**Prediction of Cardiovascular disease using Machine learning and AI**

**Introduction**

Many people die annually from CVDs than from any other cause. An estimated 17.9 million people died from CVDs in 2021, which would be 31% of all global deaths. A total of 85% is due to heart attack and stroke.

Cardiovascular diseases (CVDs) are a group of heart and blood vessel disorders.

Types of heart disease

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

Cerebrovascular disease – blockage of blood vessels supplying the brain

Peripheral arterial disease – blockage 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

**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.

**Literature review**

We could identify a couple of studies that explored the use of ML and AI for CVD prediction.

Studies show the use of data from EHR electronic health record data to train their models. Studies also use traditional machine learning algorithms, while others used deep learning techniques.

One study shows the SVM support vector machine algorithm for the prediction of CVD in patients with type 2 diabetes. Data is used from EHR records and shows higher accuracy of 85%.

Another study used data from EHRs and analyzed using a Deep learning model which also had higher accuracy compared to traditional risk assessment tools. The main parameter used was hypertension.

Random forest algorithm has also been used to predict the risk of CVD in the general population. Random Forest has the highest accuracy report compared to any other ML model.

**Discussion**

ML and AI show promising results in the prediction of CVD in many of the studies. The implemented techniques have huge potential to improve. This will help people identify a CVD at a very early stage which is a huge impact on living. My project goal is to use important measurable parameters (one can easily measure at home) in predicting CVD which is also integrated with a bot service for a better user experience.

**Preprocessing**

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

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

a: Distribution of the Numerical Values

b: All the Continuous Values

c: Categorial Values / Discrete Values

d: Missing values

e: Duplicate Values

Graphical user interface

Description automatically generated

**Chart, pie chart

Description automatically generated**

The data has the highest number of people having 120 mmHg SBP

\* No CVD, then There's more likely 48.7 % person has 120 mmHg SBP

Chart, pie chart

Description automatically generated

The data has the highest number of people having 80 mmHg DBP

\* No CVD, then There's more likely 55.3 % person has 120 mmHg DBP

Chart

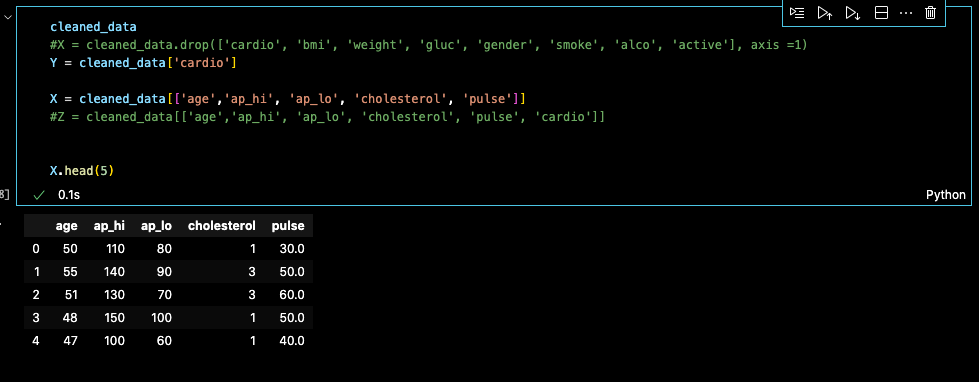
Description automatically generated

**Feature Selection**

A screenshot of a computer

Description automatically generated with medium confidence

Cleaned Data



**Observation**

We ran Logistic Regression, Random Forest, and a Simple neural network on the cleaned 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 of 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)

• 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://jpmm.um.edu.my/index.php/MJCS/article/view/35980>
* <https://ieeexplore.ieee.org/document/6558288>
* <https://link.springer.com/chapter/10.1007/978-981-10-5427-3_63>