**Final Project CSS328**

**Made by: Yermekuly Gaziz 220107028**

**Mukashev Sultan 220107032**

**Data collected from:** [**https://gender.stat.gov.kz/page/frontend/detail?id=58&slug=-47&cat\_id=3&lang=ru**](https://gender.stat.gov.kz/page/frontend/detail?id=58&slug=-47&cat_id=3&lang=ru)

*“Adult mortality, by cause of death and age group, by gender”*

*Raw data:*[death\_rate](https://docs.google.com/spreadsheets/d/1A0PseAmxW9F8NM4dyqycnF6WrHiJrZiiyk1NcYOUiVQ/edit?usp=sharing)

So, in the raw version we have different sheets, each having mortality of people by age categories and corresponding year divided by gender and address type(urban, rural) for each type of disease.

Considering this, we took each gender/address type as 0’s and 1’s, disease type as numbers in range from 0 to 17, and rewrote it as a new table using some excel formulas and queries.

*Data after our manipulations:* [tt](https://docs.google.com/spreadsheets/d/18KmGkVGSqmw67ftufQCFe-qTRJGJ2H2cH-GvuoTLIv8/edit?usp=sharing)

Now we have the following six columns: Death\_rate(per 100 000), Year, Age category, gender, address type and disease type; 27 000+ rows with these values.

**Data cleaning and Visualisation:**

We did not have any Nan values, so no need for cleaning it. But we had to replace some duplicate age categories as “85&” and “85 &” to avoid confusion.

**Preprocessing:**

* Since we have label values for age categories, we mapped all of them to corresponding indexes in list of all unique categories
* Normalization of all data was done by using MinMaxScaler.
* For using RandomForestClassifier to predict Disease types, this column was changed to categorical data instead of numerical continuous.

**Visualisations**

Can be found in our google collaboratory notebook: [finalproject.ipynb](https://colab.research.google.com/drive/10vUlTILuGUStYZ4TdM93YvEUDj0L0Nz9?usp=sharing)

**Models and Goals:**

1. **RandomForestRegressor** model to predict mortality based on other information(year, disease, gender, address type and age category).

**Results:**

Mean Absolute Error: 21.375153399150907

Mean Absolute Squared Error: 7787.508194300114

R-squared Score: 99.1468351286675

1. **RandomForestClassifier** model to predict disease type based on other features.

**Results:**

Accuracy: 0.34746676248652536

precision recall f1-score support

0 0.67 0.66 0.67 340

1 0.59 0.75 0.66 315

2 0.10 0.12 0.11 313

3 0.07 0.07 0.07 340

4 0.36 0.33 0.34 324

5 0.28 0.22 0.24 324

6 0.49 0.72 0.58 596

7 0.04 0.03 0.04 319

8 0.68 0.66 0.67 349

9 0.30 0.31 0.30 310

10 0.25 0.24 0.24 357

11 0.30 0.30 0.30 423

12 0.21 0.18 0.20 314

13 0.35 0.26 0.30 371

14 0.21 0.18 0.19 260

15 0.21 0.19 0.20 311

accuracy 0.35 5566

macro avg 0.32 0.33 0.32 5566

weighted avg 0.33 0.35 0.34 5566

1. **GradientBoostingRegressor** model to predict mortality based on other features.

**Results:**

R-squared Score: 0.8415643766247398

Tweedie Deviance Score: -7333134.542260253

1. **ARIMA** model for Time Series. Prediction of death rate for next n years.

* Overall
* Only for Disease = 6 (Vascular damage to the brain)

**Results:**

Forecasted Death Rates for the next 3 years:

27827 203.924257

27828 176.738390

27829 165.200811

Forecasted Death Rates for the next 3 years (DIsease = 6):

1656 209.103847

1657 182.335483

1658 169.406680

**About used metrics:**

In case of Tweedie deviance a lower (more negative) generally indicates a better fit of the model to the data.

In other cases, closer value to 1, indicates a better fit of the model to the data.

**Conclusion:**

We had a dataset with mortality of people from different diseases and age groups in Kazakhstan for each year from 2000 to 2022, and based on it we made a variety of machine learning models to make predictions of some of the present features.

One of the main goals was making a model to predict mortality of exactly known disease by year, gender, address type and age group. So we used a kind of regression model for it: RandomForest and GradientBoosting. Each of them performed relatively great on testing data, but RandomForest model seems to be overfitted with an R2 score of 99.14%.

Another approach was trying to predict disease type by other features by using RandomForest classifier model, but results were not so good, probably because parameters for each disease are not directly different, and the model assumes them as one.

Final and the most interesting part was applying the ARIMA (AutoRegressive Integrated Moving Average) model for analyzing Time series and making predictions about future, non-existing data. We could use it in actual business cases so I consider it a very important approach.

Overall, after finishing this project work, we greatly upgraded our knowledge about actual application of machine learning models and used all tools that we learnt from the Introduction to Data Analysis course, which was very helpful.

**Data explanation:**

*Disease types:*

1. Diseases of the circulatory system
2. Hypertension
3. Coronary heart disease
4. Angina pectoris
5. Acute myocardial infarction
6. Vascular damage to the brain
7. Neoplasms
8. Malignant neoplasms
9. Accidents, poisoning and injuries
10. Murders
11. Suicides
12. Transport accidents
13. Respiratory diseases
14. Influenza, acute respiratory infections and pneumonia
15. Diseases of the digestive system
16. Infectious and parasitic diseases

*Genders/Sex:*

0. Male

1. Female

*Address type:*

0. Urban

1. Rural

*Age categories:*

0. 0-4

1. 5-9
2. 10-14
3. 15-19
4. 20-24
5. 25-29…

Age category for every 5 years starts with 0 till 85+