**MACHINE LEARNING - MILESTONE 1**

**DAVID GOMEZ CAMARGO**

1. INTRODUCTION

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. Nearly half of all U.S. adults have some type of cardiovascular disease, a percentage that reflects recently updated guidelines for treating high blood pressure, also known as hypertension, that can lead to heart attack, heart failure and stroke.

This project can be used to explore the risk factors of CVDs in adults. The objective would be to understand how certain demographic factors, health behaviors and biological markers affect the development of heart disease. The [dataset](https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas?resource=download) contains information on age, gender, height, weight, blood pressure values, cholesterol levels, glucose levels, smoking habits and alcohol consumption of over 70 thousand individuals. Additionally, it outlines if the person is active or not and if they have any cardiovascular diseases.

1. CORRELATION

User [Amalia Nurkahasanah](https://www.kaggle.com/code/amalianurkahasanah/kardiovaskular-data-mining) searched for correlated attributes whose information contents are partially contained in another attribute (Cardiovascular diseases in this case) since more correlated the attributes are, the more interdependent they are, and the more redundance we get [Figure 1]. Also presented a correlation heatmap making it easy to visualize all 14 attributes and understand its correlation at a glance [Figure 2].

Graphical user interface

Description automatically generated

**Figure 1.** Correlation between CVDs and other features in dataset.

Chart, treemap chart

Description automatically generated with medium confidence

**Figure 2.** Heatmap between features available in dataset.

1. ROC CURVE, VIF AND GRID-SEARCH

On the other hand, user [Bubpha Woratham](https://www.kaggle.com/code/bubphaworatham/640632057-classification) generated a Receiver Operating Characteristic (ROC) curve showing the performance of various classification models at all classification thresholds (Curve plots two parameters: True Positive Rate and False Positive Rate).

Next, the Variance Inflation Factor (VIF) was estimated, which is used to detect the presence of multicollinearity. VIF measured how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related [Figure 3].

Then, a GridSearchCV hyperparameter tunning was performed to determine the optimal values for the different models. Lastly, the VIF value [Figure 4] and ROC graph [Figure 5] was presented for models after hyperparameter tunning.

Graphical user interface, text, application

Description automatically generated

**Figure 3.** Variance Inflation Factor (VIF) comparison.

Graphical user interface, application

Description automatically generated

**Figure 4.** Variance Inflation Factor (VIF) comparison after tunning.

Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated

**Figure 5.** Receiver Operating Characteristic (ROC) curve after tunning.

1. XGB CLASSIFIER

Last of all, user [Muhammad Tayyabb](https://www.kaggle.com/code/muhammadtayyab885/risk-factors-for-cardiovascular-heart-disease) applied eXtreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm. XGBoost applies a better regularization technique to reduce overfitting, and it is one of the differences from the gradient boosting.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Figure 6.** eXtreme Gradient Boosting.

After reviewing all these approaches to this dataset, I think this project can result in a great opportunity to apply multiple classification machine learning algorithms to explore the potential relations between risk factors and cardiovascular disease that can ultimately lead to improved understanding of this serious health issue and design better preventive measures.