TITANIC CHALLENGE REPORT

This code performs data exploration and pre-processing on the Titanic dataset. The dataset includes travel information and survival status of passengers. While writing this code, I had the opportunity to apply many of the things I learned in data visualization, data pre-processing, data manipulation, and data analysis.

Libraries used when writing the code. The "pandas" library is used for data manipulation and analysis. The "numpy" library is used for numerical computation and data manipulation. The "seaborn" library is used for data visualization operations. The library "matplotlib.pyplot" is used for plotting graphics. The "sklearn" library is used for machine learning modelling.

When we look at the working logic of the code, first the datasets are read from the "train.csv" and "test.csv" files. Then the "barPlot()" function is used, which can be used to plot a bar chart for any categorical feature in the dataset. Likewise, the function "plotHist()" is used, which can be used to plot a histogram for any numerical feature.

Some categorical features in the dataset e.g. gender, class, port etc. Some inquiries are made to examine the effect on survival rates. Also, using the "heatmap()" function, the temperature of correlation between features is displayed.

The "detect_outliers()" function is used to detect outliers in the data set. This function calculates the indices of outliers for each of the specified features in the dataset and returns them in a list.

Next, "Header" is created and removed from the "Name" property. Header contains the titles of the passengers and this attribute is used to estimate the age attribute. Also, the "Sex" attribute is encoded as "female" and "male".

Then an array named "aver_ages" is defined to replace the missing data. This series calculates the average age of passengers based on gender and class. Then, missing age values are replaced with these averages and the age feature is completed.

The final version of the dataset is in the form of a table with 10 features and 891 data samples. This dataset can be used for training machine learning models. For example, a classification model can predict the survival of passengers, or a regression model can predict the ages of passengers.

Finally, it is used to save the submissionto.csv command as a CSV file with the name "submission.csv" in the current working directory. The "index=False" parameter ensures that the DataFrame's index column is not included in the saved CSV file.

Finally, I uploaded the csv file Kaggle website and got 0.76315 points.

Halil İbrahim Özdemir-200254076