Patient Survival Prediction

*Interim Report

1st Rana Demir
Artificial Intelligence Engineering
Tobb ETU
Ankara, Turkey
ranademir@etu.edu.tr

2nd Zeynep Meriç Aşık
Artificial Intelligence Engineering
Tobb ETU
Ankara, Turkey
zeynepmericasik@etu.edu.tr

***Abstract*—This document is a pre-report for YAP470 Machine Learning course project. The subject of the project is the prediction of patient survival status by using machine learning. Different pre-processing techniques, models are used to obtain the most accurate prediction.**

I. Introduction

In this project, the problem that will be focused on is to predict survival state of a patient based on their health conditions. Estimating decease may prevent many negative outcomes. Looking through the diseases or other conditions a patient might have, early precautions can be taken to save lives. There are many researches on this issue since machine learning has started to be used in health care in the last few decades. With this project, it is aimed to explore the most important blood matters or illnesses which affects a patient passing away and predict it beforehand by using a machine learning model. The approach that will be used is to train the data with several models and detect the best one for this case and for the dataset which is chosen previously.

II. Problem Statement

In order to provide health care throughout the world a useful method to predict decease before it happens, a machine learning model will be trained. The dataset [7] that will be used in this project contains patients' health history who are hospitalized in detail. The general information of patients such as gender, age and height, also all of the illnesses and blood particular matter amount they have are listed such as diabetes, aids, cirrhosis, tumour, lymphoma; glucose, potassium, oxygen saturation, blood and arterial pressure. It is expected from the result to give a classified number , 0 or 1, which is 0 if the patient survives, 1 if they do not for each person hospitalized. To obtain these results, the most contributing features will be listed (feature selection is mentioned in technical approach) and will be used to train the chosen models. After training, the test results will be compared to the actual results of hospital death column in the dataset. Then, the outcomes will be evaluated with different metrics. The metrics that will most probably be used are accuracy, precision, recall, f1-score and ROC-AUC; in addition, Gini impurity index and entropy will be taken into consideration. In the end of evaluation, the model which gives the best results will be selected as the final model. From this model, it is anticipated to obtain the highest results which is expected to be greater then 90 percent but the expectations might be alternated through the experiments.

III. Technical Approach

The first step to follow is to prepare the data so that it can be used to do training and testing. In order to succeed this, the data is examined both visually and verbally. The unnecessary features which are the ones that contain single values for all samples of the data are eliminated since it does not provide any useful information to make a prediction. Data is mainly examined in two parts as categorical and numerical features. For the numerical part, the correlation between all numerical features is computed and highly correlated ones are eliminated. This process leaves only one feature from the highly correlated ones so that instead of multiple features, only one feature can be used since it represents the others. After this step, the rows that contain null values are eliminated. However, if the representativity of the data is also lost, this process will not be executed. Controlling of the representativity will be considered in the further processes. If this option seems to be not working, another way to eliminate null values is deleting features with null values. If this also does not work, filling the null values with filler methods will be executed. For the categorical part, chi squared score will be computed and according to these scores, feature selection will be applied. On the other hand, for both categorical and numerical features, up-sampling with SMOTE or down-sampling will be considered to apply on the data in a situation that data having imbalance over it. For the same case, bootstrapping can be used. After obtaining the data that will be useful for the prediction, scaling will be applied since the data has high variances in some features. Both Standard and Min Max [1] Scaler will be considered. The categorical values will be transformed to numerical values by encoding [2]. Both Label Encoder and One Hot Encoder will be applied and one of them will be used according to their evaluation results, if possible. Up to this point, pre-processing steps are taken. After all of these, data will be divided into two parts. Possible proportions are 80 and 20 percent, 75 and 25 percent for train and test respectively. Any variation close to these percentages can also be used. Cross validation will be considered with number of folds being 5

or 10. Probable models that are going to be tried for the pre-processed data is planned to be Random Forest [3], Support Vector Machine [4], Stratified Logistic Regression [5] and ensemble models containing some of these. In the case of evaluation values being lower than what is expected, other models that are not mentioned here can be used to see different results. Fitting operation is applied based on the chosen model. Then, prediction is made for each model. Evaluation metrics mentioned previously are used to evaluate the model. Then, all values are examined.

IV. Intermediate/preliminary Results

At this stage of the project, the exploratory data analysis and the pre-processing of the data has been done. The results has shown that the dataset that is chosen contains many null values, object types and highly correlated features that has to be covered. Whilst analysing the data through plots, each feature in the categorical data is examined by count plots and numerical data is examined by bar plots. It is visible that some of the features are not necessary since they serve to the same view which can be decided by looking into their correlations with each other. Also, a set of features which are efficient for the final results can be revealed through the findings of the relation with the target 'hospital death'. In the pre-processing part, the dataset has been set to appropriate for testing and training by dropping columns, scaling values, and other operations to fill in the missing data. As a result, it can be seen that only less than three quarters of the data is necessary to execute an optimized prediction process. Hence, this is planned to be used and the rest of the model will be generated with the steps mentioned in technical approach.

References

[1] https://www.mdpi.com/2227-7080/9/3/52
[2] https://www.diva-portal.org/smash/record.jsf?pid=diva2
[3] https://ieeexplore.ieee.org/document/9817303
[4] https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5
[5] https://www.jstor.org/stable/2532141
[6] https://www.nature.com/articles/s41598-020-77220-w
[7] https://www.kaggle.com/datasets/mitishaagarwal/patient