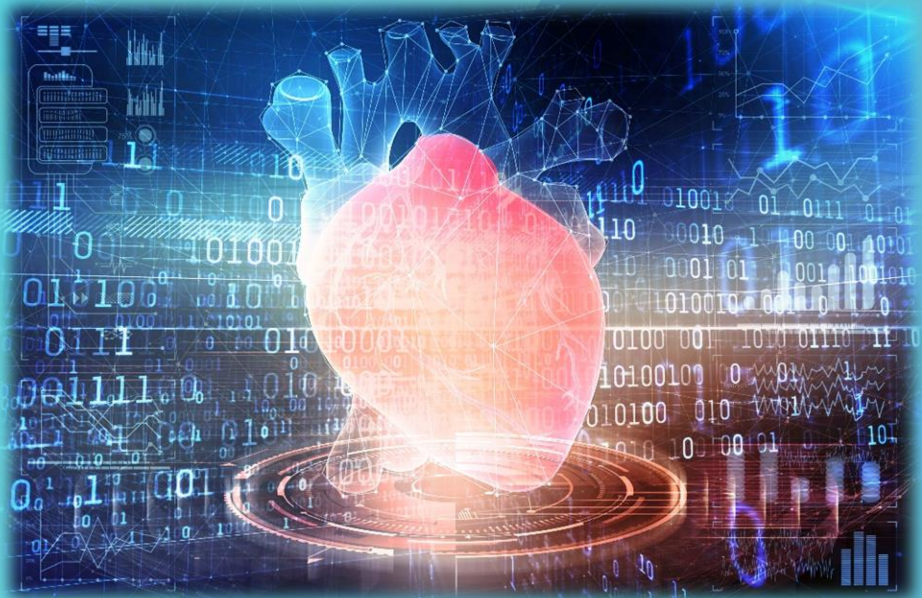
CVD PREDICTION: SUPERVISED LEARNING

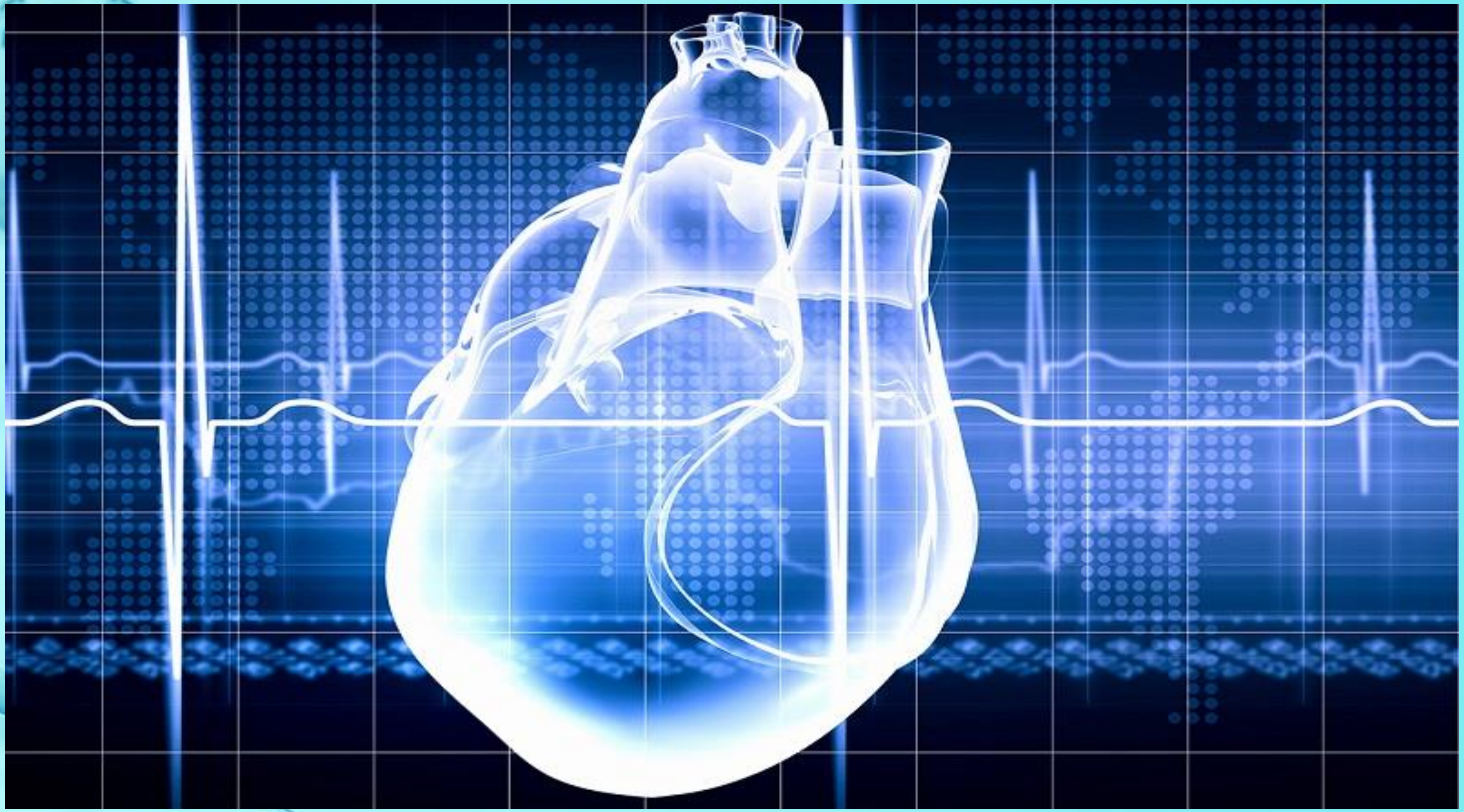


Respected Guide : - Dr. Pandi Murugan

Group members:

DEVANSH RANA - 19BAI10052
KUSHAGRA BAJPAI - 19BAI10053
RASHMI RAWAT - 19BAI10077
ARYAN DAGORE -19BAI10189

INTRODUCTION



In recent years, many achievements have been made in the study of CVD risk prediction model, but the effect of epidemiological risk factors and biomarkers may be different in different populations, the CVD model has certain population specificity. At the same time, a large number of the existing CVD prediction models use multivariable regression method to build prediction models in a linear fashion, but it generally exhibit modest predictive performance, especially for certain sub-populations. Machine learning (ML) such as random forest (RF) can improve the performance of risk predictions by exploiting large data repositories to identify novel risk predictors and more complex interactions between them.

In the studies conducted a CVD prediction model research based on a specific culture, lifestyle, behavior and genetic background. Cardiovascular events were collected through regular follow-up using the electronic health record (EHR) system, and a CVD prediction model for 3-year risk assessment of CVD was constructed using the RF algorithm based on classification and regression tree (CART).

EXISTING WORK AND LIMITATIONS

		TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Classifiers without optimization	K-NN	0.753	0.258	0.785	0.753	0.769	Absence
		0.742	0.247	0.706	0.742	0.724	Presence
	SVM	0.867	0.2	0.844	0.867	0.855	Absence
		0.8	0.133	0.828	0.8	0.814	Presence
	RF	0.847	0.225	0.825	0.847	0.836	Absence
		0.775	0.153	0.802	0.775	0.788	Presence
	NB	0.867	0.2	0.844	0.867	0.855	Absence
		0.8	0.133	0.828	0.8	0.814	Presence
	MLP	0.833	0.192	0.845	0.833	0.839	Absence
		0.808	0.167	0.795	0.808	0.802	Presence
Classifiers optimization by FCBF	K-NN	0.833	0.275	0.791	0.833	0.812	Absence
		0.725	0.167	0.777	0.725	0.75	Presence
	SVM	0.86	0.2	0	0.86	0.851	Absence
		0.8	0.14	0.821	0.8	0.81	Presence
	RF	0.847	0.25	0.809	0.847	0.827	Absence
		0.75	0.153	0.796	0.75	0.773	Presence
	NB	0.873	0.2	0.845	0.873	0.859	Absence
		0.8	0.127	0.835	0.8	0.817	Presence
	MLP	0.887	0.217	0.836	0.887	0.861	Absence
		0.783	0.113	0.847	0.783	0.814	Presence
Classifiers optimization by FCBF, PSO and ACO	K-NN	1	0.008	0.993	1	0.997	Absence
		0.992	0	1	0.992	0.996	Presence
	SVM	0.86	0.192	0.849	0.86	0.854	Absence
		0.808	0.14	0.822	0.808	0.815	Presence
	RF	0.993	0	1	0.993	0.997	Absence
		1	0.007	0.992	1	0.996	Presence
	NB	0.907	0.2	0.85	0.907	0.877	Absence
		0.8	0.0093	0.873	0.8	0.835	Presence
	MLP	0.96	0.15	0.889	0.96	0.923	Absence
		0.85	0.04	0.944	0.85	0.895	Presence

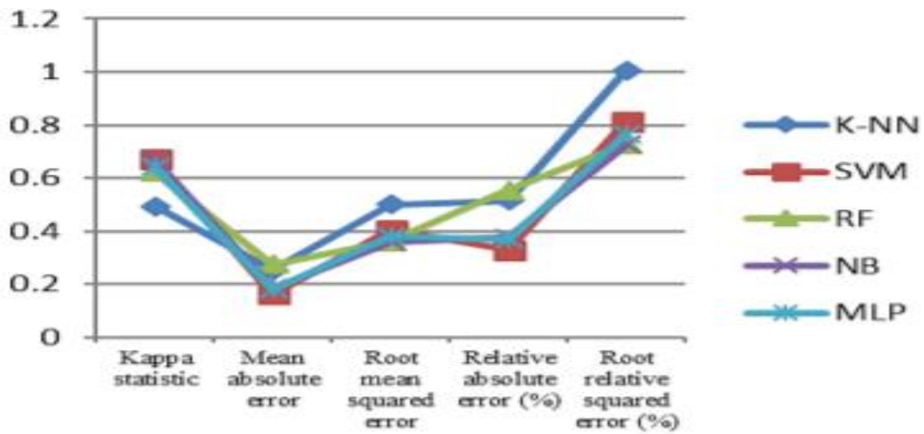


Figure. 5 Simulation error without optimization

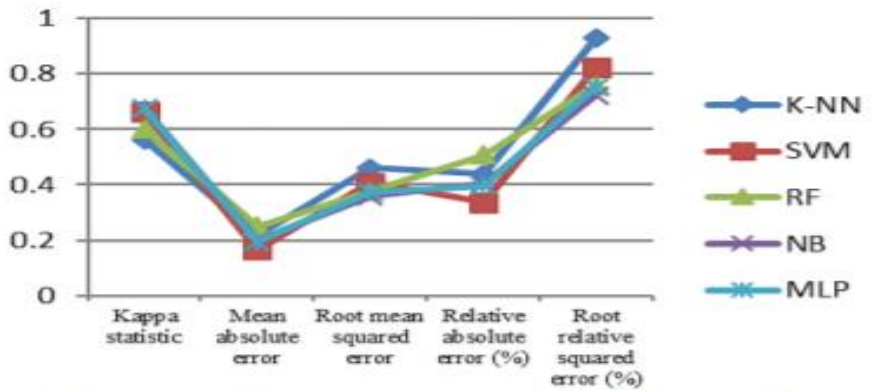


Figure. 6 Simulation error optimized by FCBF

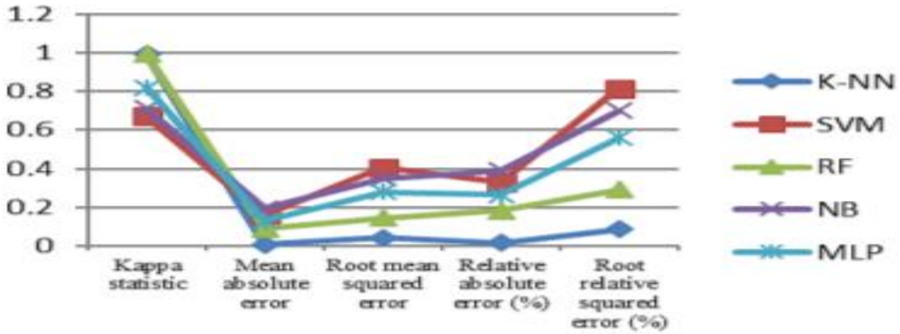
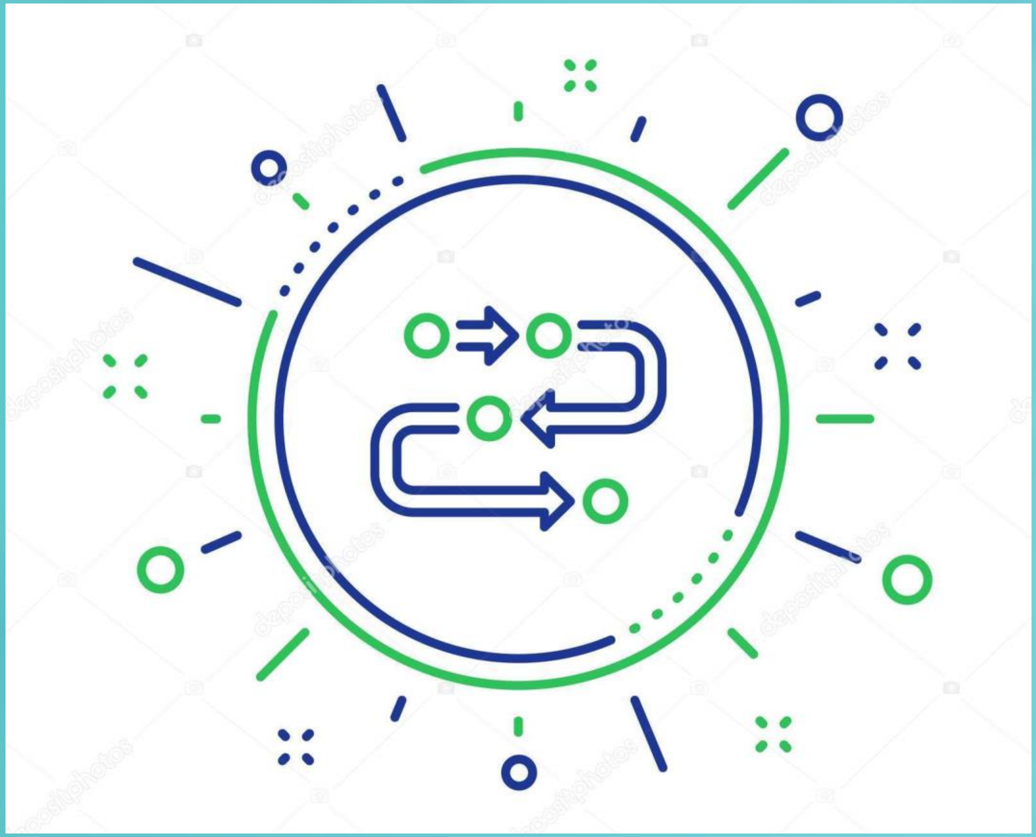


Fig. 7 Simulation error optimized by FCBF, PSO and ACO

Table 11. Performance of different methods

Model	Techniques	Disease	Tool	Accuracy
Otoom et al. [11]	Bayes Net	Heart Disease	WEKA	84.5%
	SVM			84.5%
	Functional Trees			84.5%
Vembandasamy et al. [14]	Naive Bayes	Heart Disease	WEKA	86.419%
Chaurasia et al. [13]	J48	Heart Disease	WEKA	84.35%
	Bagging	Heart Disease	WEKA	85.03%
	SVM	Heart Disease	WEKA	94.60%
Parthiban et al. [12]	Naive Bayes	Heart Disease	WEKA	74%
Tan et al. [10]	Hybrid Technique (GA + SVM)	Heart Disease	LIBSVM+WEKA	84.07%
The proposed optimized model by FCBF, PSO and ACO	K-NN	Heart Disease	WEKA	**99.65 %**
	SVM	Heart Disease	WEKA	83.55%
	RF	Heart Disease	WEKA	**99.6%**
	NB	Heart Disease	WEKA	86.15%
	MLP	Heart Disease	WEKA	**91.65%**

PROPOSED WORK AND METHODOLOGY



Logistic regression

- It is basically a **supervised classification** algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.
- Contrary to popular belief, **logistic regression IS a regression model**. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, **Logistic regression** models the data **using the sigmoid function**.

g(z) = 1/1 + e(-z)**

- Logistic Regression is a **statistical and machine-learning techniques** classifying records of a dataset based on the values of the input fields . It predicts a dependent variable based on one or more set of independent variables to predict outcomes . It can be used both for **binary classification** and **multi-class classification.**

K-nearest neighbors (KNN)

Algorithm is a type of **supervised machine learning algorithm** which can be used for both **classification** as well as **regression predictive** problems. However, it is mainly used for classification predictive problems in industry.

The following **two properties** would define KNN well :

1. **Lazy learning algorithm** − KNN is a lazy learning algorithm because it **does not have a specialized training phase** and uses all the data for training while classification.
2. **On-parametric learning algorithm** − KNN is also a **non-parametric learning algorithm** because it doesn't assume anything about the underlying data.

Support Vector Machine

- Algorithm is a type of **supervised machine learning algorithm.** SVM model is representation of different classes in a **hyperplane in multidimensional space**.
- The **goal of SVM** is to **divide** the **datasets** into **classes** to find a **maximum marginal hyperplane (MMH).**

The followings are **important concepts** in SVM:

1. **Support Vectors** − **Datapoints** that are **closest to the hyperplane** is called **support vectors**. Separating line will be defined with the help of these data points.
2. **Hyperplane** − It is **a decision plane or space** which is **divided** between a **set of objects** having **different classes.**
3. **Margin** − It may be defined as the **gap** between **two lines** on the **closet data points** of **different classes. Large margin** is considered as a **good margin** and **small margin** is considered as a **bad margin**.

Artificial Neural Network

Processing of ANN depends upon the following three building blocks −

1. **Network Topology:** A network topology is the arrangement of a network along with its nodes and connecting lines.
2. **Adjustments of Weights or Learning:** Learning, in artificial neural network, is the method of modifying the weights of connections between the neurons of a specified network. Learning in ANN can be classified into three categories namely supervised learning, unsupervised learning, and reinforcement learning.
3. **Activation Functions:** It may be defined as the extra force or effort applied over the input to obtain an exact output. In ANN, we can also apply activation functions over the input to get the exact output.

Multilayer Perception

- A multilayer perceptron (**MLP**) is a class of **feedforward artificial neural network** (ANN). **MLP** utilizes a supervised **learning** technique called backpropagation for training
- MLP can be viewed as a **logistic regression classifier** where the input is first transformed using a learnt **non-linear transformation** .
- **Non-linear transformation** projects the input data into a space where it becomes **linearly separable**. This intermediate layer is referred to as a **hidden layer**.
- A single hidden layer is sufficient to make MLPs a **universal approximator**. However we will see later on that there are substantial benefits to using many such hidden layers, i.e. the very premise of **deep learning**.

REAL TIME USAGE

Over the last few decades, heart disease is the most common cause of global death. So early detection of heart disease and continuous monitoring can reduce the mortality rate. The exponential growth of data from different sources such as wearable sensor devices used in Internet of Things health monitoring, streaming system and others have been generating an enormous amount of data on a continuous basis. The combination of streaming big data analytics and machine learning is a breakthrough technology that can have a significant impact in healthcare field especially early detection of heart disease. This technology can be more powerful and less expensive.

Cardiovascular disease (CVD) is the leading cause of death worldwide. Early prediction of CVD is urgently important for timely prevention and treatment. Incorporation or modification of new risk factors that have an additional independent prognostic value of existing prediction models is widely used for improving the performance of the prediction models.

HARDWARE & SOFTWARE REQUIREMENT

- **RAM:** A minimum of **16 GB** is required, but I would advise using 32 GB RAM if you can as training any algorithm will require some heavy Lifting. Less than 16 GB can cause problems while Multitasking.

- **CPU:** Processors above **Intel Corei7 7th Generation** is advised as it is more powerful and delivers High Performance.

- **GPU:** This is the most important aspect as Deep Learning, which is a Sub-Field of Machine Learning requires neural networks to work and are computationally expensive. Working on Images or Videos require heavy amounts of Matrix Calculations.

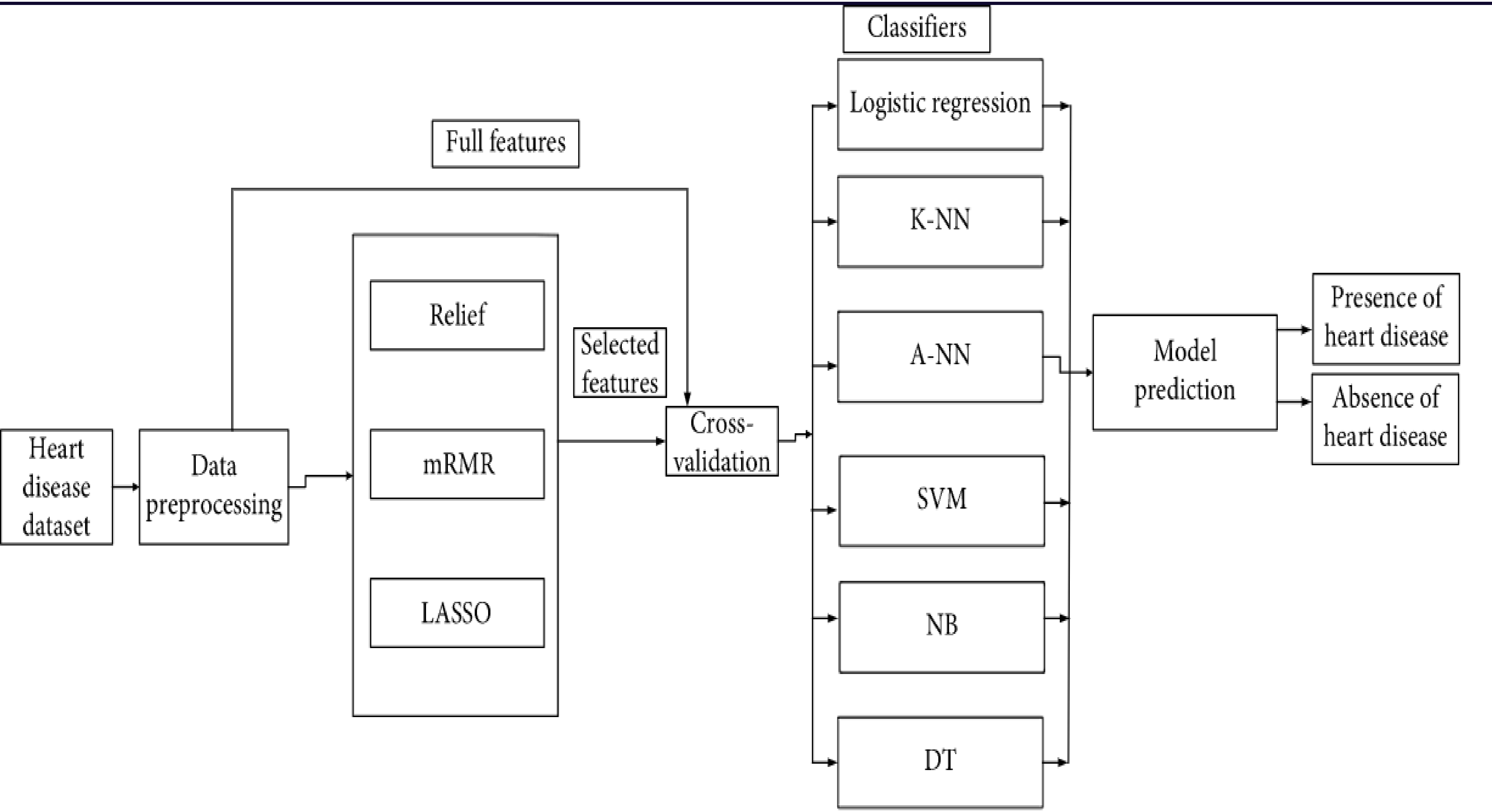
- **Storage:** A minimum of **1TB HDD** is required as the datasets tend to get larger and larger by the day. If you have a system with **SSD** a minimum of **256 GB** is advised. Then again if you have less storage you can opt for Cloud Storage Options. There you can get machines with high GPUs even.

- **Operating System:** Mostly People go for **Linux**, but Windows and MacOS can both run Virtual Linux Environment and you can work on those systems too.

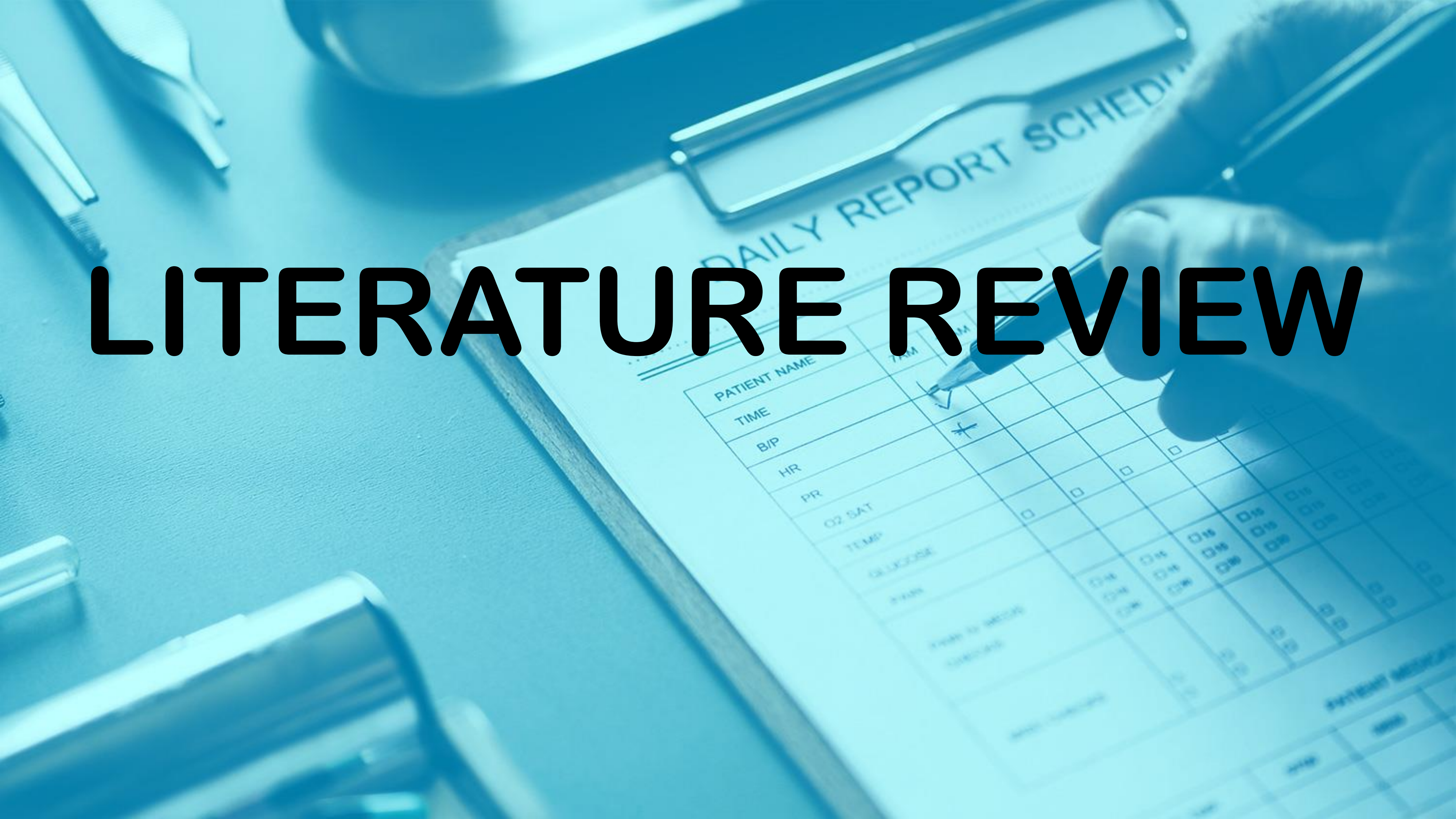
Google Cloud ML Engine

The Google Cloud ML Engine is a hosted platform to run machine learning training jobs and predictions at scale. The service treats these two processes (training and predictions) independently. It is possible to use Google Cloud ML Engine just to train a complex model by leveraging the GPU and TPU infrastructure. With Cloud ML engine, you can train your ML model in the cloud using Google's distributed network of computers. Instead of just using your laptop to train your model, Google will run your training algorithm on multiple computers to speed up the processing

System Architecture Diagram


Classifiers
Full features
Logistic regression
K-NN
Relief
Selected features
A-NN
Model prediction
Presence of heart disease
Absence of heart disease
Heart disease dataset
Data preprocessing
mRMR
Cross-validation
SVM
LASSO
NB
DT

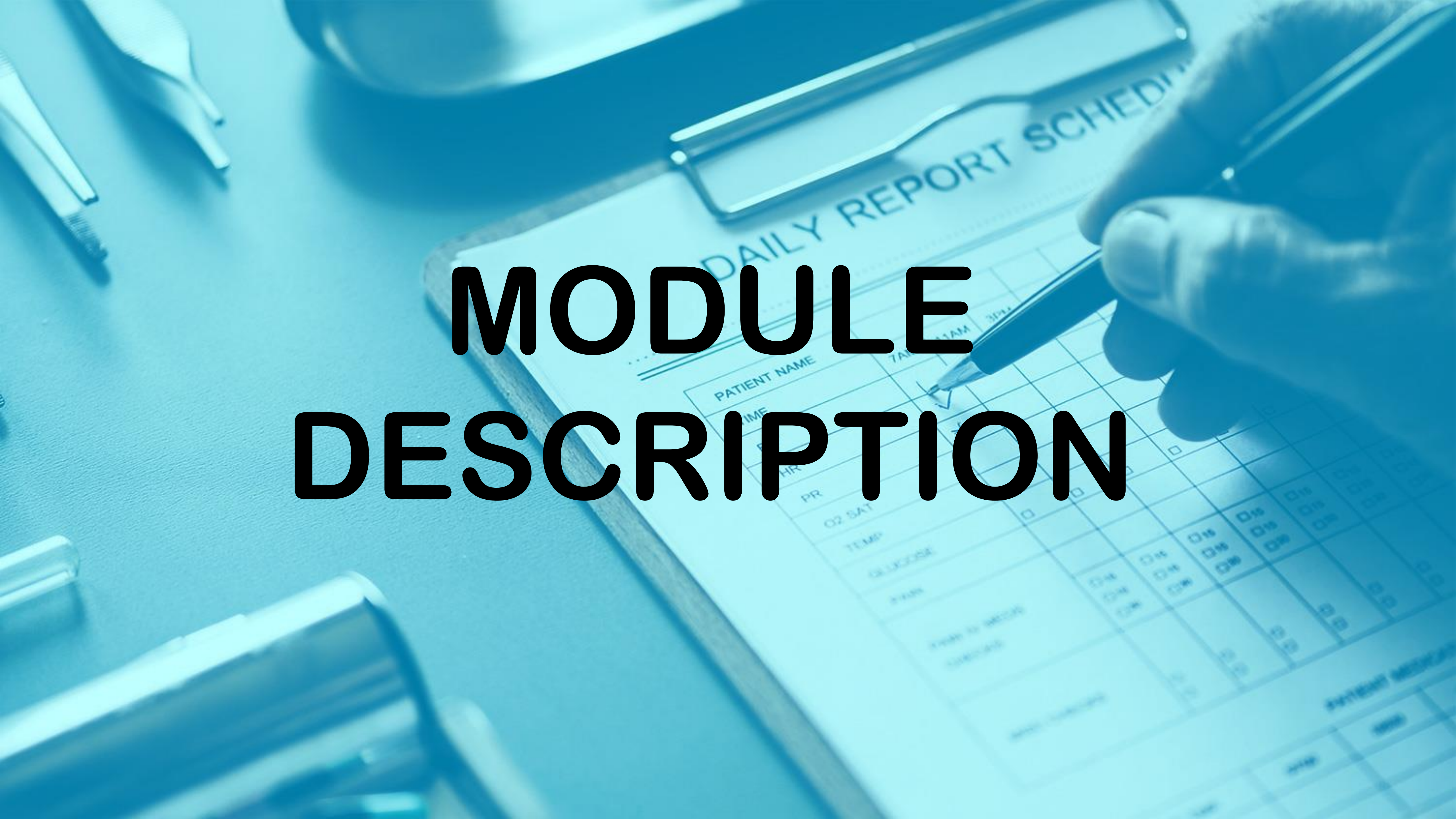
LITERATURE REVIEW



- There is a number of prediction systems proposed for different diseases and implemented using different techniques. Previous works on heart disease with different authors studied and implemented different methods and analyzed the results.
- For the implementation of the work, they have considered the data set from the UCI data repository which can also be collected from the Kaggle.
- The authors performed diseases prediction using the classification, prediction systems, using different data mining techniques and machine learning algorithms in medical centers.
- Few of the results are present below:

YEAR	AUTHOR	PURPOSE	TECHNIQUES USED	ACCURACY
2015	Sharma Purushottam et al,[15]	Efficient Heart Disease Prediction System using Decision Tree.	Decision tree classifier	86.3% for testing phase. 87.3% for training phase.
2015	Boshra Brahmi et al, [20]	Prediction and Diagnosis of Heart Disease by Data Mining Techniques.	J48, Naïve Bayes, KNN, SMO	J48 gives better accuracy than other three techniques.
2015	Sairabi H. Mujawar et al, [24]	Prediction of Heart Disease using Modified K-means and by using	Modified k-means algorithm, naive bayes algorithm.	Heart Disease detection=93%. Heart Disease
2017	Jayami Patel et al,[14]	Heart disease Prediction using Machine Learning and Data mining Technique.	LMT, UCI	UCI gives better accuracy, compared to LMT.
2017	P. Sai Chandrasekhar Reddy et al, [17]	Heart disease prediction using ANN algorithm in data mining.	ANN	Accuracy proved in JAVA.
2018	Chala Bayen et al,[12]	Prediction and Analysis the occurrence of Heart Disease using data mining techniques.	J48, Naïve Bayes, Support Vector Machine.	It gives short time result which helps to give quality of services and reduce cost to individuals.
2018	R. Sharmila et al, [13]	A conceptual method to enhance the prediction of heart diseases using the data techniques.	SVM in parallel fashion	SVM provides better and efficient accuracy of 85% and 82.35%. SVM in parallel fashion gives better accuracy than sequential SVM.

MODULE DESCRIPTION



Each *module* in Machine Learning Studio (classic) represents a set of code that can run independently and perform a machine learning task, given the required inputs. A module might contain a particular algorithm, or perform a task that is important in machine learning, such as missing value replacement, or statistical analysis.

Data Format Conversions

Feature Selection

Data Transformation

Data Input and Output

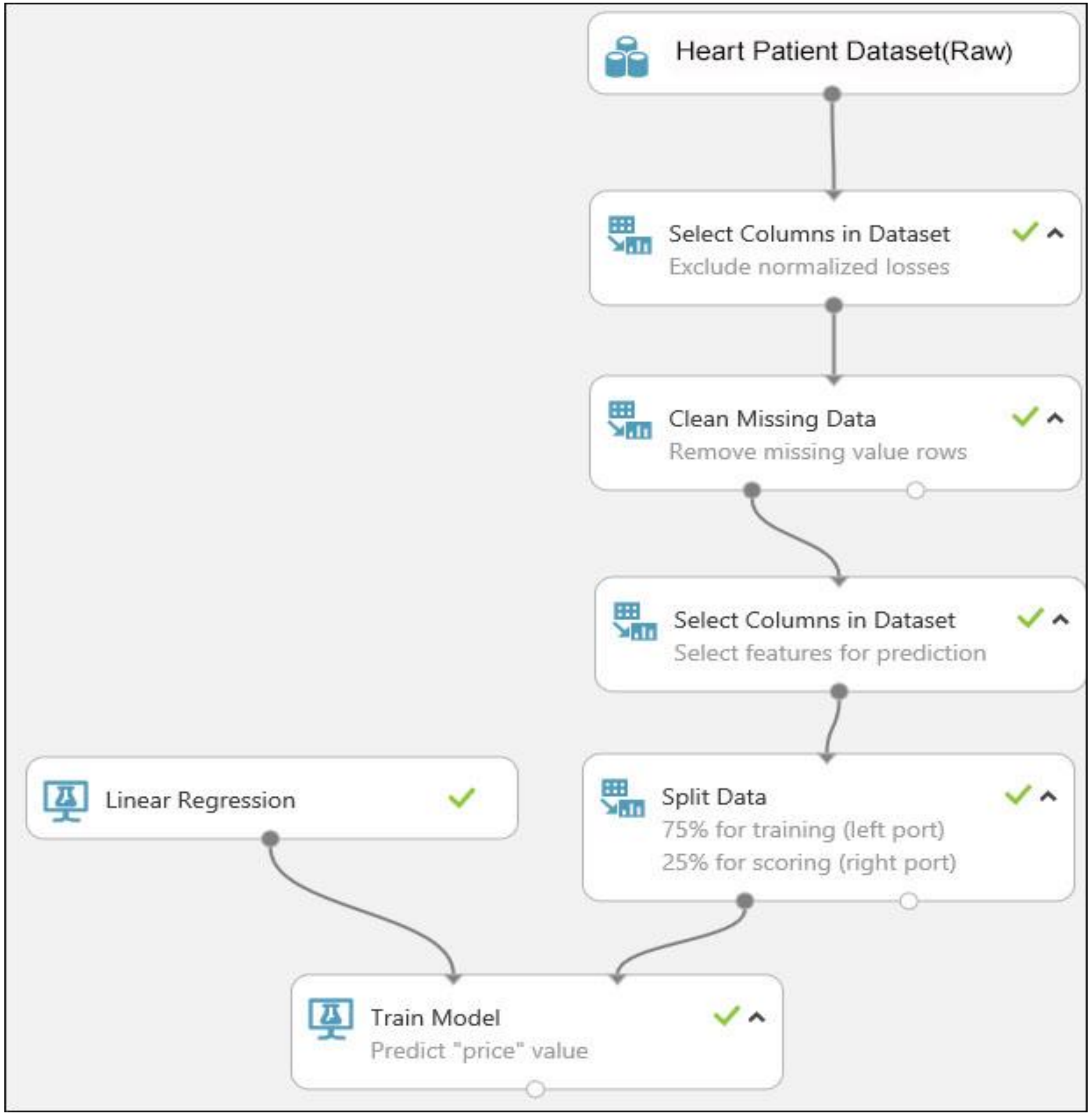
Machine Learning Modules

Python Language Modules

Statistical Functions

OpenCV Library Modules

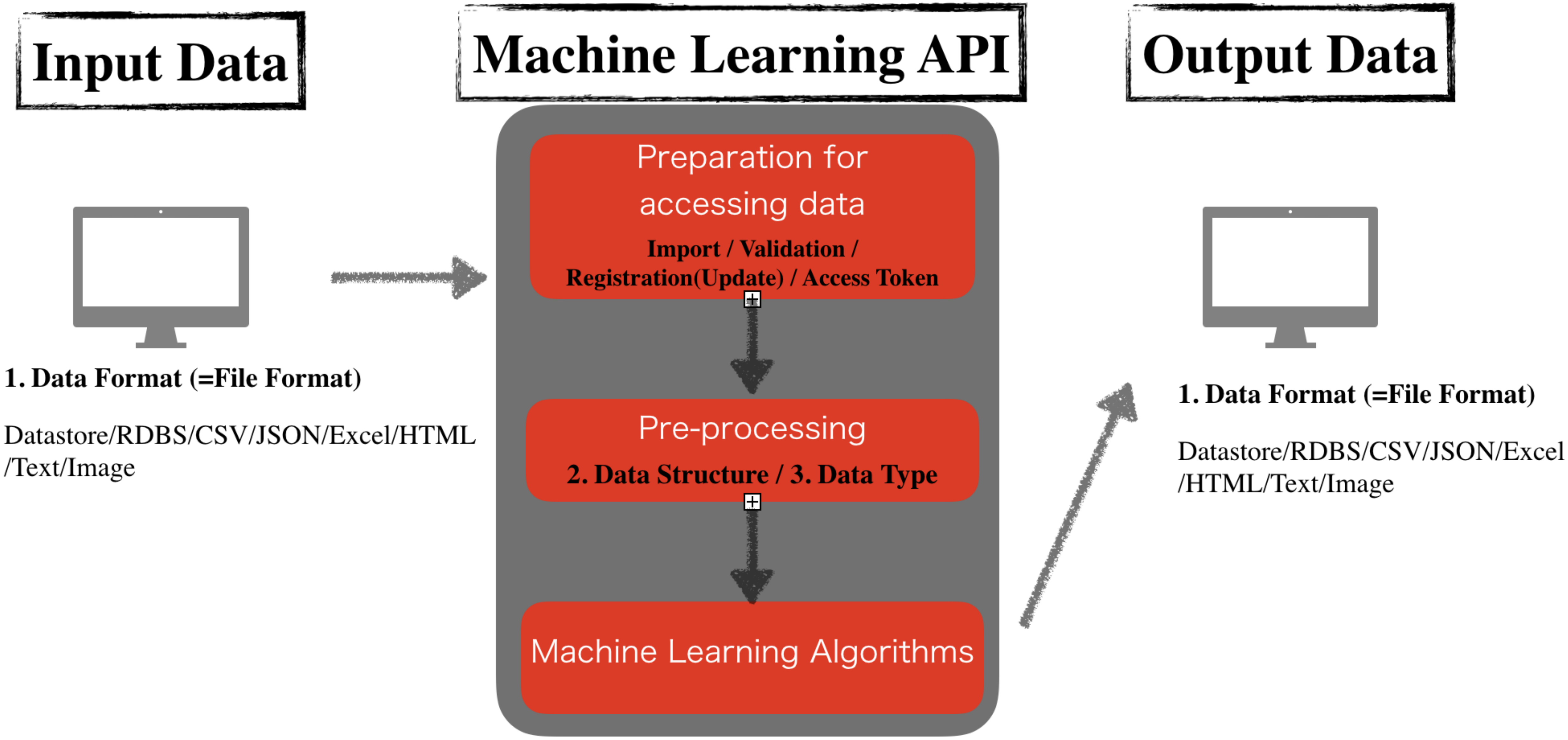
Text Analytics



Data Format Conversions

1. Convert to ARFF
2. Convert to CSV
3. Convert to Dataset
4. Convert to TSV
5. Convert to SVMlight

In our Project we have used CSV Data Format


Input Data
Machine Learning API
Output Data
Preparation for accessing data
Import / Validation / Registration(Update) / Access Token
Pre-processing
2. Data Structure / 3. Data Type
Machine Learning Algorithms
1. Data Format (=File Format)
Datastore/RDBS/CSV/JSON/Excel/HTML /Text/Image
1. Data Format (=File Format)
Datastore/RDBS/CSV/JSON/Excel /HTML/Text/Image

Feature Selection

The performance of machine learning model is directly proportional to the data features used to train it. The performance of ML model will be affected negatively if the data features provided to it are irrelevant. On the other hand, use of relevant data features can increase the accuracy of your ML model especially linear and logistic regression.

Advantages-

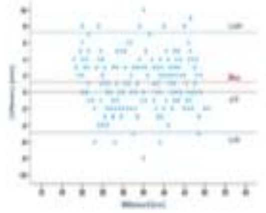
Performing feature selection before data modeling will reduce the overfitting.

Performing feature selection before data modeling will increases the accuracy of ML model.

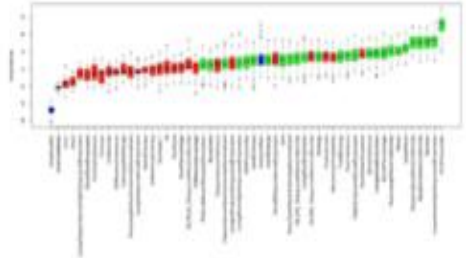
Performing feature selection before data modeling will reduce the training time

In our project we are using filter method of feature selection.

1. Analysis of reproducibility and exclusion of non-reproducible features

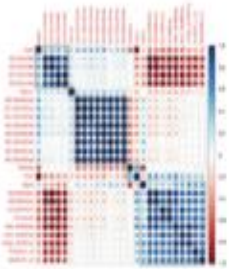


2. Calculation of „variable importance“ (ML algorithm, e.g. knock-off filter, Boruta…)



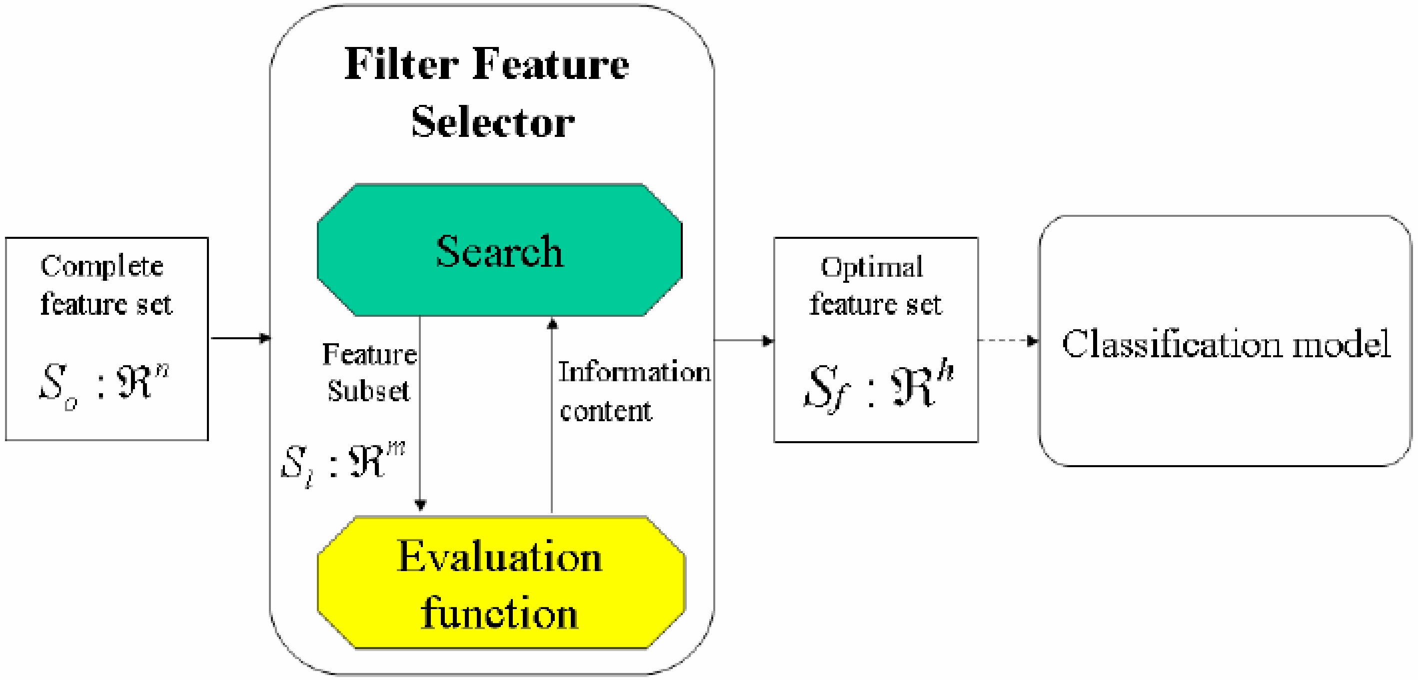
3. Data visualisation!

4. Correlation clusters

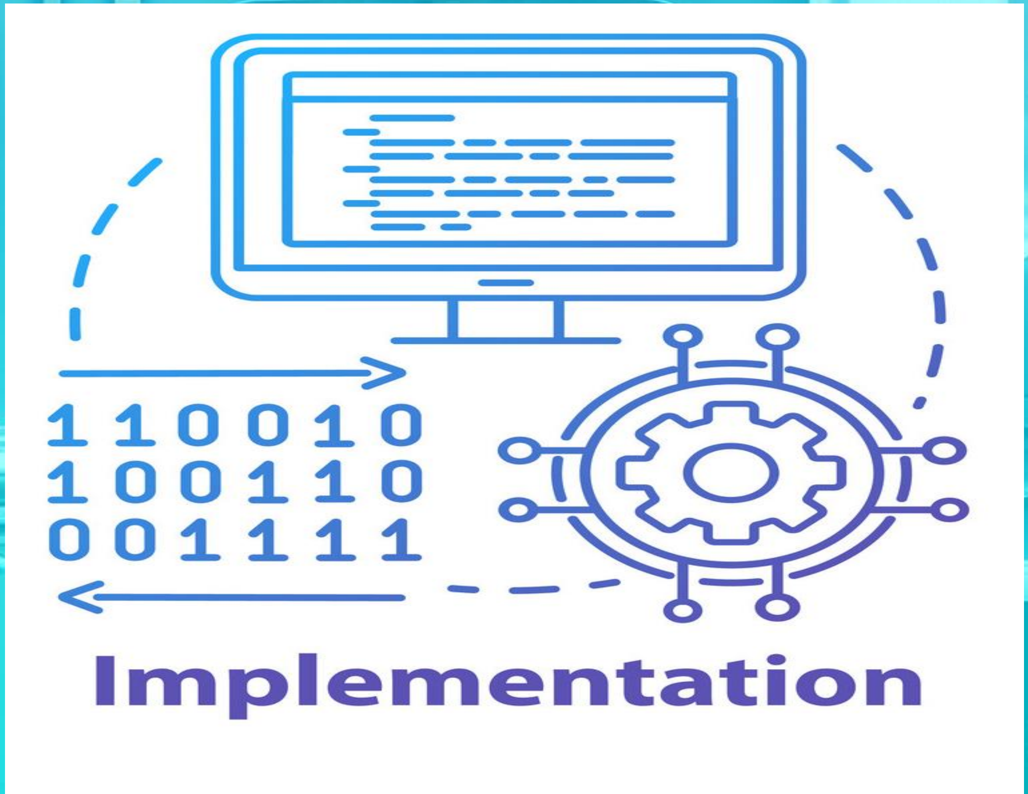


5. Selection of most representative features for each cluster

6. Model fitting with remaining features (usually n = 3-10)



Implementation and Coding



LIBRARIES

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

DATA FORMAT CONVERSION

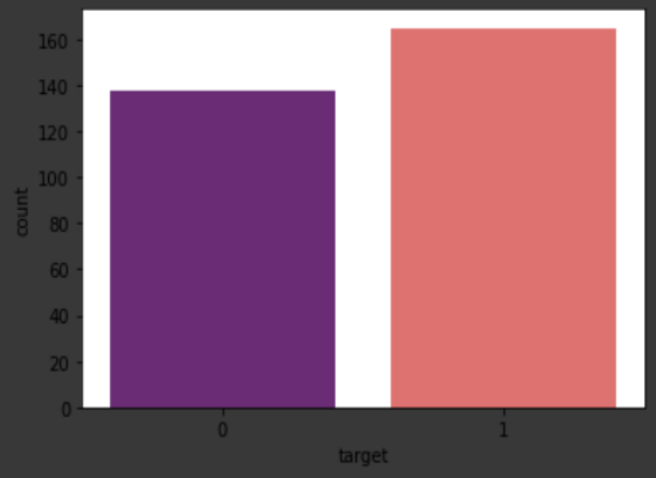
```
[ ]  df = pd.read_csv("/content/cardio_train.csv")
```

```
[ ]  df.head()
```

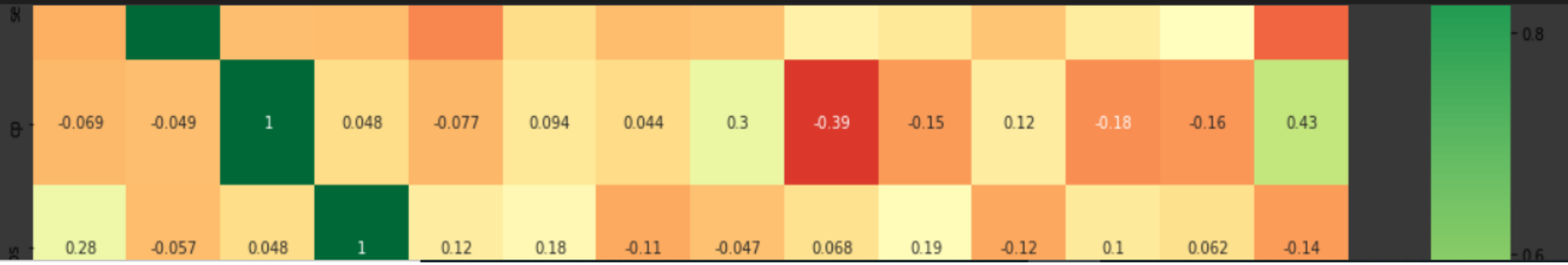
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

DATA EXPLORATION

```
[ ] sns.countplot(x="target", data=df, palette="magma")
    plt.show()
```




```
corrmat = df.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))
#plot heat map
g=sns.heatmap(df[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```



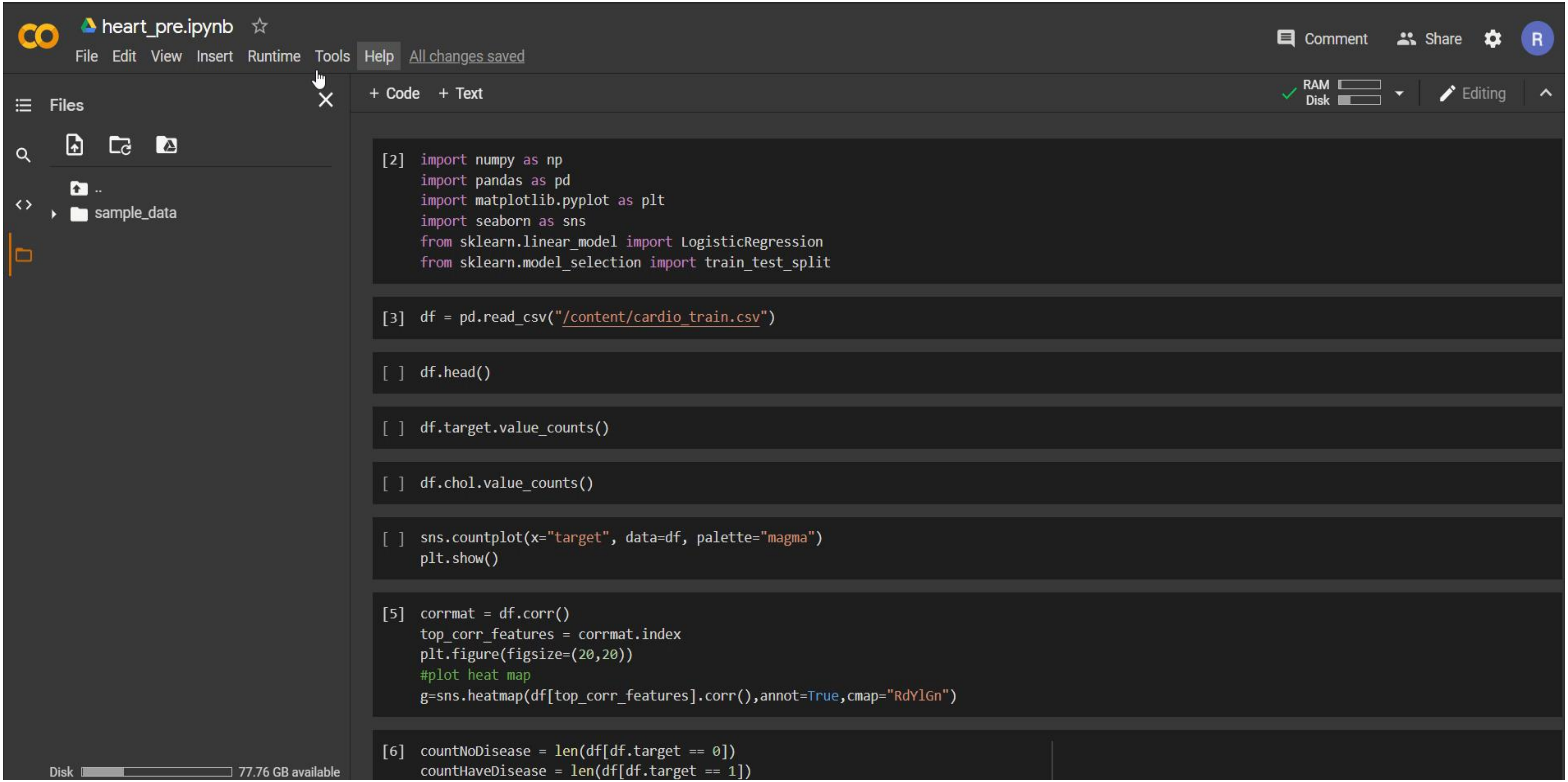
FEATURE SELECTION

```
[17] a = pd.get_dummies(df['cp'], prefix = "cp")
     b = pd.get_dummies(df['thal'], prefix = "thal")
     c = pd.get_dummies(df['slope'], prefix = "slope")
```

```
frames = [df, a, b, c]
df = pd.concat(frames, axis = 1)
df.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	cp_0	cp_1	cp_2	cp_3	thal_0	thal_1	thal_2	thal_3	slope_0	slope_1	slope_2
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1	0	0	0	1	0	1	0	0	1	0	0
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1	0	0	1	0	0	0	1	0	1	0	0
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1	0	1	0	0	0	0	1	0	0	0	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1	0	1	0	0	0	0	1	0	0	0	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1	1	0	0	0	0	0	1	0	0	0	1


DEMO VIDEO
DAILY REPORT SCHEDULE
08:15



heart_pre.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Files
+ Code + Text
RAM
Disk
Editing
sample_data
[2] import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
[3] df = pd.read_csv("/content/cardio_train.csv")
[ ] df.head()
[ ] df.target.value_counts()
[ ] df.chol.value_counts()
[ ] sns.countplot(x="target", data=df, palette="magma")
plt.show()
[5] corrmat = df.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))
#plot heat map
g=sns.heatmap(df[top_corr_features].corr(),annot=True,cmap="RdYlGn")
[6] countNoDisease = len(df[df.target == 0])
countHaveDisease = len(df[df.target == 1])
Disk 77.76 GB available

TESTING

- At last after hyperparameter tuning ,we have made our predictions on testing data by using method predict().
- From predictions we can examine whether person is suffering from Cardiovascular(Heart) disease or not.

RESULT AND ANALYSIS

CONFUSION MATRIX

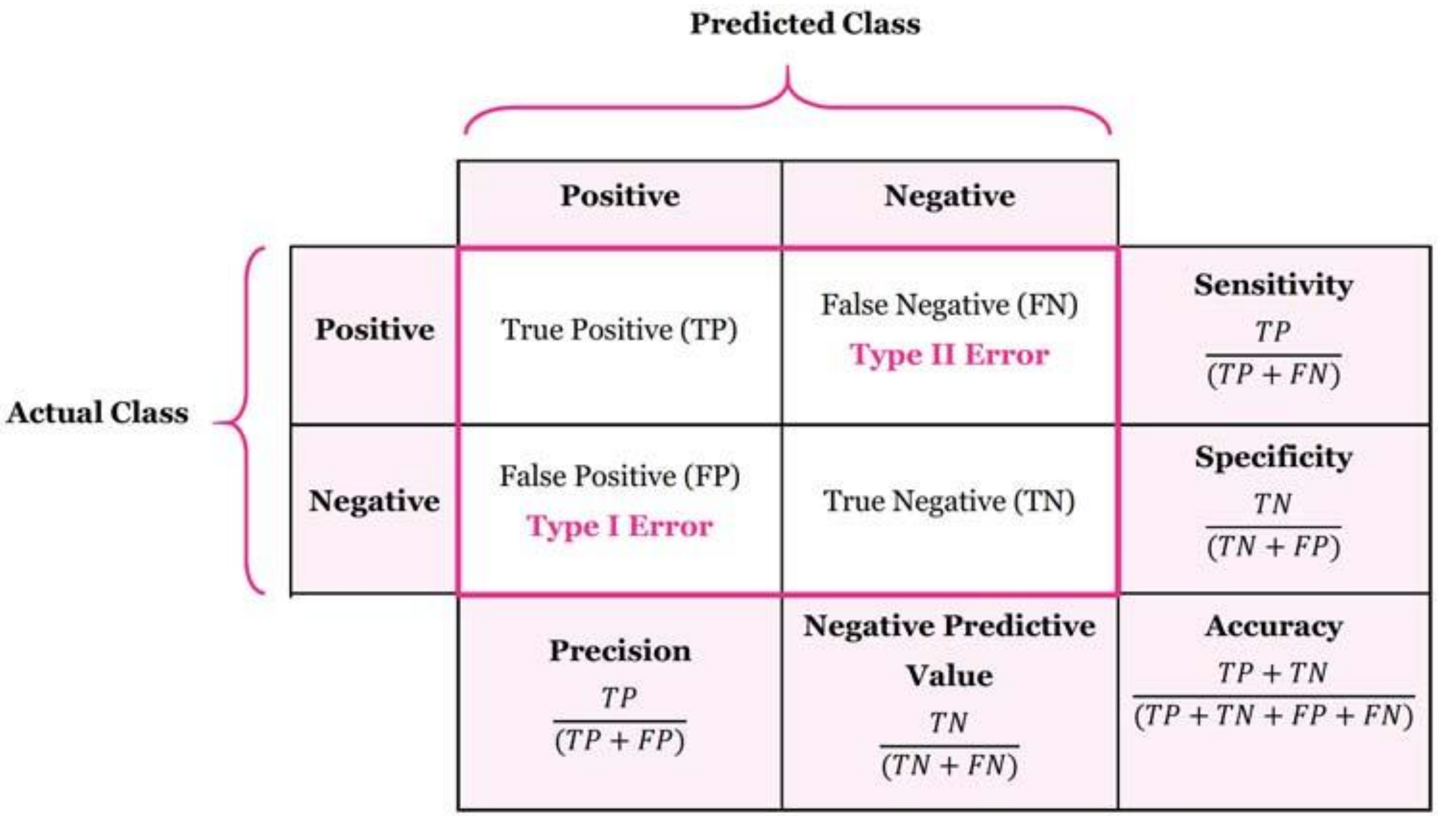
Based on the true positive and true negative in the confusion matrix we can analysis the performance. Summation of true positive and true negative suggest correct predictions, while the summation of false positive and false negative suggest incorrect predictions. Further with the calculation of true positive and true negative we can also predict whether the model is more inclined towards sensitivity or specificity.

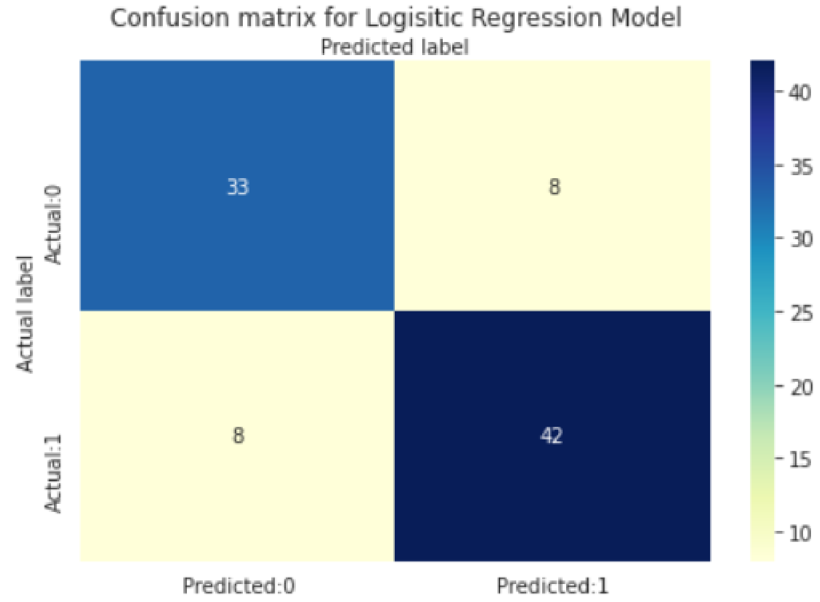
TP = The number of positively labelled data, which have been classified as "Correct".

TN= The number of negatively labelled data, which have been classified as "Correct".

FN= The number of positively labelled data, which falsely have been classified as "Negative".

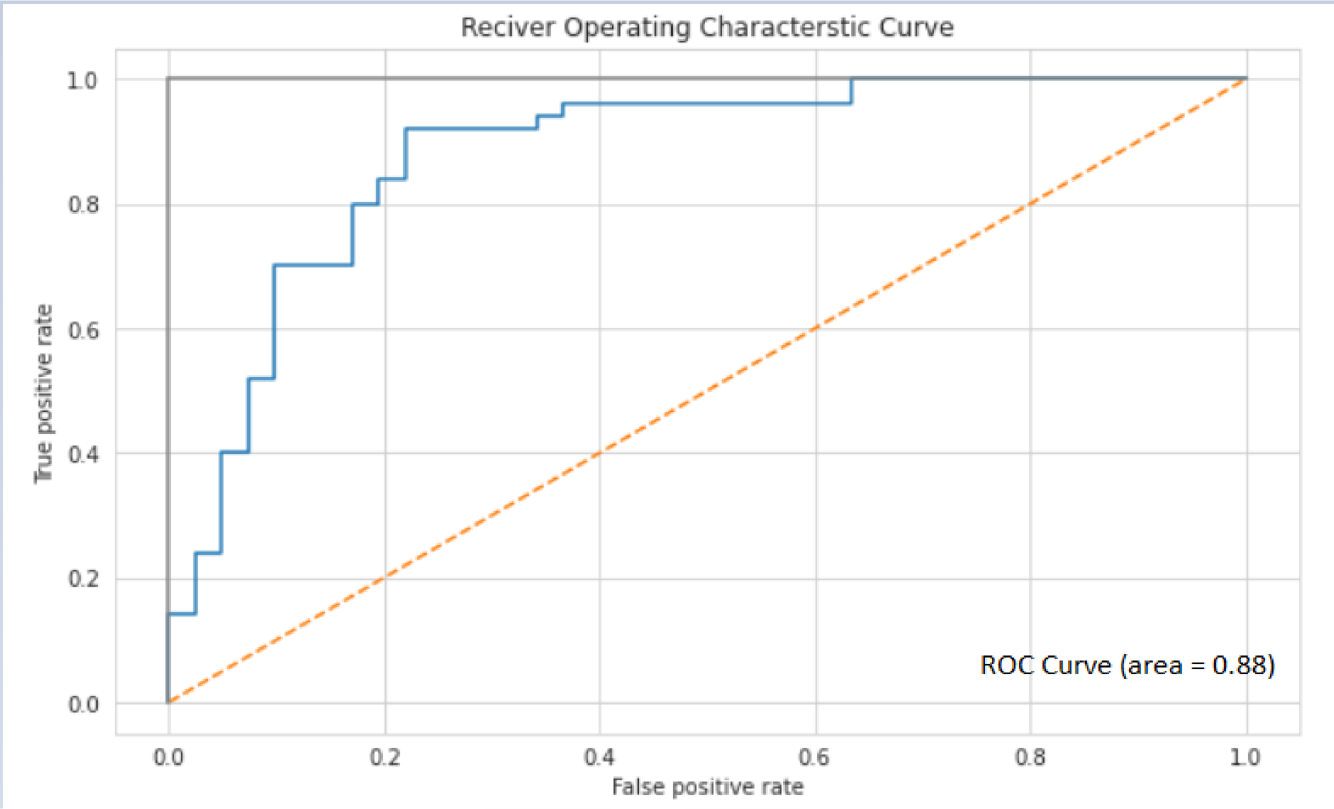
FP= The number of negatively labelled data, which falsely have been classified as "Positive".


Predicted Class
Positive
Negative
Actual Class
Positive
True Positive (TP)
False Negative (FN)
Type II Error
Sensitivity
$\frac{TP}{(TP + FN)}$
Negative
False Positive (FP)
Type I Error
True Negative (TN)
Specificity
$\frac{TN}{(TN + FP)}$
Precision
$\frac{TP}{(TP + FP)}$
Negative Predictive Value
$\frac{TN}{(TN + FN)}$
Accuracy
$\frac{TP + TN}{(TP + TN + FP + FN)}$


Confusion matrix for Logisitic Regression Model
Predicted label
Actual label
Actual:0
Actual:1
33
8
8
42
Predicted:0
Predicted:1
40
35
30
25
20
15
10

ROC CURVE

ROC curve is a simple curve that work for the binary classifier. It is mainly used to show trade-off among false positive rate and true positive rate. For a precise accuracy model, the true positive rate should be more than the false positive in every thresholds aspect. The summary of model is given by area under the curve. The indication for lower false positives and higher true negatives is shown by the smaller value on x-axis. While the higher true positives and lower false negatives is shown by the larger value on the y-axis.


Reciver Operating Characterstic Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (area = 0.88)

CONCLUSION

We can say that deep learning methods produce better results as compared to supervised learning methods. Though classification is also an important issue. A combination of different prediction models might be more accurate in the prediction of the early symptoms of cardiovascular diseases. Cardio diseases are complex and death due to these diseases increases every year. The fundamental motive of these predicting models is to achieve a high accuracy rate in heart disease prediction. The future prediction algorithm model should be based on the basis of less time complexity.

It has been observed that a properly cleaned and pruned dataset provides much better accuracy than an unclean one with missing values. Selection of suitable techniques for data cleaning along with proper classification algorithms will lead to the development of prediction systems that give enhanced accuracy. In future an intelligent system may be developed that can lead to selection of proper treatment methods for a patient diagnosed with heart disease. Data mining can be of very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

THANK YOU:)