CAPSTONE PROJECT: PROPOSAL

Mohamad Fadil Bin Mohamad Parves
(Dated: 16 May 2021)

I. DOMAIN BACKGROUND

The excessive changes in the lifestyles of inhabitants across the world through the last decades has moved the human societies from farming foods and active lives into fast foods and inactive lifestyles. The grouping of such lifestyle with growing cigarette consuming has improved the risk factors of cardiovascular diseases (CVDs).
Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. [1]
Heart failure is a common event caused by CVDs. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.
People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

II. PROBLEM STATEMENT

Healthcare is an inevitable task to be done in human life. Cardiovascular disease is a broad category for a range of diseases that are affecting heart and blood vessels. [2] The health care industry has a lot of data; hence this data should be used with machine learning algorithms in order to enhance the probability of reducing patients' risk by predicting ahead if a patient is more likely to be attacked by a specific illness or not based on their lifestyle and health records data. Hence, the problem to be solved *is to predict if a patient is going to decease in the following follow up session.*

III. DATASETS AND INPUTS

The data that will be used for this project is gathered from Kaggle [3]. This data is shared by Davide Chicco, Giuseppe Jurman in their paper which is titled Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. [1]
By using features such as age, decrease of red blood cells (anemia), level of CPK enzyme (creatinine phosphokinase), if the patient has diabetes, percentage of blood leaving the heart at each pump, if patient has hypertension, platelets in the blood, level of serum creatinine, level of serum sodium, sex, does the patient smokes, follow up period we will be able to predict either the patient will be deceased by the next follow up session or not. The reason for using this data is as it is processed data and complete for the use case.

IV. SOLUTION STATEMENT

Classification is a common practice in machine learning which is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. [4]

TABLE 1. Complete description of dataset fields

Name	Description
age	Age of the patient
anaemia	Decrease of red blood cells
creatinine_phosphokinase	Level of the CPK enzyme in the blood
diabetes	If the patient has diabetes
ejection_fraction	Percentage of blood leaving the heart at each contraction
high_blood_pressure	If the patient has hypertension
platelets	Platelets in the blood (kiloplatelets/mL)
serum_creatinine	Level of serum creatinine in the blood (mg/dL)
serum_sodium	Level of serum sodium in the blood
sex	Woman or man
smoking	If the patient smokes or not
time	Follow up period (days)

And the predicted value is either the patient survives until the next follow-up date or not. By defining the target variable, we can confirm that this is a classification problem where we need to get a 0 and 1 output for each patient where 0 means the patient survived and 1 means the patient deceased.

For this project, we will evaluate multiple machine learning algorithms such as Support Vector Machine (SVM), Gaussian Naïve Bayes, K-Nearest Neighbors, Random Forest, and Logistic Regression. At the end, comparison will be done on all the algorithms and the best model with highest precision and recall will be chosen. The project will be developed using Python, Pandas, Numpy [5], Scikit [6].

V. BENCHMARK MODEL

Random Forests outperformed all the other methods, by obtaining the top MCC (+0.384), the top accuracy (0.740), and the top ROC AUC (0.800) (Table 4). The Decision Trees obtained the top results on the true positives (sensitivity = 0.532) and on the F1 score (0.554) and was the only classifier able to predict correctly the majority of deceased patients. The linear Support Vector Machines achieved an almost perfect prediction score on the negative elements (specificity = 0.961), but a poor score on the positive elements (sensitivity = 0.072). The Artificial Neural Network perceptron, instead, obtained the top value on the Precision-Recall AUC (0.750). [1]

VI. EVALUATION METRIC

For this project, recall and precision will be used as the target output is imbalanced, using accuracy alone will not give us good idea on how well the model performed.
Precision is attempts to find what proportion of positive identifications was actually correct. [7]

$$Precision = TP \div TP + FP$$

Recall is the attempt to find what proportion of actual positives was identified correctly [7]

$$Recall = TP \div TP + FN$$

By using both of this metrics, it can define how well did the model perform on the validation set and this will give us sense of confidence on predictions.

VII. PROJECT DESIGN

The steps to solve the project will take reference from the one proposed by S.Raschka [8]. The steps are:

I. **Data collection:** This is the step where all the data will be collected as per describe in section 3 above.
II. **Data processing:** This is the step where the data will be processed, manipulated and cleaned.
III. **Data exploring:** In this step the data will be explored by performing explanatory data analysis to understand the correlation and causation in the data
IV. **Training and evaluation:** In this step, the model performance metrics will be set up. Other than that, model trainings will be done and finally all trained models will be evaluated in order to choose the best among them.
V. **Model Tuning:** In this phase, the chosen model algorithm will be tuned for the best performance
VI. **Reiterate:** Keep doing the same thing until the point to where the model can't improve anymore and giving the best output.

REFERENCES

[1] *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone (2020). Accessed 11 May 2021. Available at https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5*

[2] *Prediction of Cardiovascular Disease Using Machine Learning Algorithms (2018). Accessed 11 May 2021. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5304098/*

[3] *Kaggle. Accessed 11 May 2021. Available at https://www.kaggle.com/andrewmvd/heart-failure-clinical-data*

[4] *4 Types of Classification Tasks in Machine Learning (2020). Accessed 14 May 2021. Available at https://machinelearningmastery.com/types-of-classification-in-machine-learning/*

[5] *Numpy. Accessed 14 May 2021. Available at https://numpy.org*

[6] Scikit. Accessed 14 May 2021. Available at https://scikit-learn.org/stable/

[7] *Classification: Precision and Recall. Accessed 14 May 2021. Available at https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall*

[8] *S Raschka. Accessed 14 May 2021. Available at https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html*