# Project Group Name: Verispor

1st Erkam ÇETKİN
*Management Information Systems*
*Kadir Has University*
İstanbul, Turkey
erkamcetkin@stu.khas.edu.tr

2nd Meltem KARABASTIK
*Management Information Systems*
*Kadir Has University*
İstanbul, Turkey
meltemkarabastik@stu.khas.edu.tr

3rd Helin Kuş
*Management Information Systems*
*Kadir Has University*
İstanbul, Turkey
helinkus@stu.khas.edu.tr

4th Fatih Mehmet Alagöz
*Management Information Systems*
*Kadir Has University*
İstanbul, Turkey
fatihmehmet.alagoz@stu.khas.edu.tr

5th Uğur Kılıçdoğan
*Management Information Systems*
*Kadir Has University*
İstanbul, Turkey
ugur.kilicdogan@stu.khas.edu.tr

6th Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

***Abstract*—This data science project aims to predict the target variable based on feature variables. Our aim to build good machine learning using data science techniques. We prepared our dataset for machine learning by using data preprocessing steps. Our data preprocessing steps are Data Cleaning, Data encoding and Feature selection,Visualization Methods, Data Splitting. Finally, we built a model by using machine learning techniques and evaluated our model. Moreover, we evaluated the performance of the model we created using 3 different algorithms in machine learning and we obtained 3 different accuracy scores.**

***Index Terms*—target variable, Data Cleaning, Data encoding and Feature selection,Visualization Methods, Data Splitting, machine learning techniques, data preprocessing steps**

## I. Introduction

this project aims to build good machine learning using data science techniques. The main purpose of this project is to make the data suitable for machine learning by using data Preprocessing techniques and to achieve high accuracy results. Commonly used preprocessing steps are data cleaning, data integration, data transformation, feature selection, data encoding, handling outliers, data splitting.

## II. Data Preprocessing

Data preprocessing is one of the most important steps in data science. It includes the preparation and cleaning of raw data before it is used for analysis and modelling. in other words, it is the improvement studies on the data set. We used data preprocessing techniques to improve the quality of the data and ensure the accuracy of our model. Moreover, According to an article ”Typically, real-world data is incomplete, inconsistent, inaccurate (contains errors or outliers), and often lacks specific attribute values/trends. This is where data preprocessing enters the scenario – it helps to clean, format, and organize the raw data, thereby making it ready-to-go for Machine Learning models” [1].

### *A. Data Cleaning*

Nowadays, raw data may contain errors or missing values. Missing values reduce data quality and may adversely affect its analysis. Thanks to data cleaning, we ensure the accuracy of the data set and the improvement of the model performance. Firstly we struggled with editing the missing data in the dataset. We use pipeline method because There were too many missing data from many columns in the data set. Thanks to the pipeline, all codes appear sequentially on a single line, reducing the risk of errors. In addition, we can get more results with less code. We filled these missing data according to the type of each column. For example, for numeric values, we filled in the missing values by averaging, and for categorical variables, we filled in the missing data by taking the mode. In data science, the mean or median for numeric values is often filled with missing data to fill in missing data because the missing values are close to the mean. we fill in the missing data with the mean so that machine learning model is not affected by missing data. on the other hand, for categorical values, missing data is filled with the mode. Mode represents the most common or most frequently used value of a categorical variable. therefore, we used mode for categorical variables so that our machine learning model is not affected by missing values.

### *B. Data encoding and Feature selection*

In this step, we divided the data into 3 different categories according to their types as continuous, binary, and categorical. By categorizing the data, we can apply to appropriate data preprocessing techniques for each data type. For example, for binary variables that include yes-no, male-female we can encode them as 0s and 1s and it is a important preprocessing technique for machine learning. After that, we apply to data encoding that is a data preprocessing method. We encoded categorical values as numeric labels to make the dataset ready before machine learning. Most machine learning algorithms work with numerical values because direct use of categorical values is not in a format that machine learning algorithms can understand. The reason of this machine learning algorithms require numerical data in order to perform mathematical operations and calculations. Moreover, we examined the correlation matrix as another data preprocessing step. Firstly, We filtered the columns greater than 0.8 and less than -0.8 and visualized them to see the relationship between them. In machine learning, highly correlated pairs can cause multicollinearity problems. So, We Identify Highly Correlated Feature Pairs and extracted one feature from each highly correlated pair to eliminate redundancy. So, by preventing overfitting we made our dataset more suitable for machine learning.

### *C. Visualization Methods*

Another of the most important data preprocessing steps is data visualization. After doing the data cleaning, data integration, data transformation, future selection and data encoding preprocessing steps, we finally applied to the data visualization step. we appleid to data visualization because before machine learning we wanted to make sure the data is clean, fit and see visually if there are any errors. We use Histograms for Continuous Features because histogram provide insights into the range, central tendency, and spread of the data. They show the distribution of values for each continuous feature. So histograms help us understand the characteristics of the data and enable us to make the right decisions in data analysis and modelling. Furthermore, we use count plots for categorical features. They show the distribution of each category. Using these visualizations before machine learning, we can better understand data and review errors for the last time

### *D. Data Splitting*

Before starting machine learning, another important data preprocessing step is data splitting. The purpose of data splitting is to evaluate the performance of a machine learning model on unseen data. We divided the data as training and test set and created our model on the training set and evaluated its performance on the test set. By doing this, we get an estimate of how the model we created performs on the new data. In addition, the model we created may perform well on the training set, but may perform poorly on new data. This indicates that overfitting can exist.

## III. Machine learning models

We used some machine learning algorithms to predict the target variable based on our features. These are Logistic Regression, Logistic Regression with Standard Scaling and Random Forest Classifier. We used 3 different machine learning algorithms in our project because to compare performances, try different methods and make sure their accuracy.

### *A. Logistic Regression*

we used logistic regression as machine learning in our project. Because we are trying to find out how well our machine learning fits the real data, and since our target column is binary data. So, we used the logistic regression model, which is a classification algorithm. Finally, we trained the Logistic Regression model on the training data and evaluated performance. By using logistic regression model, Our accuracy score is ”0.9192”. Accuracy score indicates how well the model performs in predicting the target variable

### *B. Logistic Regression with Standard Scaling*

We applied standard scaling to improve the performance of the our Logistic Regression model. We used standard scaling because we ensure that all features were at the same scale. Thus, we ensured that the performance of our model could be better. The model was trained on the scaled training data and The performance of the model was evaluated using the accuracy score. By using Standard Scaling our accuracy score is 0.9193 and it shows an improvement over the previous model.

### *C. Random Forest Classifier*

Finally, We applied random forest classifier machine learning algorithm in our project. Using this algorithm we wanted our accuracy value to be slightly better. Random Forests can reduce overfitting and improve generalization performance. By using Random Forest Classifier our accuracy score is 0.9193 we can say Using the Random Forest Classifier algorithm is an effective method when predicting the target variable.

## IV. Challenges and solutions

Firstly, the fact that the columns in the dataset are not named and not knowing the logic of the dataset caused us to not fully understand the project. One of the most important difficulties we faced in our project was dealing with missing data. there was too much missing data in the dataset. But, we overcame this challenge by using the pipeline method, spending less time and code. Another challenge we faced in our project was feature selection and deletion of not useful features. Because there are so many features, it took us a while to figure out and find what features were needed. Thanks to the correlation analysis, we figured out which columns we should delete and thus we deleted the features that were not useful in the dataset. Additionally, we preserved the useful features in the dataset. By doing this, we also prevented overfitting.

## V. Conclusion and future work

As a result, we examined a complex and ambiguous dataset with multiple columns and missing data. We tried to predict the target variable by looking at the features presented to us. We used important data preprocessing steps; Data Cleaning, Data encoding and Feature selection,Visualization Methods, Data Splitting. Thus, we made the dataset ready for machine learning algorithms. Finally, we viewed the accuracy results of our predictions with machine learning. As a result, this project successfully applied data science techniques and machine learning models to analyze and predict the target variable.

## VI. Figures and Tables

The machine learning algorithms we used in our project are: Logistic Regression, Logistic Regression with Standard Scaling and Random Forest Classifier. We tabulated machine learning results with all accuracy scores and have shown the accuracy scores we got from different machine learning algorithms in table 1. The machine learning algorithms we used in our project are:

TABLE I
Accuracy Scores of Machine Learning Algorithms

| Model | Accuracy |
|---|---|
| Logistic Regression | 0.919 |
| Logistic Regression with Standard Scaling | 0.932 |
| Random Forest Classifier | 0.943 |

## References

[1] 1. UpGrad. ”Data Preprocessing in Machine Learning.” Available online: https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/ (Accessed on [18.05.2023]).