Likewise, businesses can facilitate their management and marketing events through digitalization. They sustain their events by organizing new events according to the customer base and the type of event for the management. Technology has an impact on every part of one's life, from communication to entertainment, and it continues to evolve and alter on a daily basis as a result of new breakthroughs and innovations [4]. These factors have an impact on the event industry. As new technologies become available, industry competitors are pushed to adapt to the changes or risk being left behind [5].

Companies are applying technologies to be closer to their customers, creating awareness through social media, sharing information and collaboration through websites and applications [6]. While there is no academic definition "Event Technology" following a review of literature on the topic it can be understood that event technology encompasses any application of technology within the events industry to enhance, facilitate and develop the experience [7].

## **3. Methods and technologies**

### **3.1. Diagrams and schemes**

Using Unified Modeling Language (UML), Use case for user and service partner and IDEF diagrams were developed. First, you can see our use case diagram in Figure 1, which presents possible scenarios and interactions. Main focus is on User and Service partner, since most functions developed and improved to assist them during exploitation. Users can explore events, companies, services and at the end make payment and leave reviews. Service partners can control their services, discounts and make payment as well, for using our platform. You can see that they both have profile and chatting features for service analysis and future discussion. Besides them, we have a Manager for administrations and Tech support.



### 3.2. Language and programming environment choice

When it comes to creating applications, picking the correct programming language is crucial. Otherwise, after a few hours of work, the developer may discover that the chosen language does not allow him to work efficiently in order to meet his goals.

In the end, the choice was between two JavaScript frameworks: Vue and Angular. There are differences between them, which are quite significant. However, each of them is suitable for solving problems. Angular also requires the use of TypeScript, an optional static JavaScript language. TypeScript has its obvious advantages - checking static types can be very useful for large applications. However, often for applications, the introduction of type systems can lead to more overhead than increased development performance. In such cases, it is better to use Vue, since using Angular without TypeScript can be difficult. As a result, we have chosen Vue.js as our web application language since from the beginning it was more understandable than Angular.js.

React Native enables a single JavaScript codebase for two different platforms. In this case, the React Native framework will use it to implement an iOS application, because its accessibility is much more attractive than developing on Swift/Xcode. By accessibility, we mean the monetary issue of buying hardware (macOS devices). It is feasible to create high-quality iOS applications using the React Native framework at a low cost. All common test automation frameworks are supported by React Native, therefore we can use XCTest to test our iOS application. Apple's testing framework allows us to execute unit, performance, and user interface tests. On top of that, the ability to use Keep It Functional (KIF) – the iOS functional testing framework. Fortunately, we can also automate tests with Appium and Robot Framework.

Android is the world's most popular mobile operating system. The OS has maintained its dominance in the mobile industry for a long time. It has been continuously developed since its birth. Despite Apple's stability and unchanging style, Android devices are the most preferred and are sold in big quantities. Android, unlike other operating systems, is open source. This allows companies like Samsung, Xiaomi, and Huawei to actively use Android and build their own operating systems on top of it.

## 4. Implementation part

### 4.1. Survey results

Even though statistical or general data regarding event management in Kazakhstan was not enough, all the more only Almaty, that didn't stop work because it was decided to gather own data with survey [8]. It consists of relevant questions that helped us get better results during training and testing (Figure 3). Such as general information about a participant, his preferences in organizing or attending events and overall interest in our application.

Приложение по организации мероприятий Together

Здравствуйте! Мы студенты 4го курса университета ИТУ. Данный опрос имеет отношение к нашей дипломной работе, посвященной разработке приложения по части организации ивентов. Кратко о самом проекте - это приложение, которое сокращает время на поиск и обработку информации по компаниям, проводящим разный спектр мероприятий: свадьбы, дни рождения, корпоративы и так далее. Благодаря TOGETHER наш озер может быстро создать подборку этих самых компаний в соответствии со своими пожеланиями и бюджетом.

Мы проводим опрос в целях лучшего понимания наших потенциальных пользователей: их нужд, потребностей и актуальных проблем. Просим вас ответить на несколько простых вопросов, которые не займут у вас больше 10 минут. По всем вопросам можете обращаться по электронной почте [nordnordnord@gmail.com](mailto:nordnordnord@gmail.com) (Жания)

danel.mashalkova@gmail.com (без совместного доступа)  
Сменить аккаунт

\* Обязательно

Укажите ваш пол \*

Мужчина

Женщина

Ваш род деятельности? (студент, работник [в какой профессии], и тд) \*

Мой ответ

Имели ли вы опыт в организации мероприятий (свой день рождения, день рождения близких, корпоратив, свадьба и тд) \*

Да

Нет

Если имели, то как вы оцениваете свой последний опыт? Оценка идет по нескольким критериям: поиск компаний, места проведения, развлекательной программы, меню, бюджет, общее время затраченное на организацию

Положительно - не было проблем ни с одним из пунктов

Средне - с одним или несколькими пунктами были затруднения

Негативно - по всем пунктам были проблемы, приходилось менять всё несколько раз

Figure 3: Together survey

Data mining uses ML algorithms to predict the development of case, identify patterns and assess the significance of factors. According to collected data from the survey, exploratory data analysis by utilizing ML was done. It is described in the “Machine Learning analysis” subparagraph.

## 4.2. Firebase Integration

First, it is important to create a Google Firebase project database [9], which is fortunately free and easy to share between users. All we need is a Google account and just in case enable Google Analytics and we are done. First, we want to test the Authentication segment after adding Email/Password provider. Even though this collection does not allow adding custom columns, for our current project state this user database should be enough. In future might upgrade our plan which will give access to the Functions segment with different additional features for all Firebase solutions including Authentication.

Between Firestore and Realtime databases it was decided to work with the first one since it is more NoSQL structured and has more information regarding accessing and working with data. In Figure 4 you can see collections that both user and event organizer are using while working in applications. Before sending data to firebase from our applications, manually created some collections and documents to better understand how everything is structured. One of the main is certainly organizations since it holds various information concerning its services, types of work, organizer etc. Certainly, in the future there will be far more databases to increase work efficiency for users.

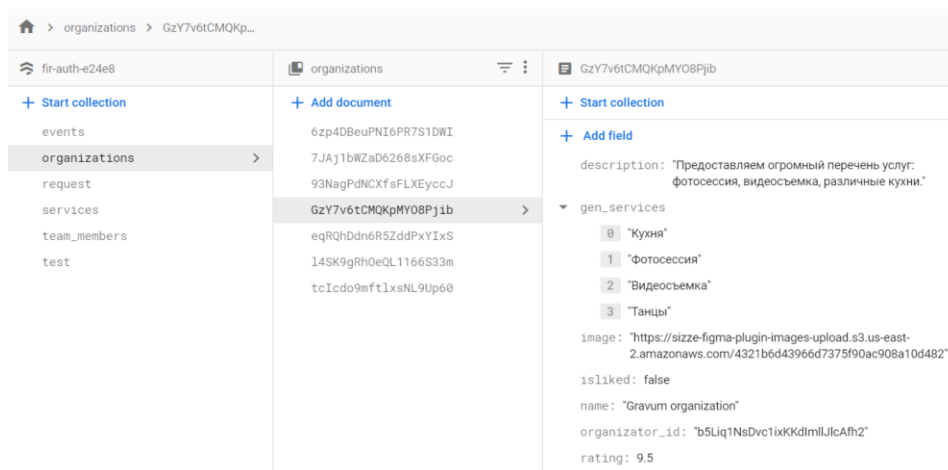


Figure 4: Firebase collections

## 4.3. Web application

Before starting to work with firebase first of all we have to connect to it using certain credential information and libraries. After creating a separate JavaScript file, we initialize the application with firebase and with the Firestore database. Also, we make connection with the analytics segment for further dashboard analysis and overview [10]. In the case of Vue JavaScript, here the difference between variables fb and db is that we use firebase to connect to the Authentication part of Firebase since it has its own separate user database, provider, sign-in methods etc. While db is for Firestore collections where we store cloud data.

Firestore SDK algorithm:

1. Import firebase library with firestore requirement as well;
2. Create variable Vue instance for firebase configurations;
3. List all needed data such as API keys, authentication domain, database URL, project id, storage bucket etc;
4. Provide constants for application and analytics initialization;
5. Export all variables and constants.

For the dashboard tab it was decided to show various graphs, so the event organizer could see his income processes and make changes himself regarding increasing staff for certain services or improving quality of work so he could respectively enhance prices. Figure 5 illustrates a pie chart which shows the ratio between categories depending on previous events costs. Graph stylization might be reminiscent of Excel charts, so for users it will not be hard to adapt to a new system. We are using vue-chartjs which is a more optimized and easier to use version of Chart.js library.

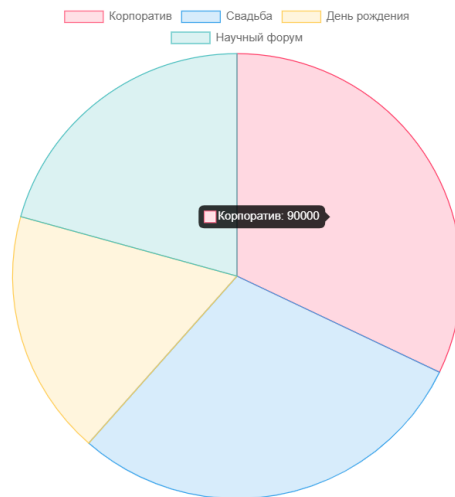


Figure 5: Pie chart

In Chart methods pseudocode you can see the methods of pie charts. Here we compare categories names with the same field in events database and in case of correspondence, we sum cost values for each one [11]. Since we constantly work with arrays it is compulsory to use map function for iteration, but for fetching data from collections it is better to use forEach since it enumerates the children's elements in the DataSnapshot to their query. Surprisingly when we tried to use a map to get data from the events database, it did not collect data correctly.

Chart methods pseudocode:

methods component:

```
setup entry point(categories) {  
  Catching all events({  
    Compare event's categories with full list  
  })  
  create constant for current user to get his id  
  open RenderChart function for graph visualization ({  
    Setting categories as labels,  
    Setting categories and current user data for datasets component [  
      if all data exists (  
        summing total cost  
      )  
    ]  
    return total  
    Listing background and border colors for pie chart  
  ] }  
})
```

Even though working with Vue JavaScript and firebase together was not easy, gained huge experience and understood how different versions and releases can affect each other and the project itself. It was noticed that the latest JavaScript frameworks tend to use cloud NoSQL databases instead of simple SQL. Which means that regarding programming and environment choices, this project is relevant.

## 4.4. iOS application

Using the React Native framework with Firestore database started to write adaptive code for iOS [12]. Firebase is a Backend as a Service (BaaS) that gives mobile app developers who use React Native a massive advantage. As React Native developers, we can use Firebase to begin developing an MVP (minimum viable product) while keeping costs down and prototype the app quickly. The react-native-firebase library is an officially recommended set of packages for bringing React Native support for all Firebase services of iOS applications.

Then initialized Firebase with our configuration values. The addition to the Firebase SDK algorithm includes the Firebase Authentication service module, which is stored in the library “@react-native-firebase/auth” and declared as auth() method. With the onAuthStateChanged() method, which is in the module allows us to subscribe to the users' current authentication status and get an event if that state changes:

For the authentication process we have two screens which are responsible for sign in/up methods: Login and Registration. A Registration screen where the user can create an account, and a Login screen where an existing user can log in.

After authentication, our user is familiarized with the functionality of the application. Customers can find basic information about our partner organizations on the main page. The documents of all event organizations are extracted in JSON array format using a filterless query to the Firestore database [13]. The array's data is subsequently transformed into a readable format and displayed on the screen.

Collections are used to hold all documents. Subcollections and nested objects, both of which can contain primitive fields like strings or complicated objects like lists, can be found in documents. Construction of the document is represented in Figure 6.

```
"database" => Provider {
  "component": Component {
    "instanceFactory": [Function anonymous],
    "instantiationMode": "LAZY",
    "multipleInstances": true,
    "name": "database",
    "serviceProps": Object {
      "DataSnapshot": [Function DataSnapshot],
      "Database": [Function Database],
      "INTERNAL": Object {
        "dataUpdateCount": [Function dataUpdateCount],
        "forceLongPolling": [Function forceLongPolling],
        "forceWebSockets": [Function forceWebSockets],
        "initStandalone": [Function initStandalone],
        "interceptServerData": [Function interceptServerData],
        "isWebSocketsAvailable": [Function isWebSocketsAvailable],
        "setSecurityDebugCallback": [Function setSecurityDebugCallback],
        "stats": [Function stats],
        "statsIncrementCounter": [Function statsIncrementCounte...(truncated to the first
```

**Figure 6:** Document construction

On the user side, the main operations include viewing data and making a request to the database for further communication with the organization. Users can easily send a request for future interactions with event organizations. Customers can search and generate lists of companies using a variety of methods, including numerous filters, search, and generation. To filter the query, the generation process employs a variety of tools. The where() method can be chained onto a collection reference to filter documents within it. A field to filter on, a comparison operator, and a value are all passed to the where() method. Cloud Firestore has extensive query functionality for identifying which documents from a collection or collection group we want to retrieve [14]. These queries can also be used with addSnapshotListener() or get(). After selecting a suitable organization based on all characteristics such as location and the range of services offered, users select a specific service and send a request to the database, where the organizer himself will call customers later and discuss the details.

The add method adds a new document to our collection and assigns it a randomly generated unique ID. We will use the set method on a DocumentReference instead if we want to specify our own ID. Firebase is worthy of the scalable solution for iOS application development utilizing React Native framework, when we want to transfer our data to several users at the same time without having to worry about disruptions. Data is served and synchronized quickly with Firebase, allowing users to access files and data from anywhere in the world.

## 4.5. Android application

We have finally come to Android's part and now we are going to look through how to integrate Firebase in our project [15]. Fortunately, Android Studio gives the possibility to make the Firebase connection process much easier without manually creating various files with credentials.

In Figure 7 you can see that in tab Tools there is a special button for this database. After clicking on it, we are automatically redirected to a website with available projects, and we choose the one that we need. Thanks to it, we do not waste our time on connection, because the environment downloads all needed libraries itself that means we can immediately start coding.

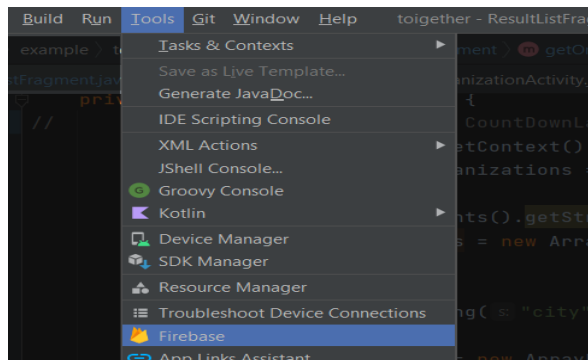


Figure 7: Firebase tool

In the Android app we have a lot of fragment pages. And here you can see the description of often used logic regarding View creation, data parsing, Firebase GET request. At the beginning of every fragment there is an `onCreateView` function. It creates and returns the view structure related to fragments. Inside of it we instantiate the user view and subsequently build the structure of the page. At the bottom you can notice that we call `getOrganizations()` function.

While a user goes through generating compilation his chosen data is saved into `SharedPreferences`. It's a simple object that links to a file containing key-value pairs and provides read and write methods for them. City, event category, service list and other keep saved in certain file. It allows us to carry the user's choice through pages while he is turning them. Also, this feature would help us to save the user's interests for future recommendations, even after he closes the application. The data is ready to be used in a query. By declaring a variable `db` we get a default instance of the database. Cloud Firestore's strong query feature helps to select which documents from a collection you want to retrieve. The query returns all documents matching the request in ascending document ID order by default. Your data can be sorted in whatever way you want. It is possible to combine filters: `orderBy()`, `limit()`, `whereEqualTo()`. Seemingly, it is much easier and better than ordinary sql queries. Nevertheless, there are a lot of disadvantages and problems. For instance, our organization's essence has 3 arrays as attributes. For sorting arrays as it is demonstrated below the `whereArrayContainsAny()` function is used [16]. But we can-not use it more than once. Firestore limitations do not allow the use of array filters more.

Overall, for Android applications Firebase is a great option for cloud database usage. Various tutorials and materials assisted us through the whole coding process. We could easily find errors description and how to fix them no matter how unexpected they were. It was rather hard to consider assembling certain attributes for both sides: client and entrepreneur, which are compulsory for all platforms.

## 5. Machine learning analysis

Besides creating one web and two phone applications a consumer interest forecasting using Machine Learning for event industry in Kazakhstan was built. Machine learning techniques allow to predict different features in this sphere, such as the number of products/services from partners to be obtained during a defined future period or more accurately define our target group each time season [17, 18].

Possibilities are endless, so we settled to start with overall analysis and after getting a clear picture, choose target value. In this case, a software system can learn from data for improved analysis [19]. The first task when initiating the demand forecasting project is to provide meaningful insights [20].

The process includes the following steps:

- gather available data;
- briefly review the data structure, accuracy, and consistency;
- run a few data tests and pilots;
- look through a statistical summary.

In Figure 8 you can see the first columns of a whole data table and its information. You might think that some information is not useful, for example the number of pets in a family. But it is possible that if a person has too many animals in his house, then he does not organize events at his house very often or even attend some in order not to leave his pets alone for a long time. After all, even the small details together can drastically affect the final result [21]. Some values we had to extract manually from participants' answers, which took time, but as a result we got a very solid database to work with.

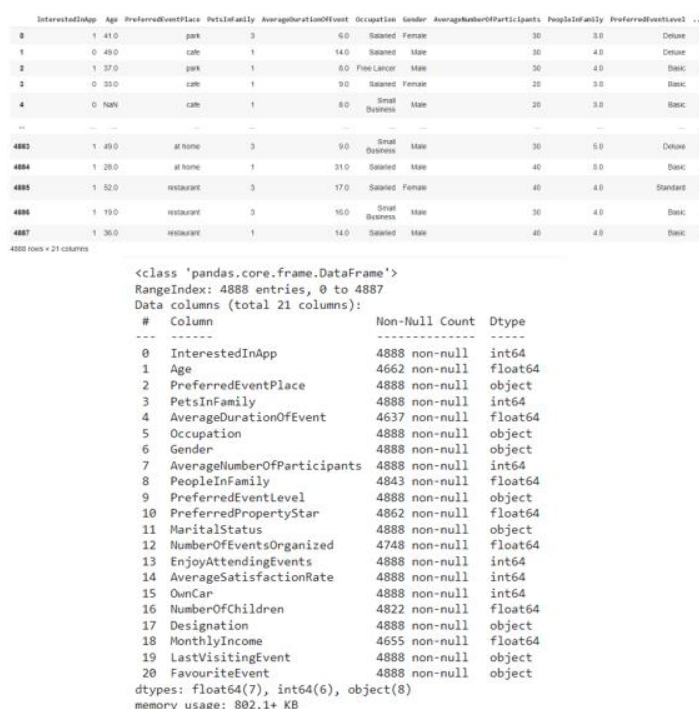


Figure 8: Data table and information

After that listing categorical features data:

- InterestedInApp [1 0]
- PreferredEventPlace ['home' 'cafe' 'hotel' 'restaurant' 'townsquare' 'office' 'park' 'beach']
- PetsInFamily [3 1 2]
- Occupation ['Salaried' 'Free Lancer' 'Small Business' 'Large Business']
- Gender ['Female' 'Male']
- AverageNumberOfParticipants [30 20 10 40 50]
- PreferredEventLevel ['Deluxe' 'Basic' 'Standard' 'Super Deluxe' 'King']
- MaritalStatus ['Single' 'Divorced' 'Married' 'Unmarried']
- EnjoyAttendingEvents [1 0]
- AverageSatisfactionRate [2 3 5 4 1]
- OwnCar [1 0]
- FavouriteEvent ['concerts' 'weddings' 'festivals' 'birthdays' 'parties' 'exhibitions']

Now we can choose some columns and create plots. Paid special attention to the column



InterestedInApp and as the name suggests it shows whether participants would download our app or not. 0 equals “yes” and 1 – “no”, so we show the number of observations for categorical data, and as parameter for encoding we set InterestedInApp. As a result, we receive a 5x3 array of subplots which in total 15 bar plots.

In Figures 9 and 10 it was decided to show in total eight of fifteen graphs since the picture would be too big. From the first graph it seems that more people are interested in application rather than not. Then we have PreferredEventPlace, in which can be noticed that the most popular variants are home, restaurant, park and town square. Probably because first, second and third are places where young people spend time and can afford while second could be adults with higher income. Third chart is Occupation where it is predictable that most people are employed and either have their business or work. Last is gender and the results show that among participants were a bit more women.

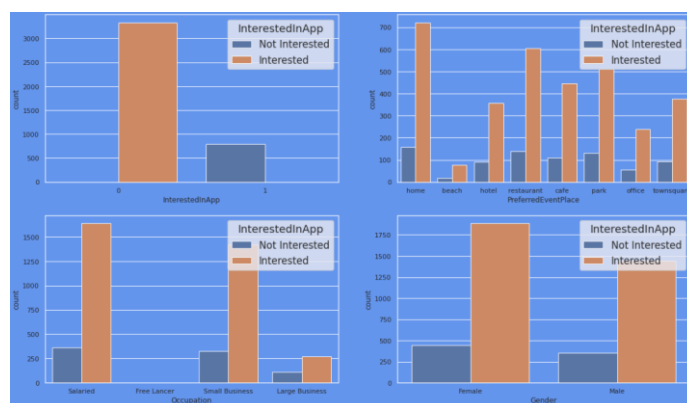


Figure 9: First bar plots

Other four presents less significant but still considerable parameters such as preferred organization level, martial status, average satisfaction rate and car availability. Among them probably organization level could play important role. But it depends on whether participant is organizer or attender because in first case MartialStatus can be significant feature since event budget is one of the key factors.

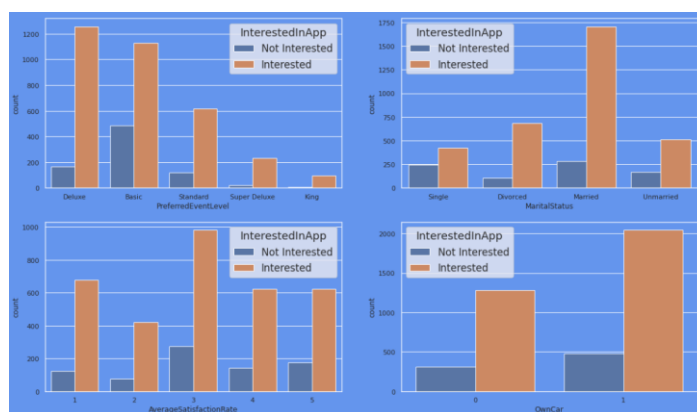


Figure 10: Second bar plots

Before we run a model, we need to fit it to a different set of data than the data that will be tested it on. In classification algorithms, balancing data is a common aspect of the data science process [22]. Data resampling is a set of approaches for transforming a training dataset in order to balance or improve the distribution of classes. Standard machine learning methods can be trained directly on the altered dataset without any modifications once the dataset has been balanced. This enables a data preparation strategy to meet the difficulty of imbalanced classification, even with substantially imbalanced class distributions. In our case, we decided to balance the target value by under-sampling one class in favor of the other. As a result, we have balanced classes that are ready to be split into train/test datasets.

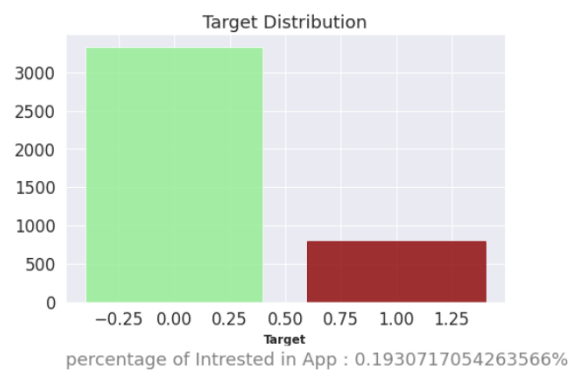
To get the numeric values for the features, we first did a label encoding. Our dataset has eight

categorical columns. We converted categorical data to numeric using the LabelEncoder() function from sklearn.preprocessing library in the transform method. First making copy of dataset then check if columns exist and after that put them in fit\_transform function. Repeat this process in recursion form and return output.

Using SelectKBest from sklearn.feature\_selection decided to determine best features and their score with chi2 as score\_func:

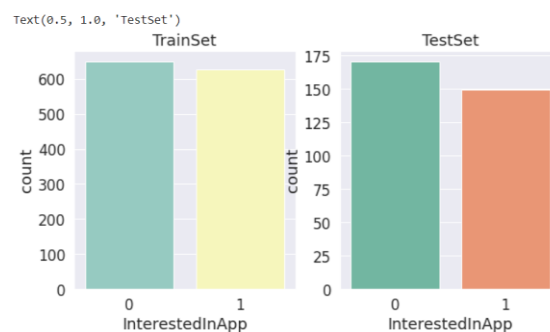
- MonthlyIncome – 64526.316713
- Age – 223.966488
- EnjoyAttendingEvents – 214.101731

After determining features, we need to determine target distribution and then in Figure 11 plotting graph for accurate resampling analysis.



**Figure 11:** Target Distribution

The final step in data preprocessing is to split the resampled dataset into train and test sets (Figure 12). The train-test split is used to estimate the performance of machine learning algorithms for prediction-based applications. This method is a simple procedure for comparing our own ML model outcomes to machine results. We use the train\_test\_split() function from the sklearn library and split our dataset in an 80/20 ratio. As a target value we use the IntrestedInApp column, which consists of binary values: 1 and 0.



**Figure 12:** Train and test sets

Forecasting algorithms are not "one-size-fits-all." Several machine learning algorithms are frequently used in demand forecasting characteristics. Machine learning models are chosen based on a variety of parameters, including the business purpose, data type, data quantity and quality, forecasting time, and so on.

You can see our list of trained models of these particular machine learning approaches as applied to our clients and the evaluation for them:

- AdaBoost

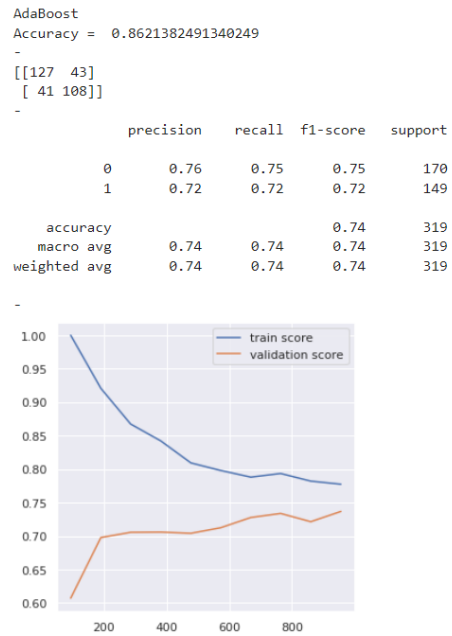


Figure 13: AdaBoost metrics

- K-Nearest Neighbors Classifier

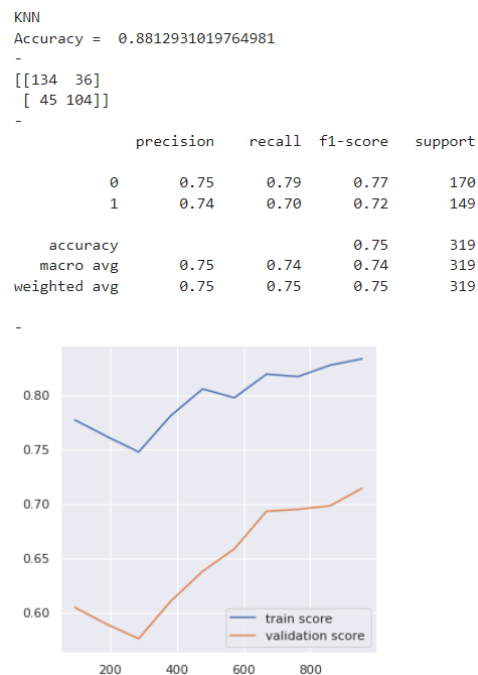
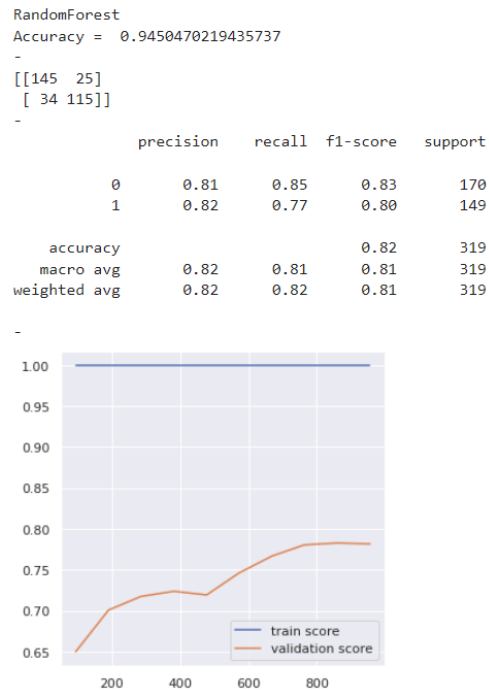


Figure 14: K-Nearest Neighbors metrics

- Random Forest



**Figure 15:** Random Forest metrics

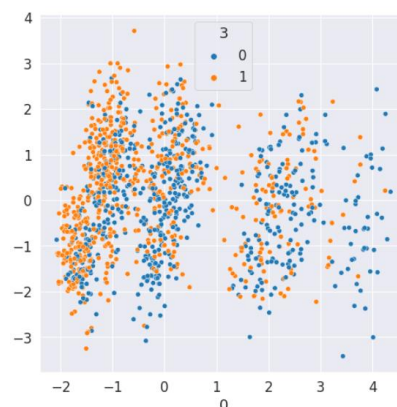
The type of product is a key consideration for the demand model. A perishable product demand model, for example, should not overstate demand because the excess goods will be wasted. Instead, the modeling error should always be set to a lower amount of inventory than an actual demand.

For additional observations Principal component analysis (PCA) was made with dataframe (Figure 16) where concatenated transformed X\_train and y\_train. plotted using scatterplot in Figure as well as scatterplot\_3d.

	0	1	2	3
0	-1.258290	1.195632	-0.577769	1
1	-0.715275	-0.157995	0.307032	1
2	3.642519	-1.287342	0.124866	0
3	0.291203	-0.189715	1.120294	0
4	-0.854242	2.900300	1.452473	1

**Figure 16:** PCA

Also printed plots using scatterplot in Figure 17 as well as scatterplot\_3d in Figure 18. Using update\_traces checked available traces in chart but can be visualized as legend.



**Figure 17:** Scatterplot

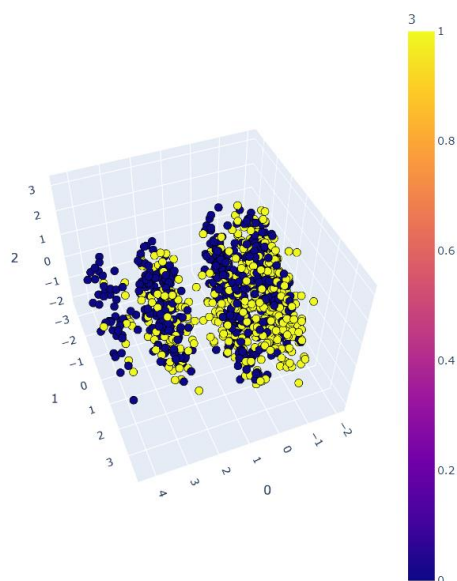


Figure 18: Scatterplot\_3d

## 6. Conclusion

The purpose of the thesis is to design and develop the event management system for the multifunctional cross-platform application. The first part of the work describes characteristics and analyzes the project structure. Moreover, in this part, the need to create the cross-platform application is justified and analysis of the event management system is conducted. The second part is devoted to technologies used during the project development. And the third part includes the stages of the designing and the structure of the project.

Main task goals: analysis and approach. Each of them has an approximate deadline, responsible team or person, in what type it should be formalized and prelim budget. Among important tasks we can highlight “Advert strategy development” and “Contact collection system”.

The main problem of users, which is solved by using the TOGETHER application, is to reduce the time to search for companies that will meet all the client's requirements regarding service, design, locations, and entertainment programs. Each partner offering their services on our platform undergoes a thorough check, including compliance with quarantine conditions, if these measures remain relevant at the time of the launch of the application on the market.

Entire possible range of events was covered, both entertaining and more personal: weddings, corporate parties, birthdays, team buildings, presentations, matinees, parties, company openings, festivals and fairs, exhibitions and press conferences, social events, private events that also include commemoration and burial ceremonies.

During the project development, programming languages, application frameworks, and cloud database such as JavaScript, Vue, React, Android Studio, Firebase were used. The characteristics and capabilities of all listed languages were given. Also, there is their functionality and ability to realize the set of tasks.

Since all strengths and drawbacks of the project were unknown, it was difficult to pick a programming language which would best suit our objectives. We've studied a number of relevant programming languages, starting from basics of C++ and ending with the PHP environment. It was decided to choose a language based on its relevance and compatibility with Firebase, since after some searching we found out that some languages have far more information about firebase integration than others. In solving the set of goals concerning information system, the following tasks were accomplished: the analysis on the relevance of the problem and the comparison of the world-wide and local analogues of the product.

As a result of the project, all important requirements were fulfilled, and the main thesis work

objectives were fully implemented. In addition, the developing tools of the information system were discussed in terms of program developers and programming languages.

Overall, all of the project's features have been incorporated to help users. This project's output can benefit a large number of people. It has generated a large amount of documentation that will be useful in the future. Even if this system has flaws, it can be improved in the future. Finally, all event-related data may be handled and viewed. As a result, consumers will not miss an event that they are interested in, and organizers will be able to submit plans and track their status.

## 7. References

- [1] Anuarbekova A. (2021) “Trends and Event-industry challenges”. URL: <https://businessmir.kz/2021/02/25/trendy-i-vyzovy-event-industrii/>.
- [2] Kairatova, G. K. “Role of event management in the economic development in Kazakhstan” in “Young Scientist”. — 2016. — № 25 (129). — С. 284-288. URL: <https://moluch.ru/archive/129/35791/>.
- [3] Nurmakov A. (2016) Kazakhstan and the Global Industry of Mega Events: A Case of Autocratic DManagement. In: Makarychev A., Yatsyk A. (eds) Mega Events in Post-Soviet Eurasia. Mega Event Planning. Palgrave Macmillan, New York. URL: [https://doi.org/10.1057/978-1-137-49095-7\\_6](https://doi.org/10.1057/978-1-137-49095-7_6).
- [4] Kemal Birdir, Sevda Birdir, Ali Dalgic and Derya Toksoz. (2020). “Impact of ICTs on Event Management and Marketing”, p. 357.
- [5] R. Waring. (2014). “The aisles are a changing Irish Marketing Journal”, p. 40.
- [6] Oxford: Meyer & Meyer. Kang, J., Tang, L. and Fiore, A. M. (2014). “Enhancing consumer-brand relationships on restaurant Facebook fan pages: Maximizing consumer benefits and increasing active participation. International Journal of Event Management”, pp. 145-155.
- [7] L. Smit. (2020). “Event Management: Putting theory into practice”, [blog].
- [8] TOGETHER survey. URL: <https://docs.google.com/forms/d/e/1FAIpQLSc3FCEk1kwfhCtY1Niv9ZUkIM727LbZq8AH5EIDXaFhEHGigg/viewform>
- [9] Use Firebase in Progressive Web Apps. URL: <https://firebase.google.com/docs/web/pwa>.
- [10] Evan You (2022) The Vue Instance. URL: <https://v2.vuejs.org/v2/guide/instance.html>.
- [11] Evan You (2022) Composition API: setup(). URL: <https://vuejs.org/api/composition-api-setup.html>.
- [12] Add Firebase to your Apple project. URL: <https://firebase.google.com/docs/ios/setup>.
- [13] Firebase JSON Config. URL: <https://rnfirebase.io/app/json-config>.
- [14] Artem Diashkin (2020) Firebase Basics with React in Examples. URL: <https://medium.com/litslink/react-js-firebase-basics-in-examples-cfc980e6b144>.
- [15] Connect to Firebase. URL: <https://developer.android.com/studio/write/firebase>.
- [16] Public class Query in Firebase. URL: <https://firebase.google.com/docs/reference/android/com/google/firebase/firestore/Query>.
- [17] Letham Benjamin, Rudin Cynthia and Madigan, David (2013) Sequential event prediction. URL: <https://dspace.mit.edu/handle/1721.1/88080>.
- [18] Anna Pastushko (2021) Forecasting of periodic events with ML. URL: <https://towardsdatascience.com/forecasting-of-periodic-events-with-ml-5081db493c46>.
- [19] What we know about event technology and machine-learning. URL: <https://blog.impactpointgroup.com/what-we-know-about-event-technology-and-machine-learning>.
- [20] Event Management Guide 2022. URL: <https://www.eventleaf.com/event-management/event-management-guide>.
- [21] Integrating Artificial Intelligence into Event Management. URL: <https://cleproductions.com/integrating-artificial-intelligence-into-event-management/>.
- [22] Prince Canuma (2022) How to Deal With Imbalanced Classification and Regression Data. URL: <https://neptune.ai/blog/how-to-deal-with-imbalanced-classification-and-regression-data>.