**Research Paper**

Introduction:

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels, and they include:

Coronary heart disease – a disease of the blood vessels supplying the heart muscle

Cerebrovascular disease – a disease of the blood vessels supplying the brain

Peripheral arterial disease – a disease of blood vessels supplying the arms and legs

Rheumatic heart disease – damage to the heart muscle and heart valves from rheumatic fever, caused by a bacteria

Congenital heart disease – non-function of heart structure existing at birth

Deep vein thrombosis and pulmonary embolism – blood clots in the leg veins, which can cause blockage and move to the heart, lungs, and various other parts of the body

Many people die annually from CVDs than from any other cause.

An estimated 17.9 million people died from CVDs in 2021, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke

Project goal

Our Project goal is to predict the possibility of a person having cardiovascular disease or not based on various parameters specified in the dataset provided by Svetlana Ulianova on Kaggle.

Preprocessing

The dataset had impurities which had to be rectified using various preprocessing/PCA methods

In our Data Analysis, We will try to analyze to find out the below

\* Missing / Duplicate Values

\* All the Continuous Values

\* Distribution of the Numerical Values

\* Categorial Values / Discrete Values

\* Cardinality of Categorial Values / Discrete Values

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

A picture containing timeline

Description automatically generated

Graphical user interface

Description automatically generated

Chart, pie chart

Description automatically generated

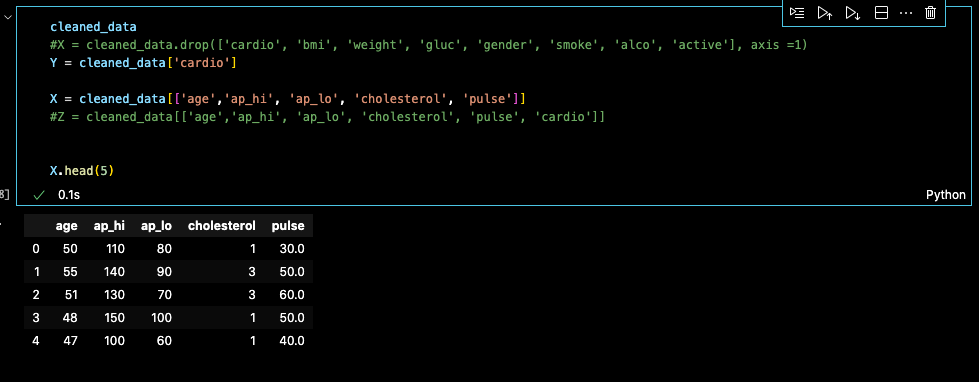
> \* Our Population has the highest no of people having 80 mmHg Diastolic Blood Pressure

> \* If a person is not having CVD, then There's more likely (55.3 %) that he/she has 120 mmHg Diastolic Blood Pressure

> \* If a person is having CVD, then There's more likely (42.5 %) that he/she has 120 mmHg Systolic Blood Pressure with a second most likely case (31.9 %) of having 90mmHg Diastolic Blood Pressure

A screenshot of a computer

Description automatically generated with medium confidence



We ran Logistic Regression, Random Forest, and a Simple neural network on the preprocessed data set.

The conclusion states Simple neural network had the highest accuracy of 0.73 followed by Random Forest at 0.72. The precision and recall were also at 0.72 for RF. We also went ahead and applied NLTK to the output of models. NLTK is a powerful tool to preprocess text data for further analysis like with ML models. It helps convert text into numbers. The output of the NLTK program would pop-up a user input where we need to give all the basic inputs like age, systolic blood pressure, diastolic blood pressure, cholesterol level, and pulse rate, separated by commas. The output would either be 1(Predicted to have cardiovascular disease) or 0 (Predicted not to have cardiovascular disease)

Conclusion:

• Accuracy of prediction for CVD is 73%

• Precision Value of the model is 72%

• Recall is 88%

• Overall F1 score is 72%

Reference:

• Kaggle

• <https://www.kaggle.com/code/sulianova/eda-cardiovascular-data>

• <https://ieeexplore.ieee.org/document/6558288>