Cardiovascular Disease Prediction

Group: 32

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Presentation By:

Arnav Gupta	**2021236**
Karan Gupta	**2021258**
Shivesh Gulati	**2021286**
Vishal Singh	**2021575**

Outline

- Introduction and Motivation
- Literature Review
- Dataset Description
- Models and Methodologies
- Result Analysis and Conclusion
- Timeline
- Contributions

Motivation

Introduction

The main goal behind this project is to detect the presence of cardiovascular disease, among people, using easy to determine parameters like Blood Pressure, cholesterol levels, glucose levels etc. using machine learning models.

Motivation

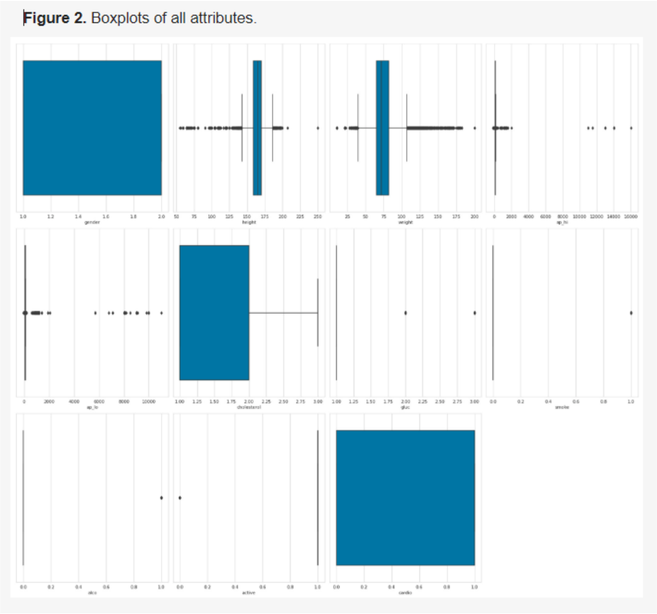
Cardiovascular diseases are one of the leading causes of fatalities worldwide. The primary motivation behind this study is to develop a machine learning model capable of predicting cardiovascular diseases(CVD) in an individual, using easy-to-determine parameters such as age, glucose levels, weight, and blood pressure indices. This can serve as a tool for early detection of the risk of cardiovascular diseases among individuals, allowing them to take preventive measures and seek medical attention at an early stage to reduce further risk.

Literature Review

Research Paper-1

Aim: To develop a machine learning model to predict Cardiovascular Disease Accurately

- The dataset used, consisting of **70,000 records**, was taken from Kaggle.com.
- The study uses **Decision trees, XGBoost, Random forest,** and **multilayer perceptron** models. These were optimized using GridSearchCV.
- Since dataset was mostly categorical, various techniques for outlier detection were used, resulting data had **57,155** instances.
- Feature engineering such as binning(converting continuous data to categorical) was applied to age, height, weight, systolic blood pressure, and diastolic blood pressure.



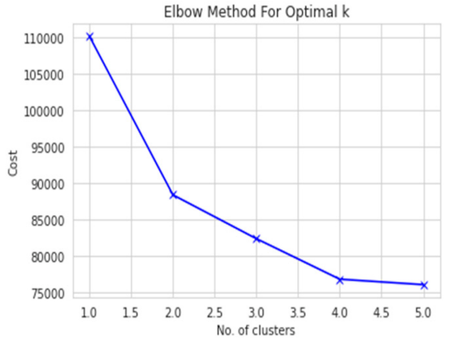
kaggle

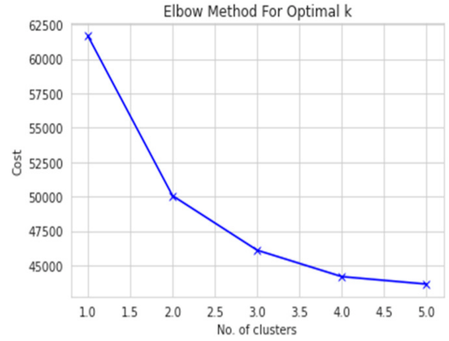
Images taken from : https://www.mdpi.com/1999-4893/16/2/88

Literature Review

Research Paper-1(continued)

Aim: To develop a machine learning model to predict Cardiovascular Disease Accurately





- Two derived parameters, viz Body Mass Index (**BMI**) and Mean Arterial Pressure (**MAP**), were derived from age, weight, Systolic Blood Pressure, Diastolic Blood Pressure, respectively.
- K-Modes clustering was used over K-means as it was not suitable for categorical dataset. optimal number was found using elbow curve method.
- The dataset was divided on gender basis for better prediction.
- The dataset was split into **80:20** for training and testing.
- MLP was best to perform with an accuracy of **87.28%** and other algorithms like XGBoost and random forest also showed comparable performances.

Table 5. The evaluation metrics resulting from different classifiers.

Model	Accuracy		Precision		Recall		F1-Score		AUC
	Without CV	CV	Without CV	CV	Without CV	CV	Without CV	CV	
MLP	86.94	87.28	89.03	88.70	82.95	84.85	85.88	86.71	0.95
RF	86.92	87.05	88.52	89.42	83.46	83.43	85.91	86.32	0.95
DT	86.53	86.37	90.10	89.58	81.17	81.61	85.40	85.42	0.94
XGB	87.02	86.87	89.62	88.93	82.11	83.57	86.30	86.16	0.95

To conclude, this research paper applied models that achieved an accuracy of **>86%**, which could be further improved in the future to get more accurate predictions of CVD

Images taken from : https://www.mdpi.com/1999-4893/16/2/88

Literature Review

Research Paper-2

Aim: To understand the effectiveness of various BP indices in predicting Cardiovascular Diseases using Machine Learning Models.

- The dataset used in this study, although not known in its entirety, was obtained from the ADVANCE study and consisted of the following attributes: age, sex, gender, SBP, MAP, DBP, PP, smoking, total cholesterol (HbA1c), and duration of mellitus (type-2 diabetes)
- COX Proportional Hazard Regression Models were used to find the Hazard Ratio and the 95% confidence interval.
- The ability of different BP indices, to discriminate among the participants was measured using AUC analysis.
- For multivariate analysis, RIDI (Relative Integrated Discrimination Improvement) was used, which measured the increase in the discrimination obtained when new variables were introduced to the prediction model.

Literature Review

Research Paper-2

Conclusions: The overall conclusions obtained from the study were as follows:

- It was observed during the study that for baseline BP estimates pulse Pressure (PP) and Mean Arterial Pressure (MAP) were more effective in CVD prediction.
- In the case of achieved BP estimates, Systolic Blood Pressure (SBP) and Mean Arterial Pressure (MAP) performed better compared to other metrics.
- Overall, if a combination of Blood Pressure variables is used, it was concluded that SBP and PP are superior to MAP and DBP in predicting CVD-related events.
- DBP was observed to be the worst for predicting CVD-related events, mainly because of the fact that DBP tends to remain constant or even decrease after the age of 50-60 years

Shape Size and Derived Features

Summary of the dataset features

Dataset Source — Kaggle

Original Shape and Size — 70,000 rows
13 columns

Target Variable — A binary variable “cardio“
(0 for no CVD, 1 for presence of CVD)

Dropped Column — Patient ID

Derived Features — Pulse Pressure (PP) = ap_hi-ap_lo
Mean Arterial Pressure (MAP) = ap_lo+1/3(PP)

Final Shape and SIze — 70,000 rows and 14 columns (excluding Patient Id and adding PP and MAP)

Feature Description

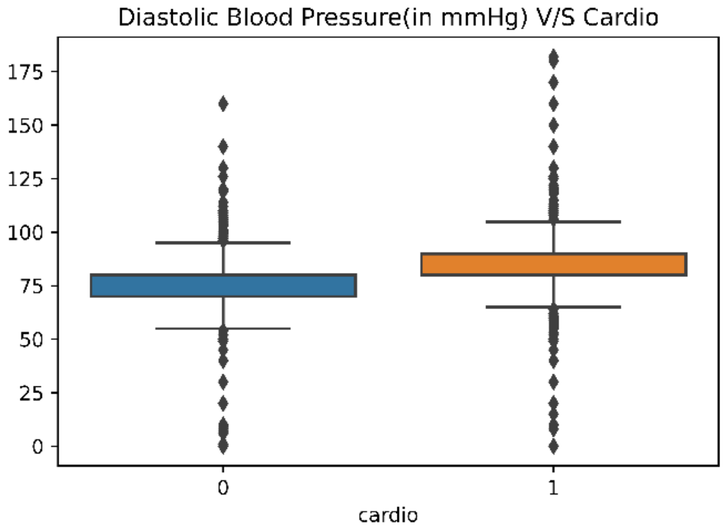
Feature	Description	Type of Feature	Data-Type	Unit of measurement	Value Range
age	Age	Objective Feature	integer	Days	Any Integer Value >= 0
height	Height	Objective Feature	integer	Centimeters (cm)	Any Integer Value > 0
weight	Weight	Objective Feature	float	Kilograms (kg)	Any Floating Value > 0
gender	Gender	Objective Feature	Categorical Code	-	1: Female 2: Male
ap_hi	Systolic Blood Pressure	Examination Feature	integer	Millimetre(s) of Mercury (mmHg)	Any Integer Value > 0
ap_lo	Diastolic Blood Pressure	Examination Feature	integer	Millimetre(s) of Mercury(mmHg)	Any Integer Value > 0
cholesterol	Cholesterol	Examination Feature	Categorical Code	-	1:Normal 2:Above Normal 3:Well Above Normal
gluc	Glucose	Examination Feature	Categorical Code	-	1:Normal 2: Above Normal 3:Well Above Normal
smoke	Smoking	Subjective Feature	Categorical Code	-	0: Non-Smoker 1: Smoker
alco	Alcohol Intake	Subjective Feature	Categorical Code	-	0: Does not drink alcohol 1: Drinks Alcohol
active	Lifestyle	Subjective Feature	Categorical Code	-	0: Sedentary Lifestyle 1: Active Lifestyle
PP	Pulse Pressure	Derived Feature	integer	Millimeter(s) of Mercury (mmHg)	Any Integer Value > 0
MAP	Mean Arterial Pressure	Derived Feature	float	Millimeter(s) of Mercury(mmHg)	Any Floating Value > 0

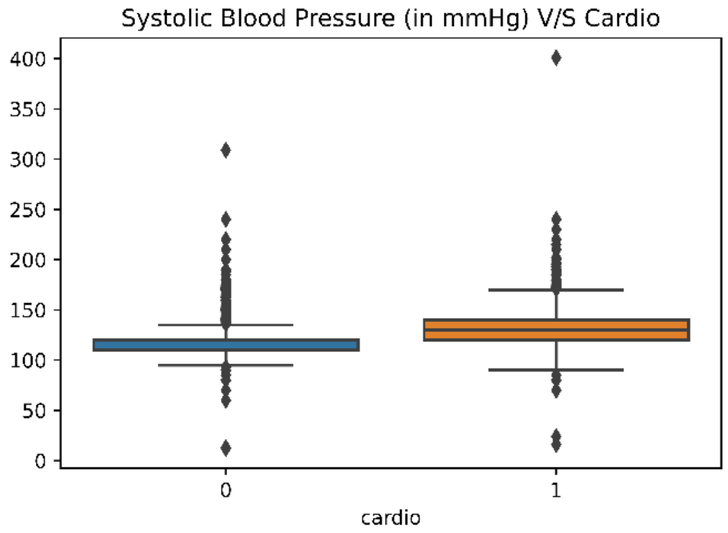
Table 1. Table describing the dataset features

Box Plots

Inference

The number of outliers, in case of Diastolic Blood Pressure are high



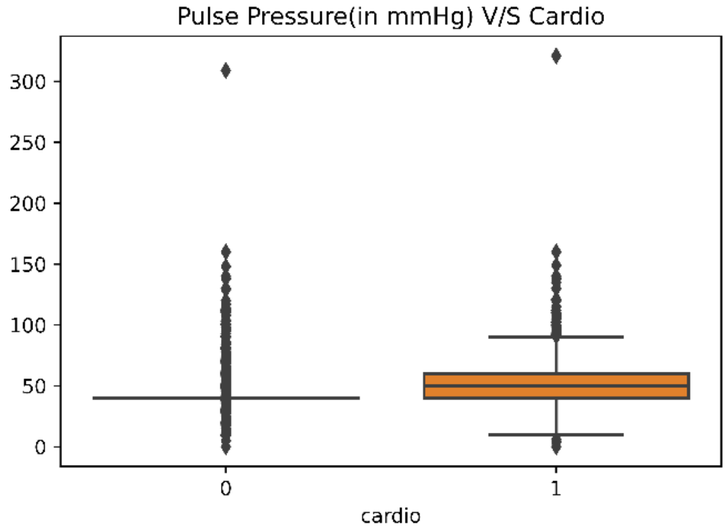


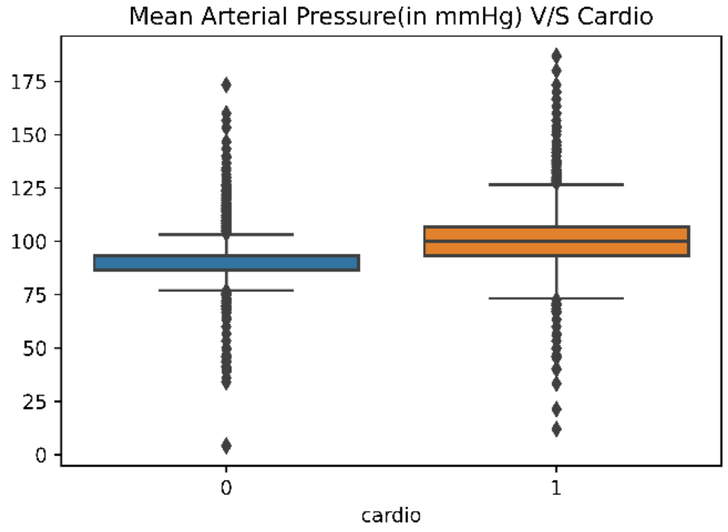
Inference

The number of outliers, in case of Systolic Blood Pressure are higher than Dyastolic Blood pressure, but the quartiles are better defined.

Inference

The number of outliers, in case of pulse pressure are the highest.

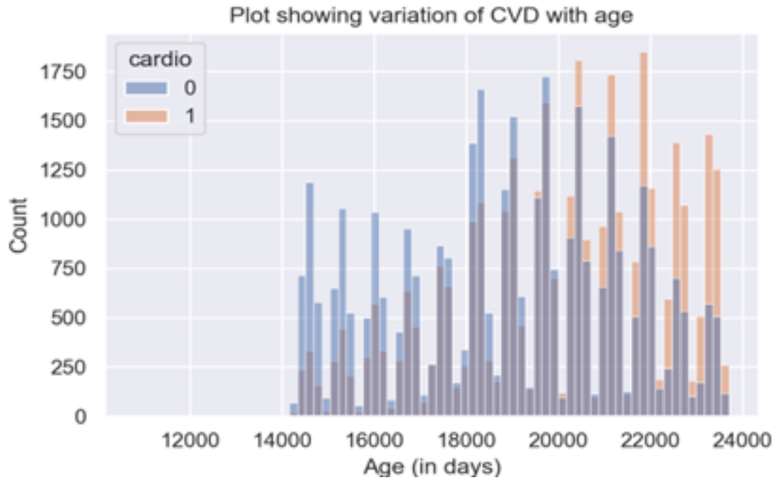




Inference

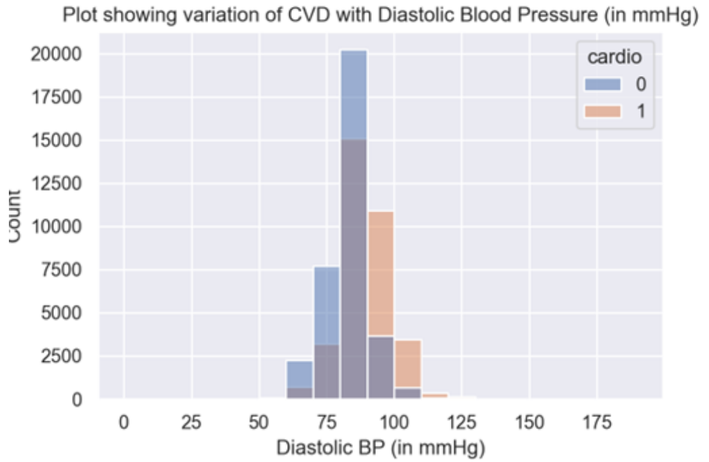
Mean Arterial Pressure, shows the least number of outliers, and has significant size of the box.

Histograms

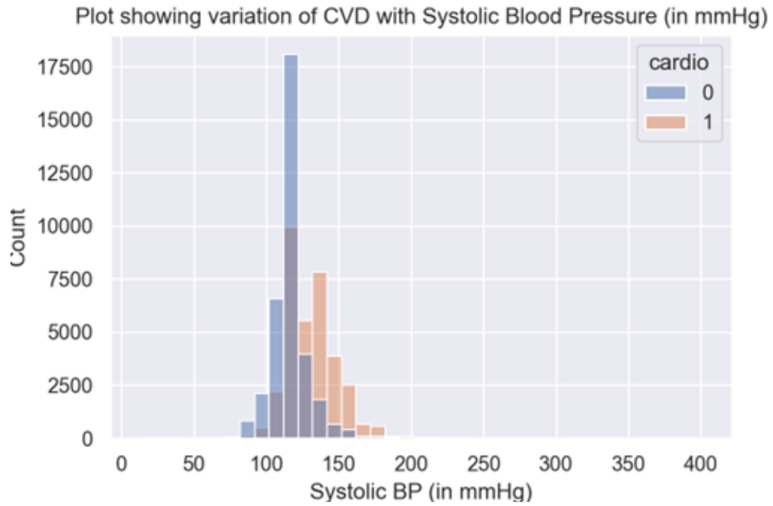


Graph-1

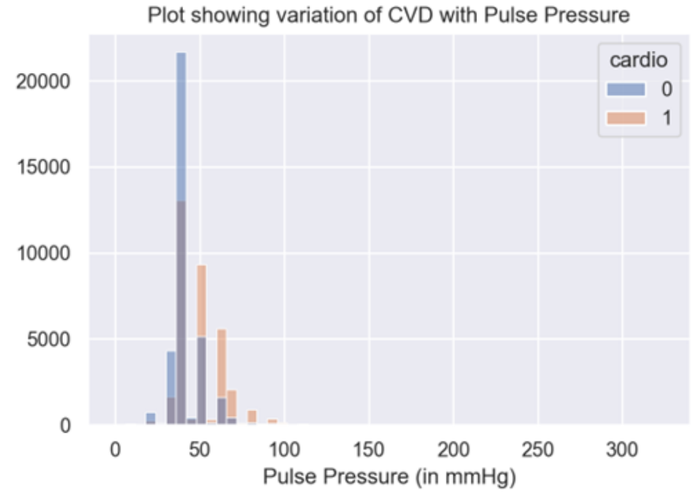
Graph-2



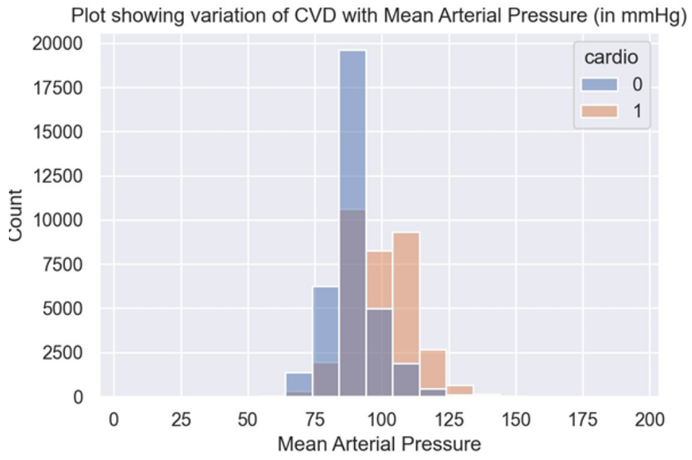
Graph-3



Graph-4



Graph-5



Graph-6

Inferences from Histograms

Histograms were drawn to analyze the distribution of participants, with or without CVD, in specific ranges of the various numerical features. The insights obtained are as follows:-

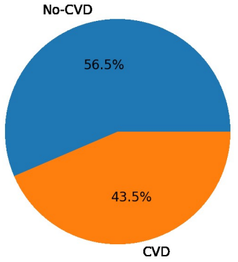
- As the age bracket increases, the number of people with CVD is higher than those without CVD, especially for people beyond 22500 days (61 years). **(See Figure 1)**

- As the weight bracket increases, the proportion of people having CVD is significantly higher as compared to those not having CVD, especially in the 75 kg and beyond range **(See Figure 2).**

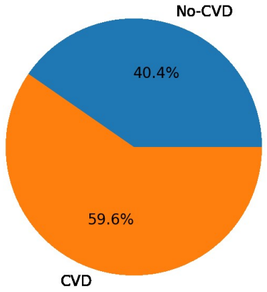
- In populations with higher systolic, diastolic, and pulse pressures, cardiovascular disease (CVD) frequency is higher. Specifically, systolic BP >=140 mmHg, diastolic BP>=90 mmHg, and pulse pressure >=50 mmHg are associated with increased CVD prevalence. **(See Figures 3,4,5,6)**.

Pie Charts

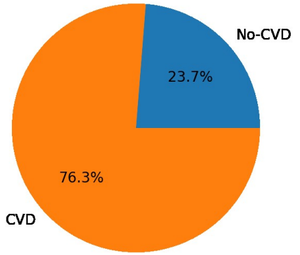
% of people with and without CVD having normal cholesterol



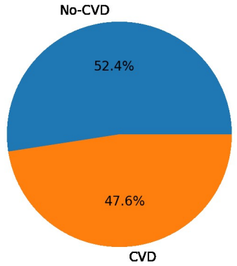
% of people with & without CVD having above-normal cholesterol



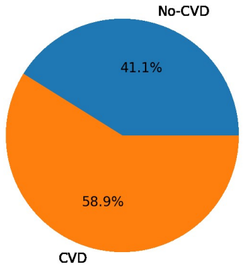
% of people with & without CVD having well-above-normal cholestrol



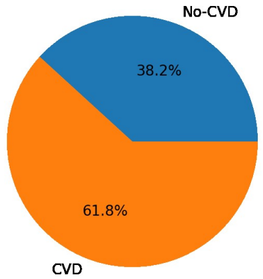
% of people with and without CVD having normal glucose



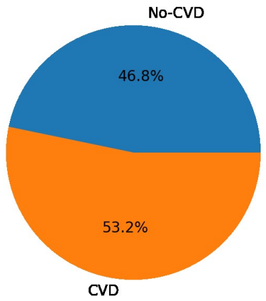
% of people with & without CVD having above-normal glucose



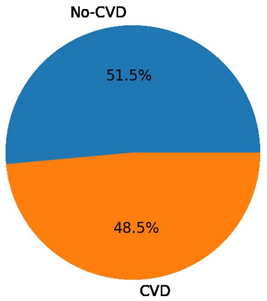
% of people with & without CVD having well above normal glucose



% of people with & without CVD having low activity levels



% of people with and without CVD having high activity levels



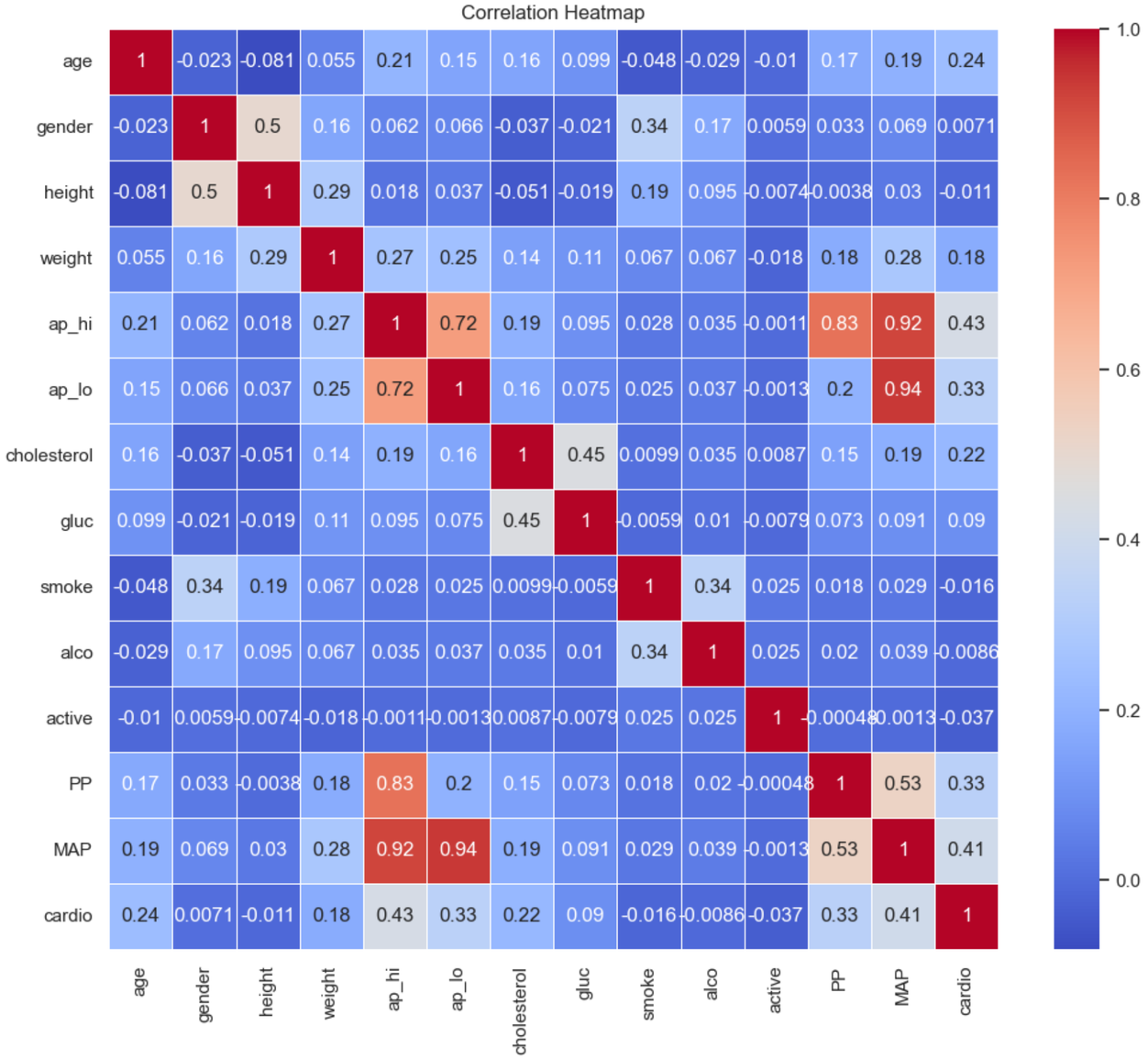
Inferences

Pie charts were used to analyze the distribution of participants, with or without CVD, in various categories of categorical features :-

1) Proportion of people not having CVD in the normal cholesterol level category is higher as compared to the people having CVD in the above-normal and well-above normal cholesterol categories the proportion of people, having CVD, is higher, as compared to those not having CVD. A similar trend can be seen in glucose levels as well.

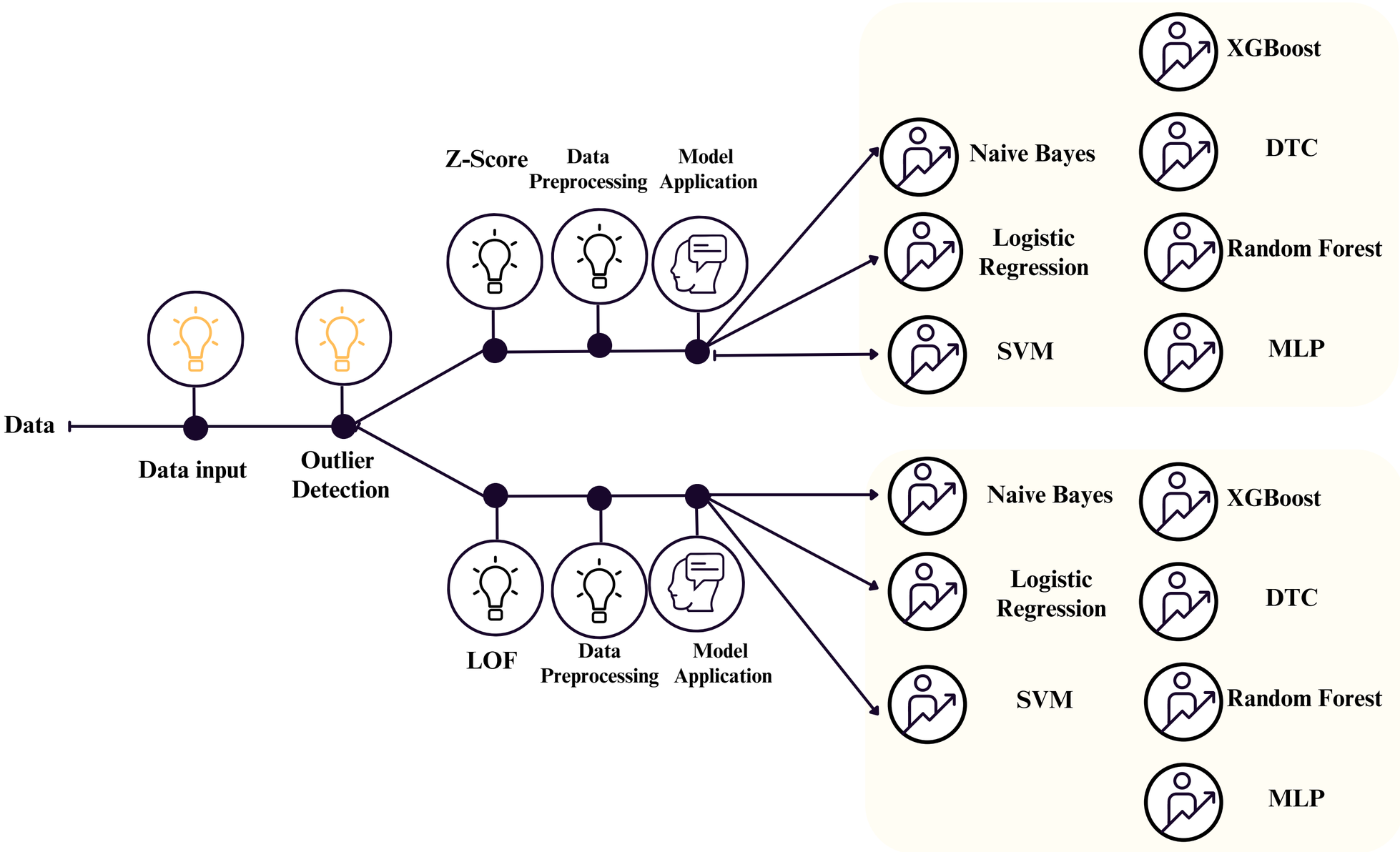
2) Out of the people with an active lifestyle, the proportion of people not having CVD is higher than those with CVD. Similarly, the people with a sedentary lifestyle have a higher proportion of people who have CVD than the ones not having CVD.

Correlation Heatmap



- Ap_lo and Ap_hi have a strong correlation(close to 1) as greater diastolic pressure means a greater systolic pressure.
- PP and MAP both have a strong correlation with ap_hi and ap_lo.
- Ap_hi and map have a moderate correlation (around 0.5) with CVD as higher blood pressure generally means a greater risk of CVD.
- PP ap_lo and cholesterol have somewhat moderate(around 0.3) correlation with CVD PP and MAP have a moderate correlation with each other

Flow Chart


Data
Data input
Outlier Detection
Z-Score
Data Preprocessing
Model Application
LOF
Data Preprocessing
Model Application
Naive Bayes
Logistic Regression
SVM
XGBoost
DTC
Random Forest
MLP
Naive Bayes
Logistic Regression
SVM
XGBoost
DTC
Random Forest
MLP

Data Cleaning

Outlier Detection and Cleaning

Manual CAP

For the original and derived features of Blood Pressure, Values not in the range [0,500] were manually dropped.

70,000 → 68,727

Z-Score

Applied Z-Score on the Manually Capped Data

68,727 → 65,048

LOF

Applied Local Outlier Factor on the Manually Capped Data

68,727 → 54,981

Z-Score

Formula

Score

Mean

$$Z = \frac{x - \mu}{\sigma}$$

SD

Lower bound=-2.75

Upper bound=2.75

Final size: 65,048

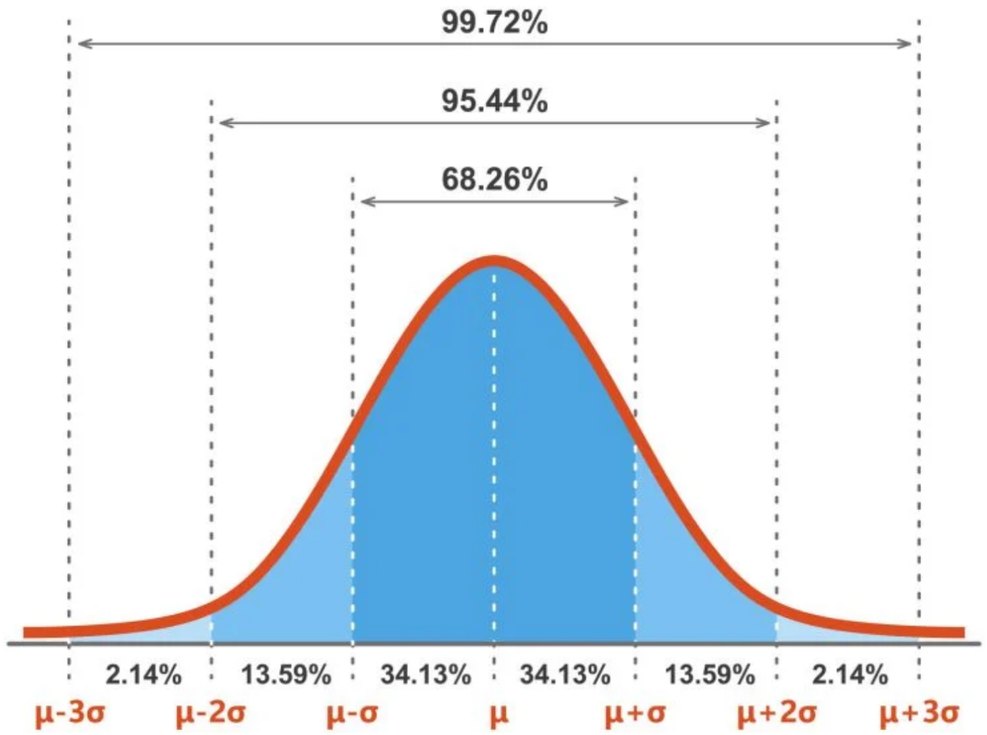


Image Source : https://www.simplypsychology.org/z-score.html

Local Outlier Factor

Formula

For a given Data set

$$D_n = \{ (x_i, y_i) | x_i \in R^2, y_i \in \{X, Y, Z\} \}$$

Local Outlier Factor for each data point is given by

$$LOF(x_i) = \frac{\sum_{x_j \in N(x_i)} lrd(x_j)}{|N(x_i)|} \times \frac{1}{lrd(x_i)}$$

$|N(x_i)|$: Number of elements in the neighborhood of x_i

$lrd(x_i)$: Local Reachability Density of x_i

www.mlpoint.com

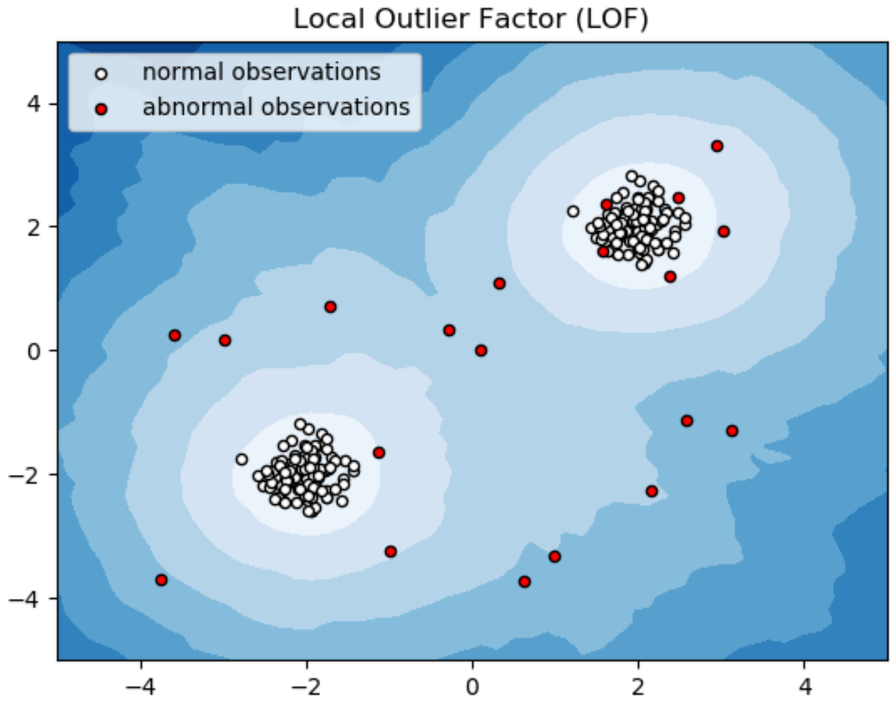


Image Source: https://scikit-learn.org/0.19/auto_examples/neighbors/plot_lof.html

Final size: 54,981

Data Pre-Processing

Train-Test-Split

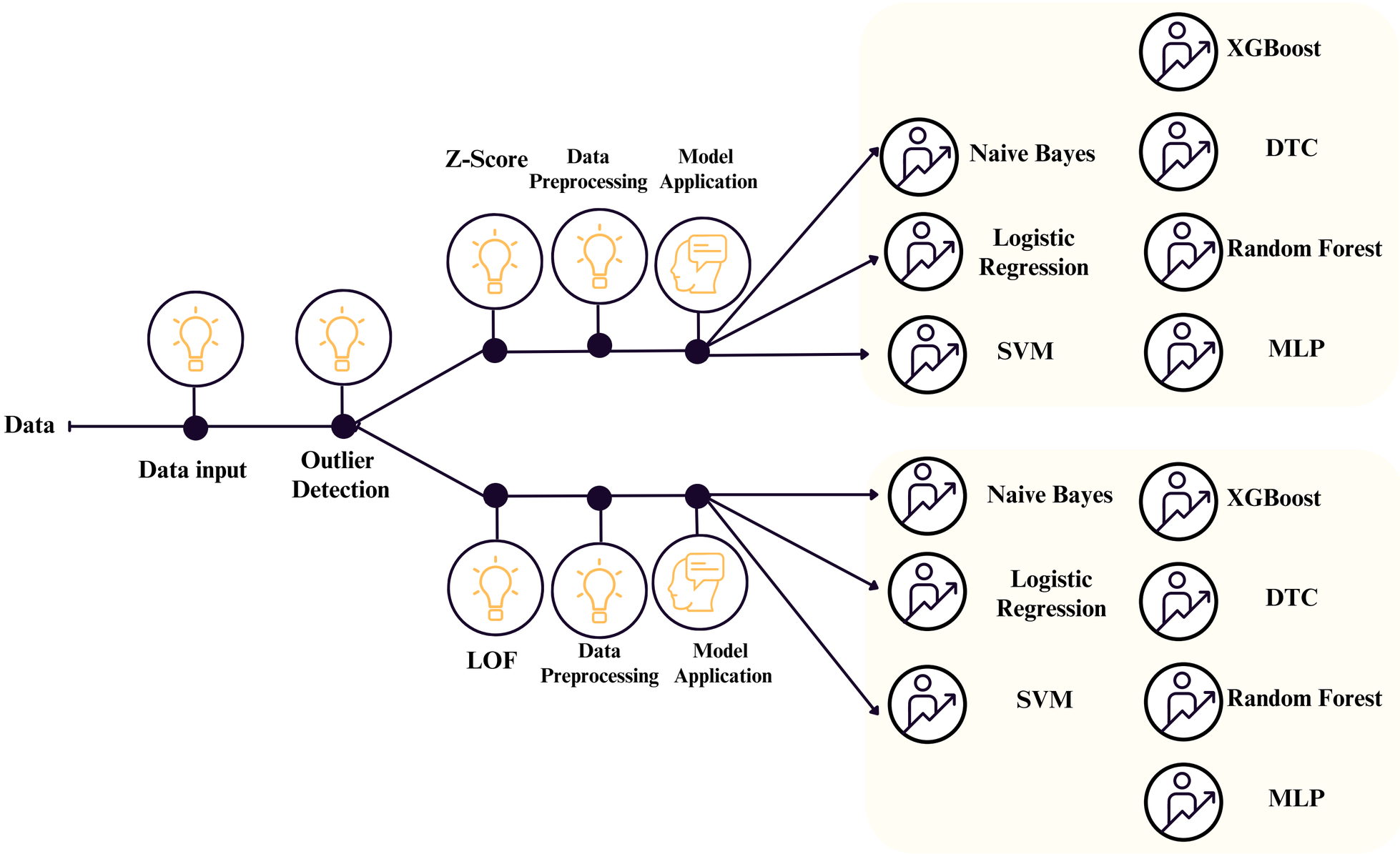
70:30

Data Standardization

StandardScaler()

Image Source: https://stackoverflow.com/questions/39120942/difference-between-standardscaler-and-normalizer-in-sklearn-preprocessing

Flow Chart


Data
Data input
Outlier Detection
Z-Score
Data Preprocessing
Model Application
LOF
Data Preprocessing
Model Application
Naive Bayes
Logistic Regression
SVM
XGBoost
DTC
Random Forest
MLP
Naive Bayes
Logistic Regression
SVM
XGBoost
DTC
Random Forest
MLP

Models and Methodologies

Classification Problem

- Naive Bayes Classifier
- Logistic Regression Classifier
- Support Vector Machine
- Multi Layer Perceptron
- Random Forest
- XGBoost
- Decision Tree

Models and Methodologies

Data Processing

Standard Data

The Original Data had Label Based Encoded Features and Models were applied.

One Hot Encoding

The Original Label Based Features were converted to One Hot Encoding and the Models were applied.

One Hot Encoding + PCA

The One Hot Encoded data was used with PCA and the Models were applied. The optimal number of components for PCA, were determined using K-Fold Cross Validation

.

Models (Stats)

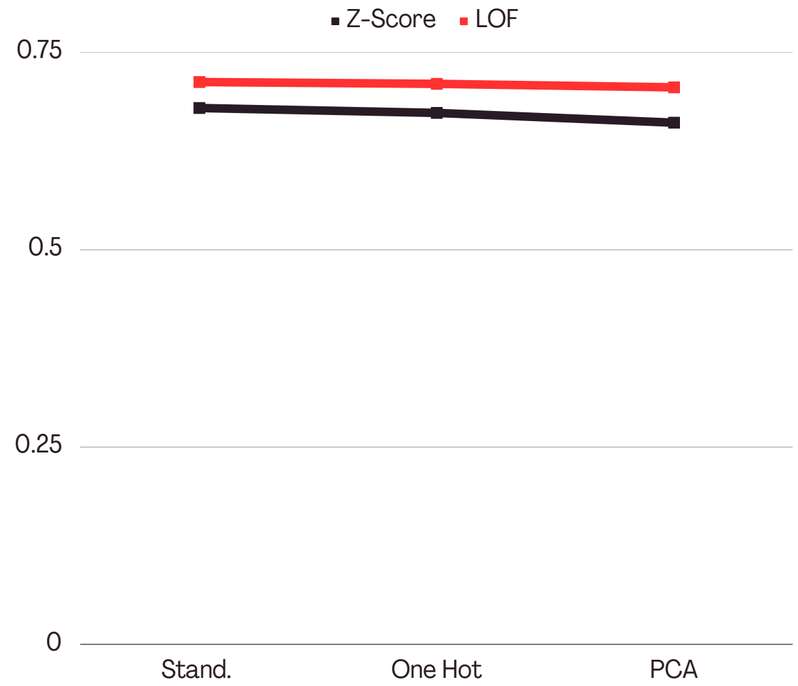
Naive Bayes

Z-Score Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Gaussian Naive Bayes	Standard Data	0.7209	0.7191	0.7570	0.7592	0.6159	0.6148	0.6792	0.6794
	One Hot Encoded	0.7169	0.7139	0.7520	0.7529	0.6113	0.6088	0.6744	0.6732
	One Hot Encoded+ PCA(components=3)	0.7126	0.7105	0.7611	0.7633	0.5841	0.5825	0.6610	0.6608

LOF Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Gaussian Naive Bayes	Standard Data	0.7275	0.7313	0.7763	0.7889	0.6403	0.6493	0.7018	0.7123
	One Hot Encoded	0.7236	0.7290	0.7712	0.7857	0.6371	0.6476	0.6978	0.7100
	One Hot Encoded+ PCA(components=3)	0.7217	0.7253	0.7694	0.7823	0.6343	0.6426	0.6954	0.7056



F1-Score

Models (Stats)

Logistic Regression

Z-Score Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Logistic Regression	Standard Data	0.7244	0.7206	0.6480	0.6433	0.7444	0.7448	0.6928	0.6903
	One Hot Encoded	0.7251	0.7219	0.6470	0.6438	0.7461	0.7468	0.6930	0.6915
	One Hot Encoded+ PCA(components=13)	0.7251	0.7219	0.6470	0.6438	0.7461	0.7468	0.6930	0.6915

LOF Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Logistic Regression	Standard Data	0.7311	0.7350	0.6810	0.6902	0.7575	0.7688	0.7172	0.7274
	One Hot Encoded	0.7316	0.7371	0.6808	0.6914	0.7586	0.7716	0.7176	0.7293
	One Hot Encoded+ PCA(components=13)	0.7316	0.7370	0.6808	0.6913	0.7586	0.7715	0.7176	0.7294

F1-Score

Model Parameters

- Solver: newton-cg
- Regularization Strength: 10

Models (Stats)

Support Vector Machine

Z-Score Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Support Vector Machines	Standard Data	0.7312	0.7281	0.6391	0.6338	0.7620	0.7642	0.6952	0.6929
	One Hot Encoded Data	0.7325	0.7274	0.6450	0.6393	0.7608	0.7595	0.6981	0.6942
	One Hot Encoded+ PCA(components=15)	0.7330	0.7276	0.6456	0.6395	0.7614	0.7597	0.6988	0.6944

LOF Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Support Vector Machines	Standard Data	0.7370	0.7388	0.6821	0.6892	0.7669	0.7760	0.7220	0.7300
	One Hot Encoded	0.7378	0.7392	0.6834	0.6900	0.7675	0.7761	0.7230	0.7305
	One Hot Encoded+ PCA(components=12)	0.7377	0.7388	0.6838	0.6899	0.7672	0.7754	0.7231	0.7302

Model Parameters

- Regularisation Strength: 1
- Kernel: rbf

Models (Stats)

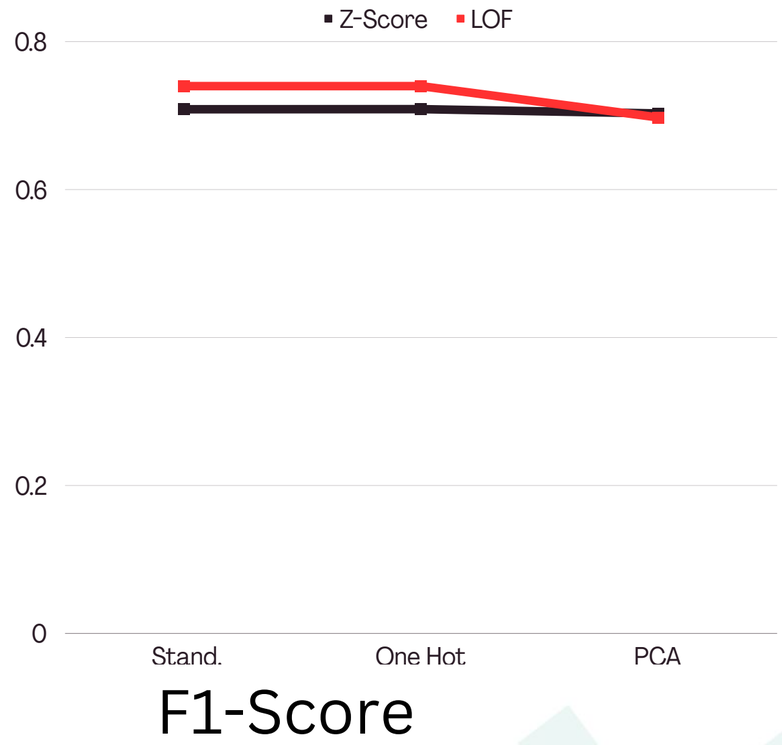
Decision Tree

Z-Score Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data + GridSearchCV

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
decision tree	standard	0.72554	0.72262	0.70317	0.69659	0.71860	0.72093	0.71080	0.70855
	standard + onehot	0.72551	0.72272	0.70326	0.69690	0.71852	0.72095	0.71081	0.70872
	standard +onehot +pca + hyper	0.72336	0.71734	0.69845	0.69076	0.71737	0.71546	0.70779	0.70289

LOF Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Decision Tree	Standard Data	0.73479	0.73877	0.72631	0.73545	0.74195	0.74429	0.73405	0.73984
	One Hot Encoded	0.73479	0.73877	0.72631	0.73545	0.74195	0.74429	0.73405	0.73984
	One Hot Encoded+ PCA(components=12)	0.73747	0.72298	0.67360	0.65964	0.75303	0.73985	0.71110	0.69745



F1-Score

Model Parameters
- Max Depth: 7
- Min Samples Split: 5
- Min Samples Leaf: 1
- Gini impurity

Models (Stats)

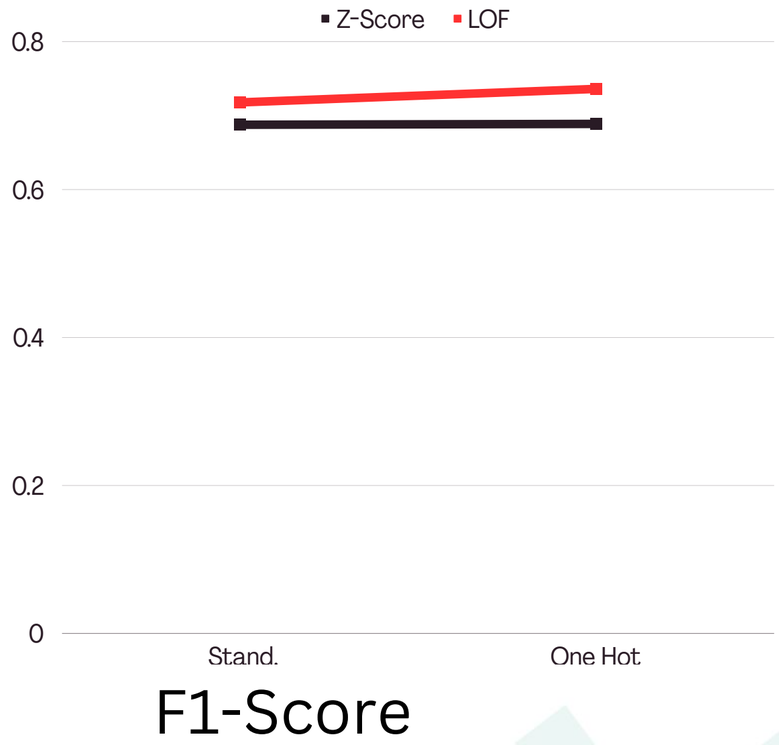
Random Forest

Z-Score Data
- Standard Data
- One Hot Encoded Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Random forest	standard	0.99806	0.70458	0.99693	0.67171	0.99903	0.70429	0.99798	0.68761
	standard +onehot	0.99780	0.70602	0.99665	0.6720	0.99876	0.70635	0.99770	0.68876

LOF Data
- Standard Data
- One Hot Encoded Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Random forest	Standard Data	0.99787	0.71525	0.99675	0.71708	0.99902	0.71855	0.99788	0.71781
	One Hot Encoded	0.79112	0.74132	0.75472	0.71420	0.81684	0.75932	0.78455	0.73607



F1-Score

Model Parameters

- max depth=15
- max features='log2'
- min samples leaf=4
- Gini impurity
- n estimators=300
- Bootstrap=True

Models (Stats)

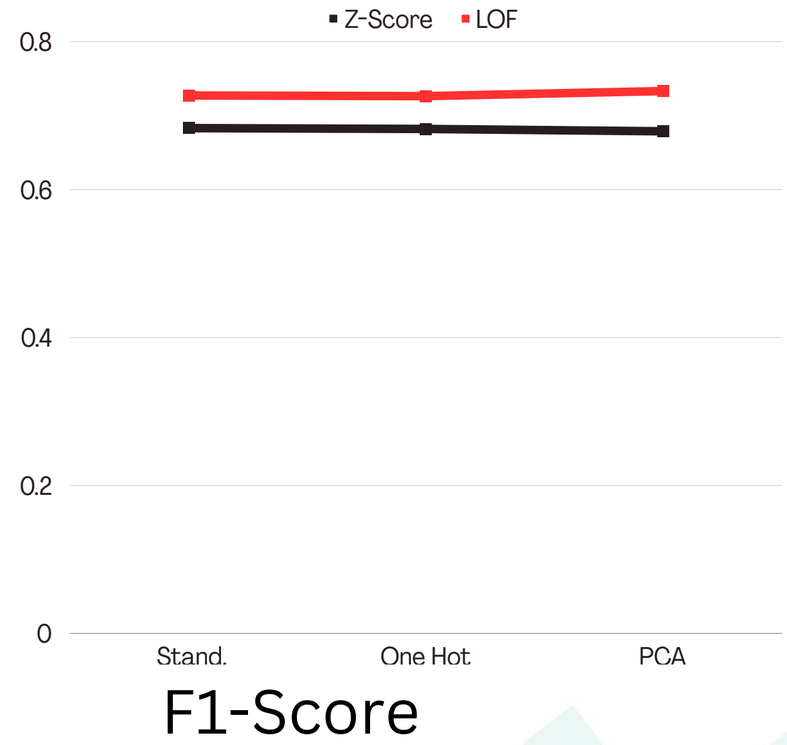
XGBoost

Z-Score Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Xgboost	standard	0.72575	0.72282	0.76239	0.76486	0.62217	0.61708	0.68518	0.68307
	standard +onehot	0.72611	0.72144	0.76234	0.76222	0.62332	0.61698	0.68586	0.68195
	standard + one-hot+pca	0.72448	0.72047	0.76403	0.76483	0.61581	0.6101	0.68196	0.67875

LOF Data
- Standard Data
- One Hot Encoded Data
- One Hot + PCA Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Xgboost	Standard Data	0.73305	0.73659	0.76171	0.76237	0.68434	0.69511	0.72095	0.72719
	One Hot Encoded	0.73380	0.73701	0.76413	0.76553	0.68238	0.69091	0.72094	0.72631
	One Hot Encoded+ PCA(components=12)	0.74697	0.73853	0.76718	0.75645	0.71481	0.71132	0.74007	0.73319



F1-Score

Model Parameters

- learning rate: 0.2
- max depth: 3
- n estimators: 100

Models (Stats)

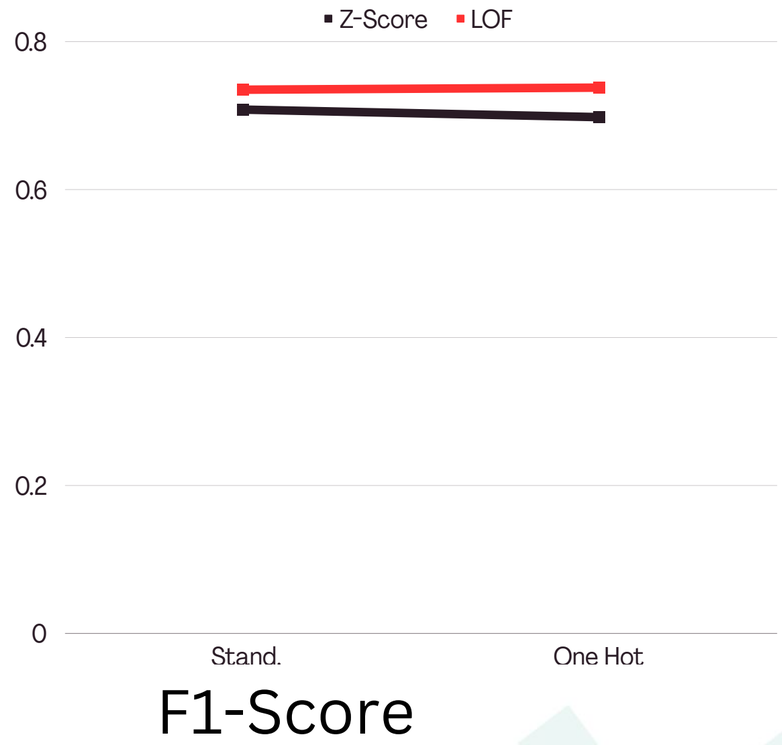
Multi Layer Perceptron

Z-Score Data
- Standard Data
- One Hot Encoded Data + GridSearchCV

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
MLP	standard	0.73507	0.72482	0.69964	0.68992	0.73522	0.72750	0.71699	0.70821
	one hot encoded + hyperparameter	0.74256	0.72564	0.67057	0.65488	0.76388	0.74710	0.71419	0.69795

LOF Data
- Standard Data
- One Hot Encoded Data

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Multi-Layer Perceptron	Standard Data	0.74419	0.73743	0.71996	0.72116	0.75980	0.74950	0.73935	0.73505
	One Hot Encoded	0.74629	0.73786	0.73172	0.73040	0.75677	0.74543	0.74403	0.73784



F1-Score

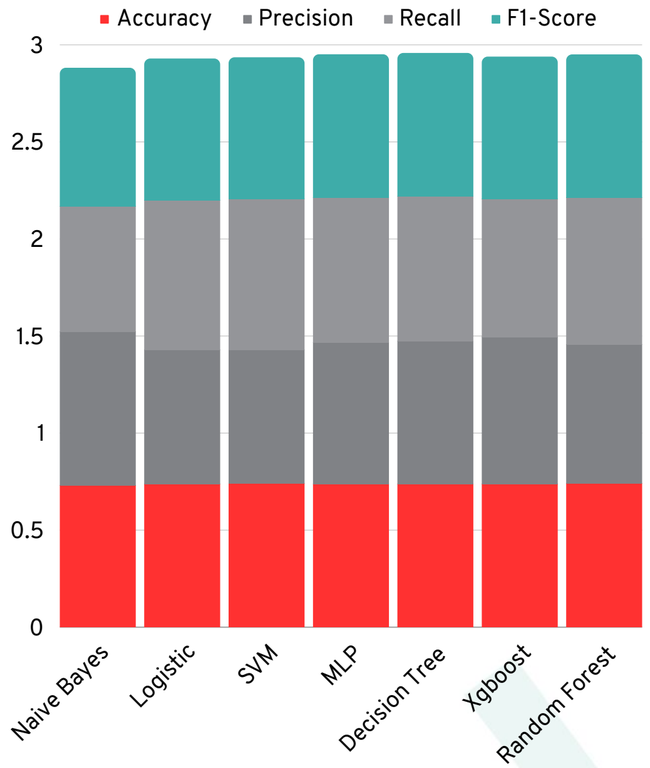
Model Parameters

- Neurons in hidden layer-1 : 300
- Neurons in hidden layer-2: 250

Stats - Z-Score

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Naïve bayes	Standard Data	0.72097	0.71919	0.75703	0.75921	0.61599	0.61486	0.67927	0.67945
	One Hot Encoded	0.71690	0.71396	0.75209	0.75294	0.61132	0.60882	0.67444	0.67326
	One Hot Encoded + PCA(components=3)	0.71260	0.71053	0.76114	0.76335	0.58417	0.58257	0.66102	0.66082
Logistic	Standard Data	0.72444	0.72067	0.64800	0.64334	0.74442	0.74482	0.69287	0.69037
	One Hot Encoded	0.7251	0.72195	0.64704	0.64387	0.74614	0.74680	0.69306	0.69152
	One Hot Encoded + PCA(components=13)	0.7251	0.72195	0.64704	0.64387	0.74614	0.74680	0.69306	0.69152
SVM	Standard Data	0.73120	0.72810	0.63916	0.63381	0.76209	0.76423	0.69523	0.6929
	One Hot Encoded	0.73252	0.7274	0.64511	0.63931	0.76088	0.75952	0.69823	0.69425
	One Hot Encoded + PCA(components=14)	0.73307	0.72769	0.64580	0.63963	0.76148	0.75980	0.6988	0.69456
Xgboost	Standard Data	0.72575	0.72282	0.76239	0.76486	0.62217	0.61708	0.68518	0.68307
	One Hot Encoded	0.72611	0.72144	0.76234	0.76222	0.62332	0.61698	0.68586	0.68195
	One Hot Encoded + PCA(components=8)	0.72448	0.72047	0.76403	0.76483	0.61581	0.6101	0.68196	0.67875
Decision tree	Standard Data	0.72554	0.72262	0.70317	0.69659	0.71860	0.72093	0.71080	0.70855
	One Hot Encoded	0.72551	0.72272	0.70326	0.69690	0.71852	0.72095	0.71081	0.70872
	One Hot Encoded + PCA(components=12)	0.72336	0.71734	0.69845	0.69076	0.71737	0.71546	0.70779	0.70289
Random for-est	Standard	0.99806	0.70458	0.99693	0.67171	0.99903	0.70429	0.99798	0.68761
	One Hot Encoded	0.99780	0.70602	0.99665	0.6720	0.99876	0.70635	0.99770	0.68876
MLP	Standard Data	0.73507	0.72482	0.69964	0.68992	0.73522	0.72750	0.71699	0.70821
	One Hot Encoded	0.74256	0.72564	0.67057	0.65488	0.76388	0.74710	0.71419	0.69795

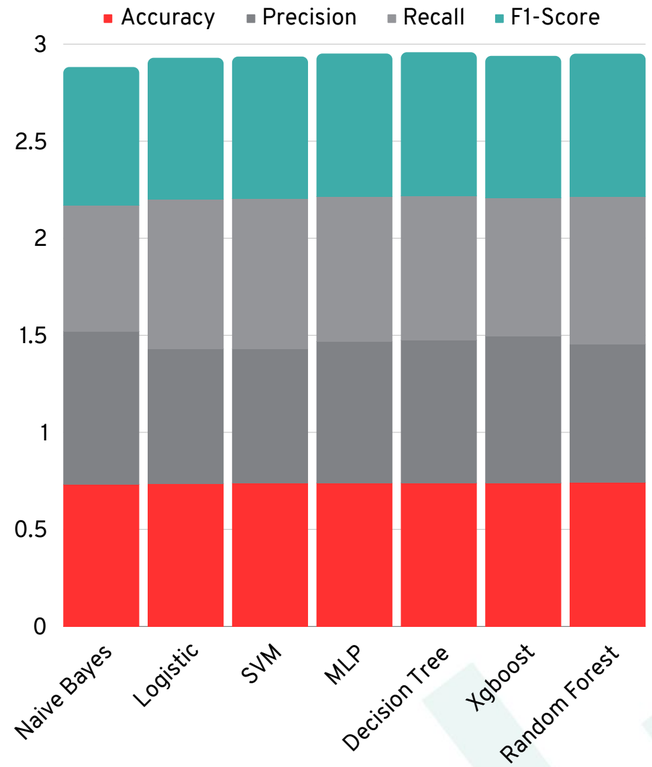
Table 2. Metrics on the dataset cleaned using Z-Score



Stats - LOF

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Naïve Bayes	Standard Data	0.72788	0.73089	0.77897	0.77872	0.64221	0.65262	0.70401	0.71012
	One Hot Encoded	0.72385	0.72555	0.77291	0.77273	0.64004	0.64686	0.70023	0.70421
	One Hot Encoded+ PCA(components=3)	0.72141	0.73022	0.77009	0.77765	0.63747	0.65238	0.69753	0.70953
Logistic Regression	Standard Data	0.73221	0.73495	0.68475	0.69379	0.76007	0.76043	0.72045	0.72558
	One Hot Encoded	0.73229	0.73610	0.68408	0.69367	0.76059	0.76240	0.72031	0.72642
	One Hot Encoded+ PCA(components=13)	0.73229	0.73622	0.68403	0.69391	0.76062	0.76246	0.72029	0.72658
SVM	Standard Data	0.73666	0.73883	0.68243	0.69199	0.76898	0.76795	0.72313	0.72800
	One Hot Encoded	0.73822	0.73919	0.68588	0.69451	0.76957	0.76707	0.72532	0.72899
	One Hot Encoded+ PCA(components=13)	0.73837	0.73889	0.68691	0.69499	0.76921	0.76628	0.72574	0.72890
Xgboost	Standard Data	0.73305	0.73659	0.76171	0.76237	0.68434	0.69511	0.72095	0.72719
	One Hot Encoded	0.73380	0.73701	0.76413	0.76553	0.68238	0.69091	0.72094	0.72631
	One Hot Encoded+ PCA(components=12)	0.74697	0.73853	0.76718	0.75645	0.71481	0.71132	0.74007	0.73319
Decision Tree	Standard Data	0.73479	0.73877	0.72631	0.73545	0.74195	0.74429	0.73405	0.73984
	One Hot Encoded	0.73479	0.73877	0.72631	0.73545	0.74195	0.74429	0.73405	0.73984
	One Hot Encoded+ PCA(components=12)	0.73747	0.72298	0.67360	0.65964	0.75303	0.73985	0.71110	0.69745
Random forest	Standard Data	0.99787	0.71525	0.99675	0.71708	0.99902	0.71855	0.99788	0.71781
	One Hot Encoded	0.79112	0.74132	0.75472	0.71420	0.81684	0.75932	0.78455	0.73607
MLP	Standard Data	0.74419	0.73743	0.71996	0.72116	0.75980	0.74950	0.73935	0.73505
	One Hot Encoded	0.74629	0.73786	0.73172	0.73040	0.75677	0.74543	0.74403	0.73784

Table 3. Metrics on the dataset cleaned using LOF



Analysis

- Using F1 Score or Accuracy as evaluation metric: Tradeoff between Precision and Recall.
- False Negatives are more harmful than false positives in our case. Thus, a higher recall is preferred.
- Best Recall was observed in the case of the **Support Vector Machine (0.76921),** obtained after applying PCA on one hot encoded data cleaned using LOF after tuning the hyperparameters.
- It can be observed that the F1 scores of all the models trained on the data with LOF as a method of outlier removal were higher than the models trained on data with Z-Score as the outlier removal method.
- PCA was not applied to MLP since it is capable of handling complex relationships between the features and the data points. Applying PCA preserves the global structure, whereas the local structure is distorted. However, the initial layers of the MLP are capable of recognizing the local structure within the data; hence, distorting it might render these initial layers useless, resulting in lower performance metrics.

Conclusions

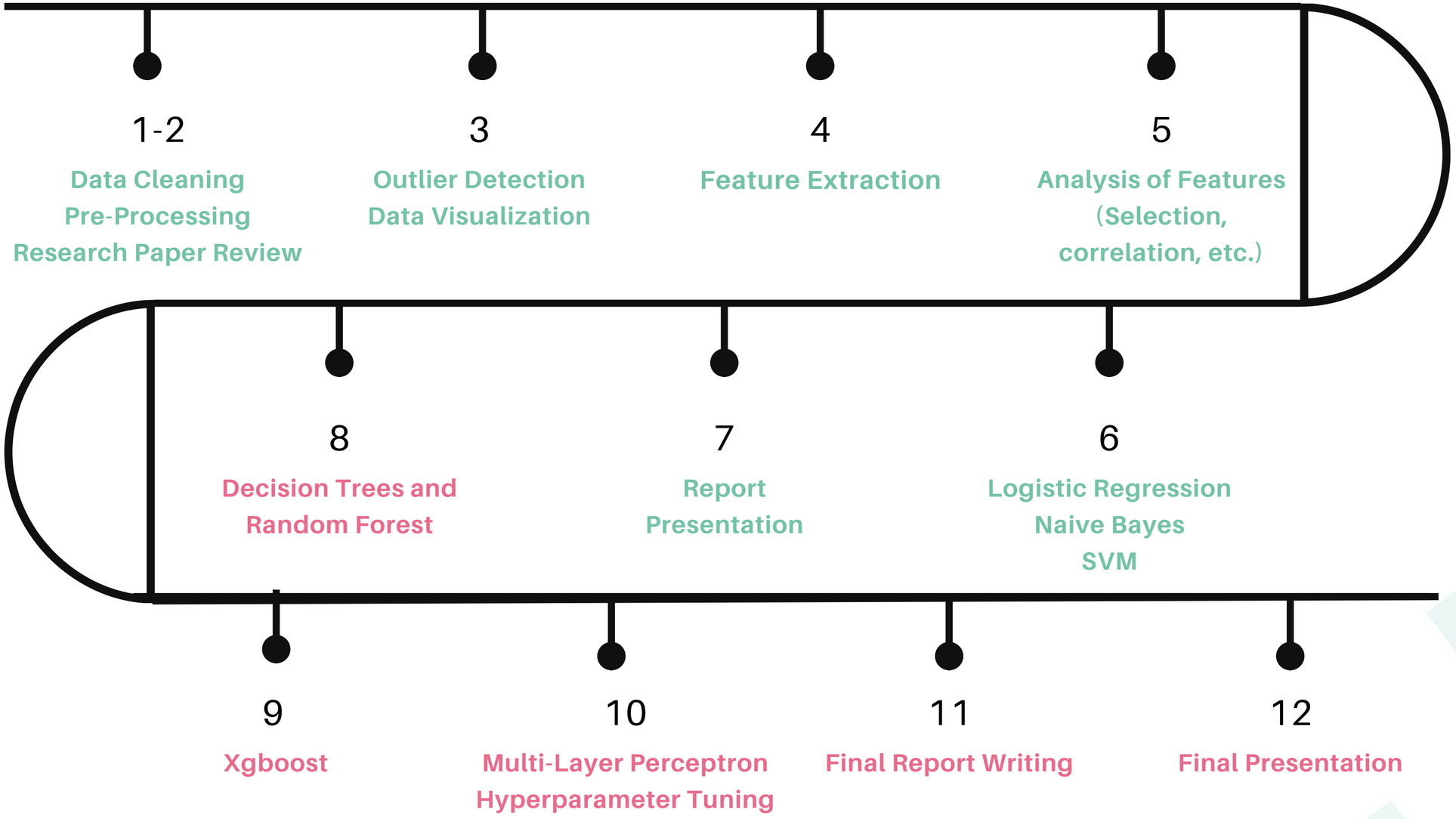
- From the above analysis, it can be concluded that out of all the models applied, the best performance (considering the F1 score as the overall evaluation metric) was achieved by the **Decision Tree (0.73984)** followed by **Multi-Layer Perceptron (0.73784)**.

- If we observe the accuracies of the model, we can observe that the best accuracy was observed in the case of **Random Forest** with an accuracy of **0.74132,** which was closely followed by **Multi-Layer Perceptron 0.73919.**

Our Learnings

- Solving some complex machine learning problems using classical machine learning models.
- Learning how to perform Exploratory Data Analysis, in order to gain valuable insights into the dataset.
- Learning how to set parameters for the model while balancing the tradeoff between computational complexity and performance.
- Applying different regularization techniques, while ensuring a proper balance between bias and variance.

Timeline


1-2
Data Cleaning
Pre-Processing
Research Paper Review
3
Outlier Detection
Data Visualization
4
Feature Extraction
5
Analysis of Features
(Selection,
correlation, etc.)
6
Logistic Regression
Naive Bayes
SVM
7
Report
Presentation
8
Decision Trees and
Random Forest
9
Xgboost
10
Multi-Layer Perceptron
Hyperparameter Tuning
11
Final Report Writing
12
Final Presentation

Contributions

Arnav Gupta

2021236

Contributions:

- Data Cleaning
- Multivariate Exploratory Data Analysis
- Outlier Detection using Z-Score
- Support Vector Machines
- Decision Tree Classifier
- Slides

Karan Gupta

2021258

Contributions:

- Data Cleaning
- Multivariate Exploratory Data Analysis
- Naive Bayes
- Random Forest
- Report Writing
- Slides.

Shivesh Gulati

2021286

Contributions:

- Literature Review (RP-2)
- Univariate Exploratory Data Analysis
- Logistic Regression
- Multi Layer Perceptron
- Report Writing
- Slides.

Vishal Singh

2021575

Contributions:

- Literature Review (RP-1)
- Outlier Detection using Local Outlier Factor
- Data Preprocessing
- Support Vector Machines
- XGBoost
- Slides.