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.

Outline

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

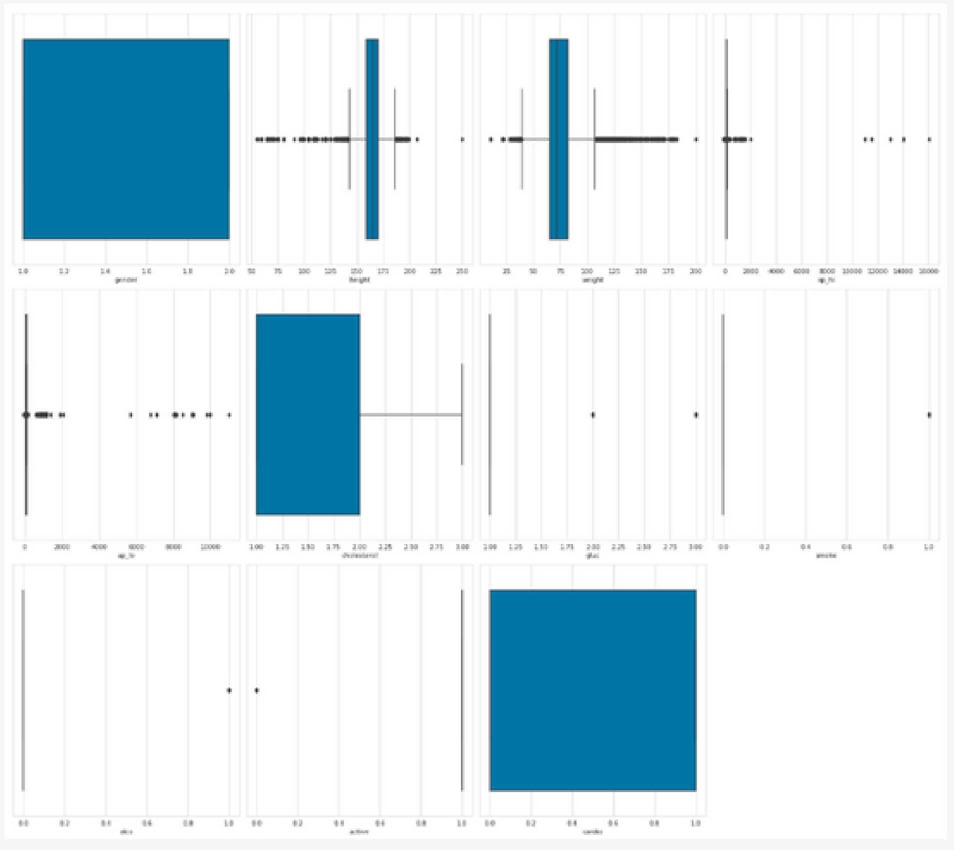
Literature Review

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.

Figure 2. Boxplots of all attributes.



kaggle

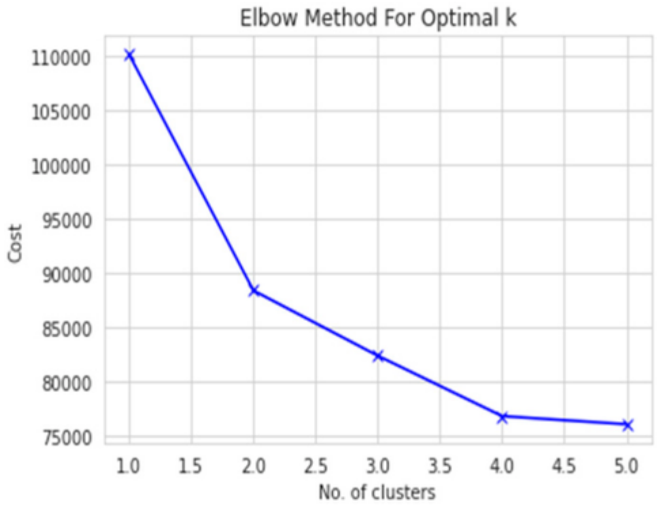
Images taken from : https://www.mdpi.com/1999-4893/16/2/88

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.



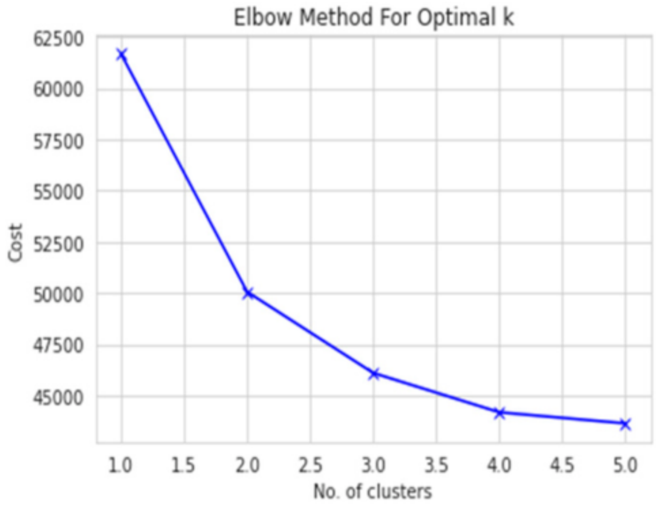


Table 5. The evaluation metrics resulting from different classifiers.

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

Outline

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

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

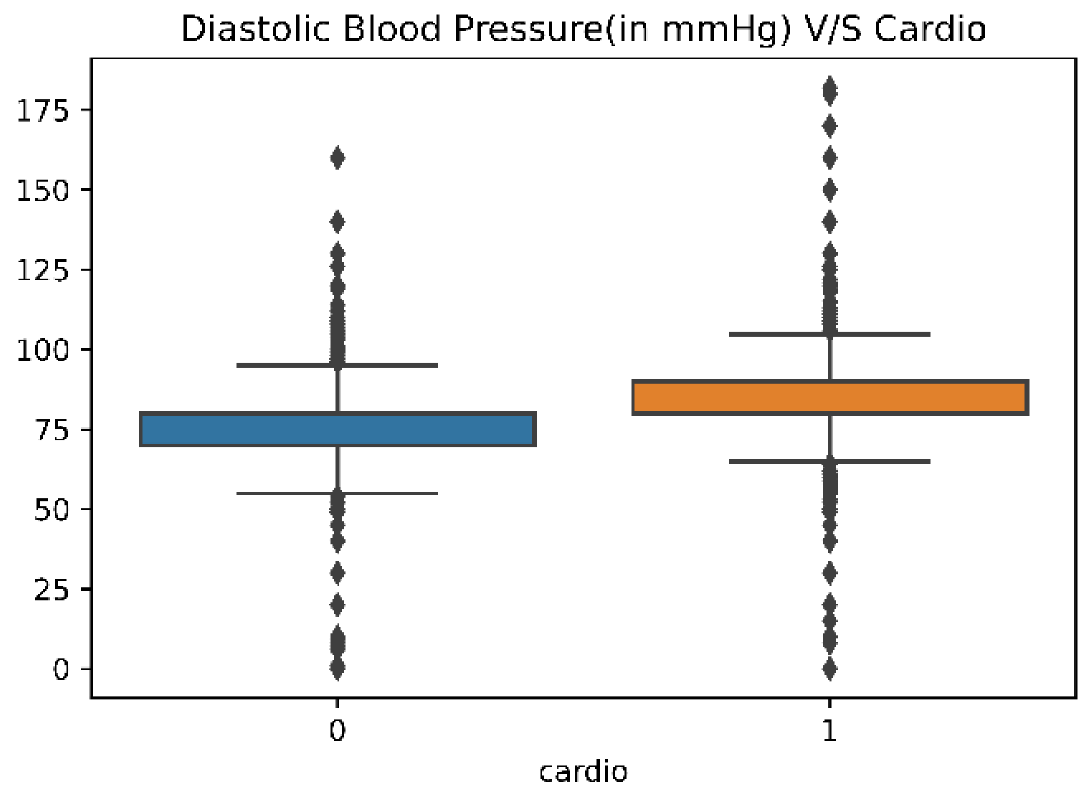
Feature	Description	Type of Feature	Data-Type	Unit of measurement	Value Range
age	Age	Objective Feature	integer	Days	Any Integer Value >= 0
height	Height	Objective Feature	integer	Centimeters (cm)	Any Integer Value > 0
weight	Weight	Objective Feature	float	Kilograms (kg)	Any Floating Value > 0
gender	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

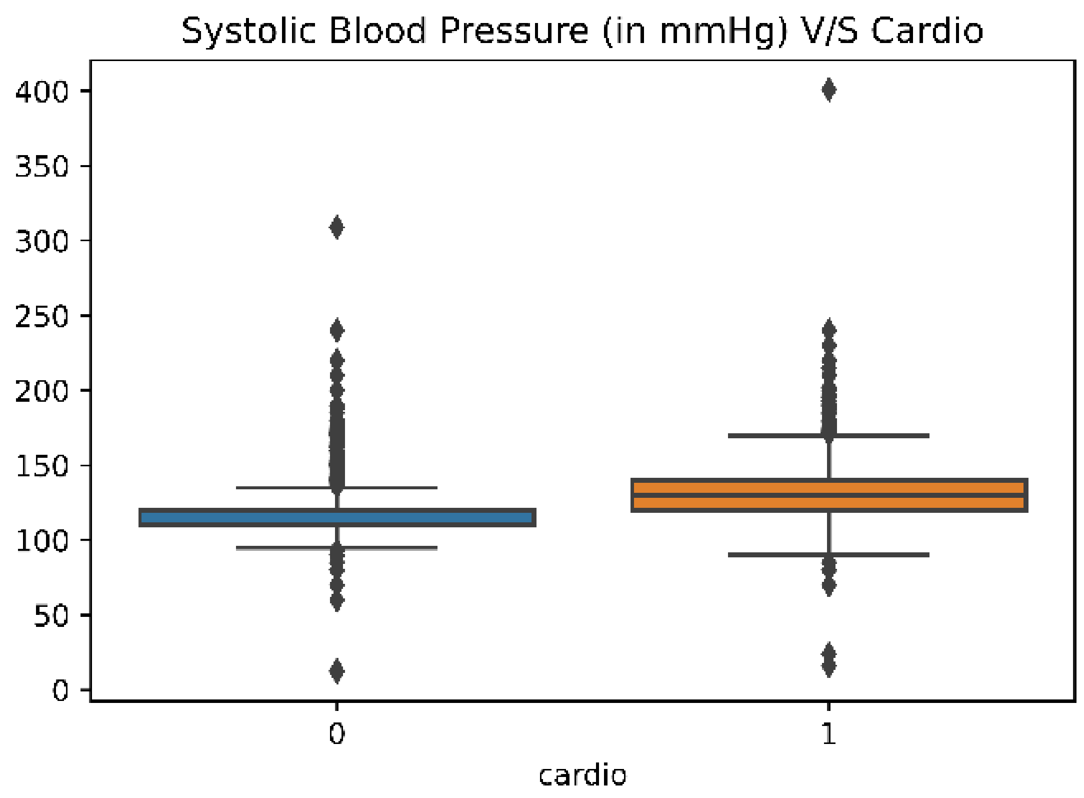
Table 1. Table describing the dataset features

Box Plots

Inference

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



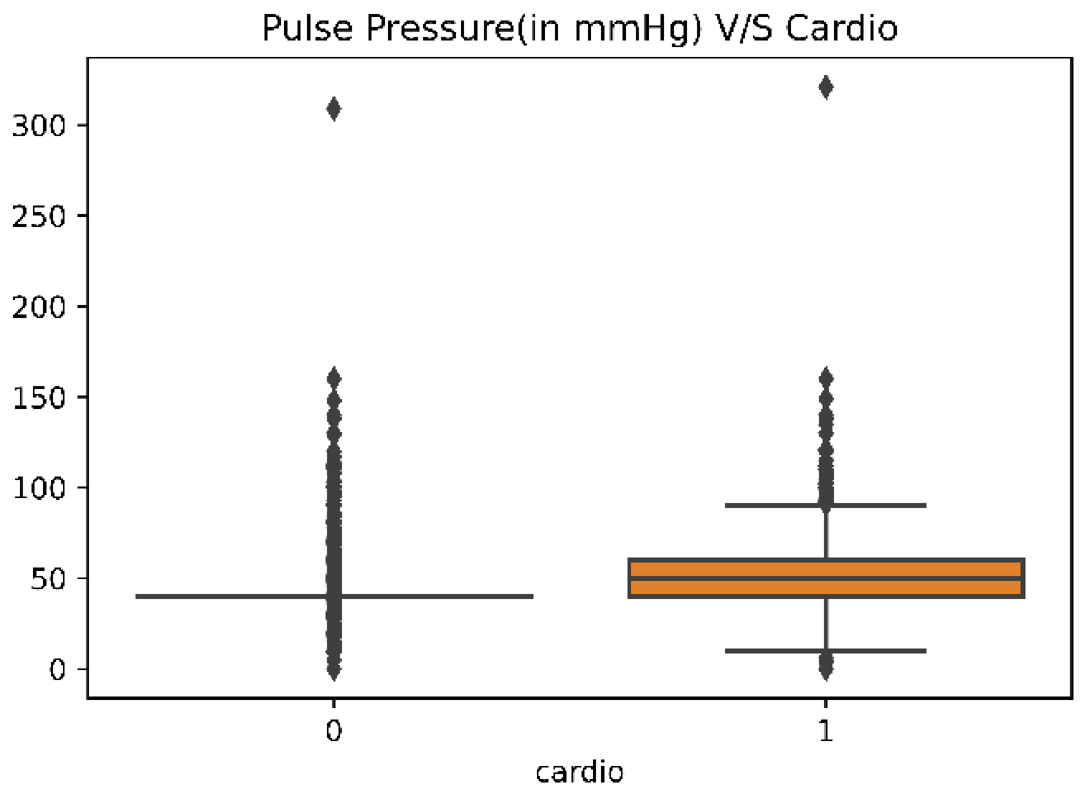


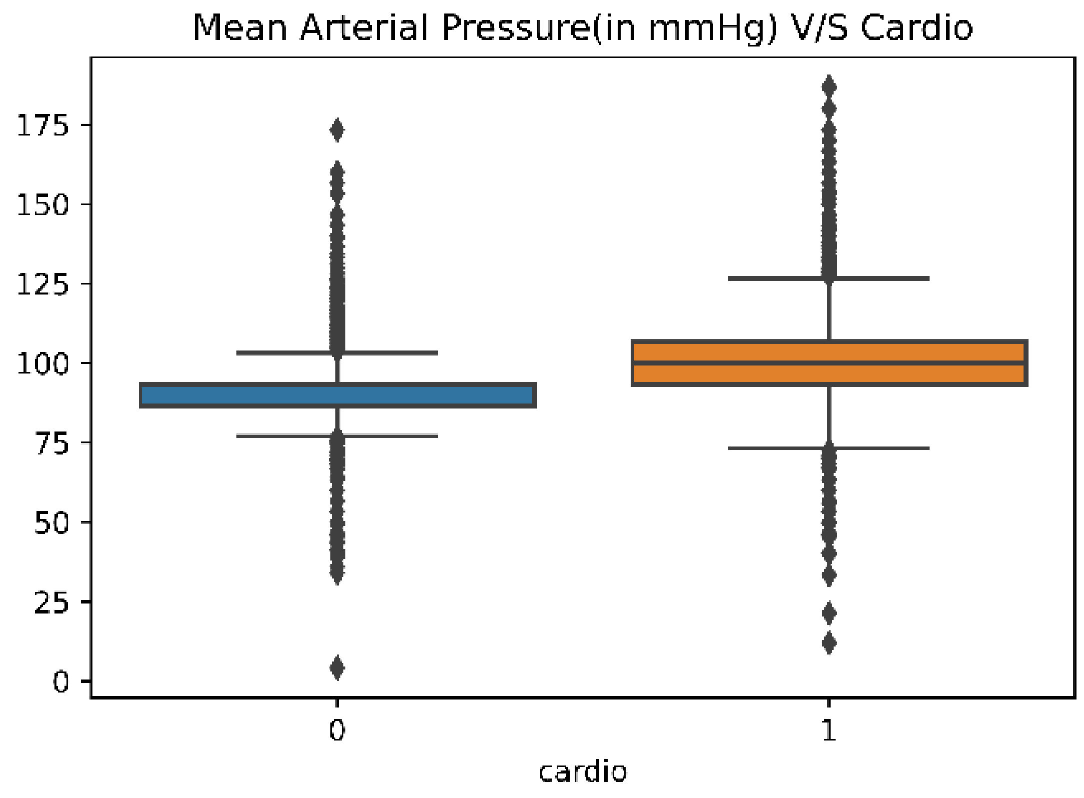
Inference

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.

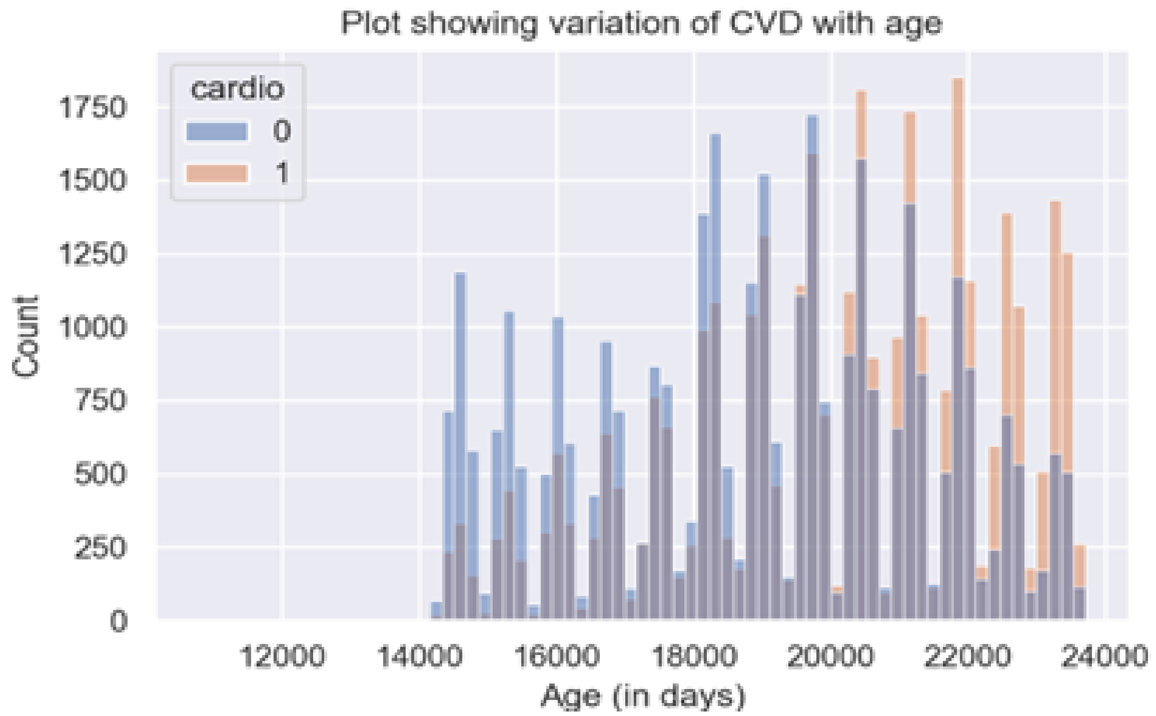




Inference

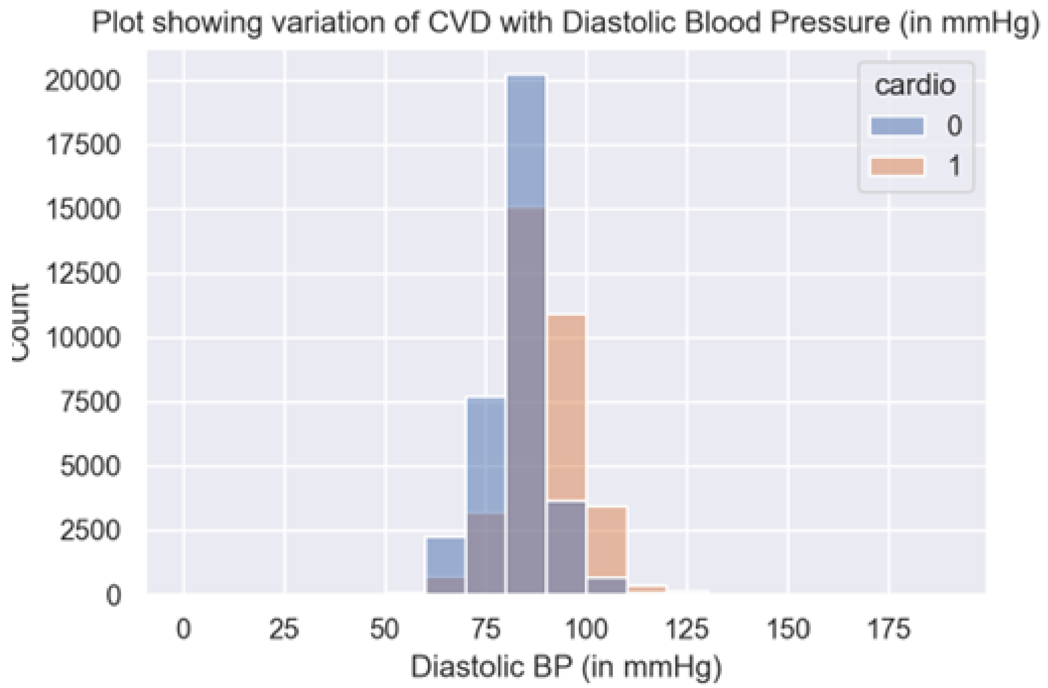
Mean Arterial Pressure, shows the least number of outliers, and has significant size of the box.

Histograms

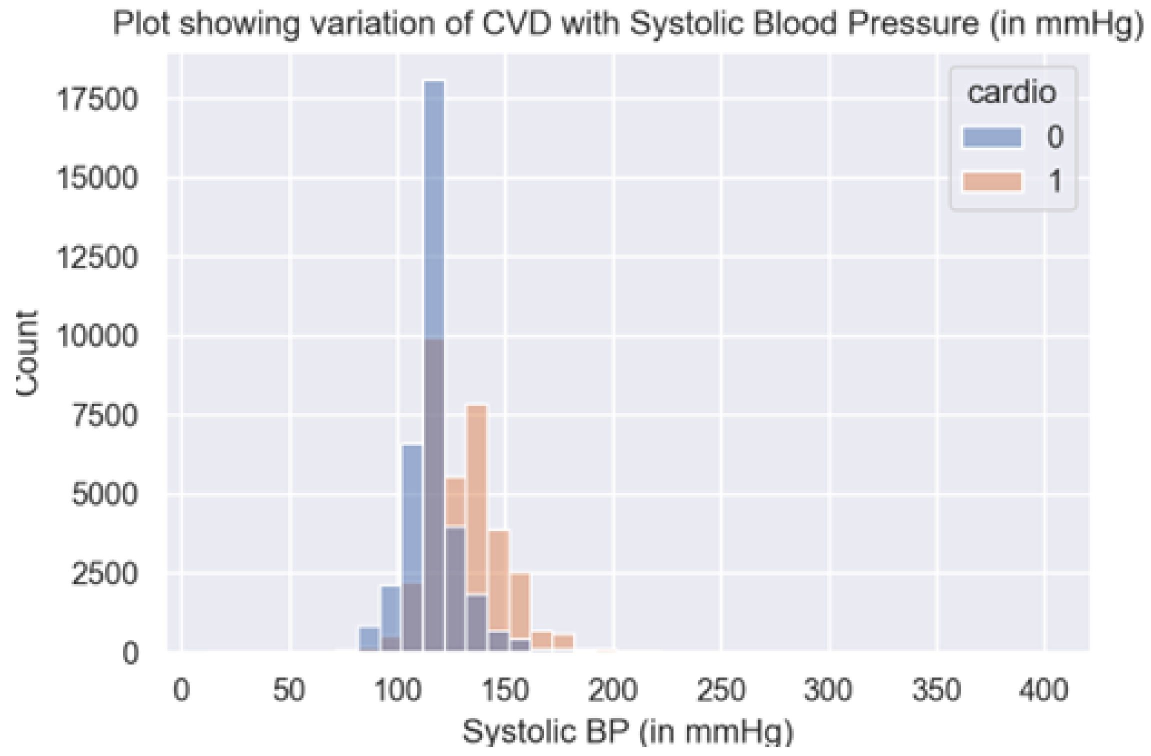


Graph-1

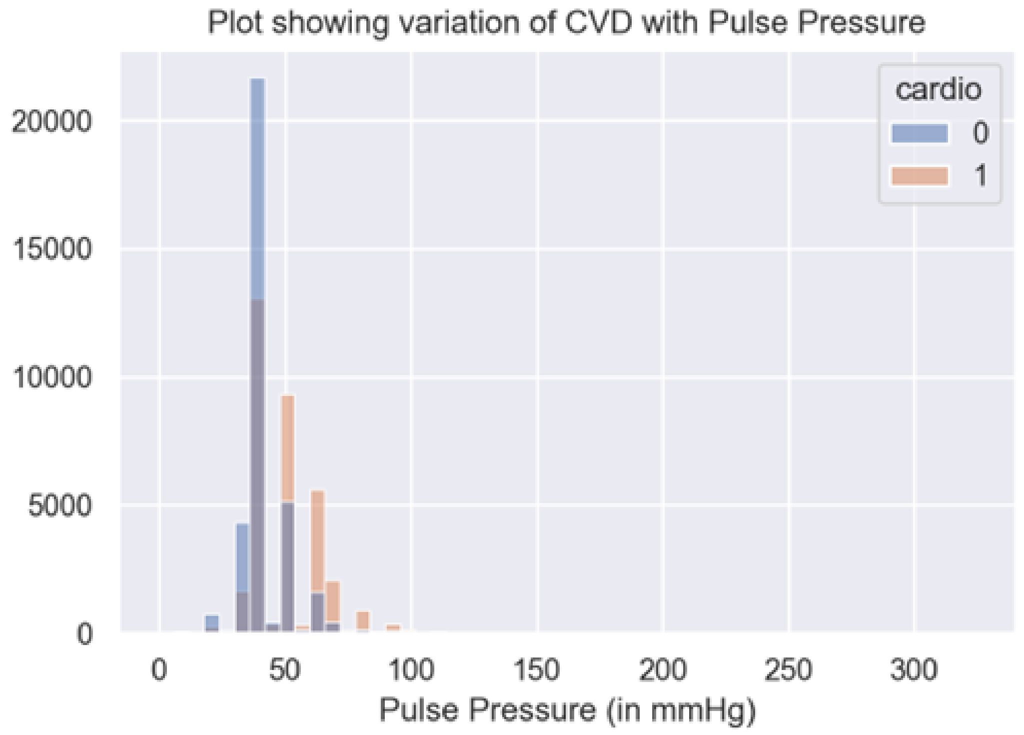
Graph-2



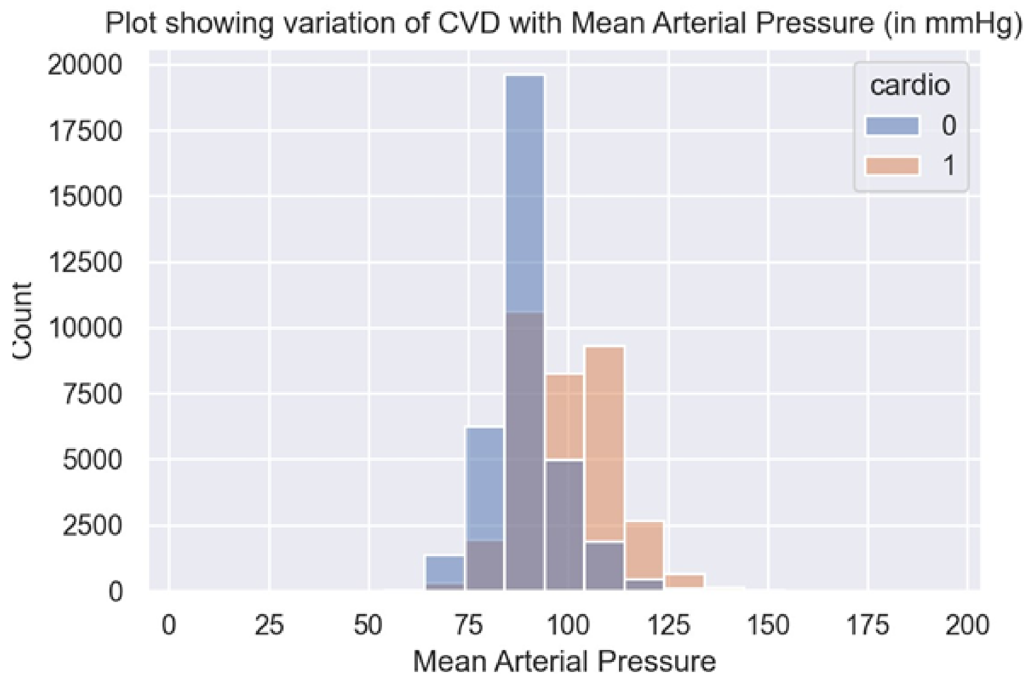
Graph-3



Graph-4



Graph-5



Graph-6

Inferences from Histograms

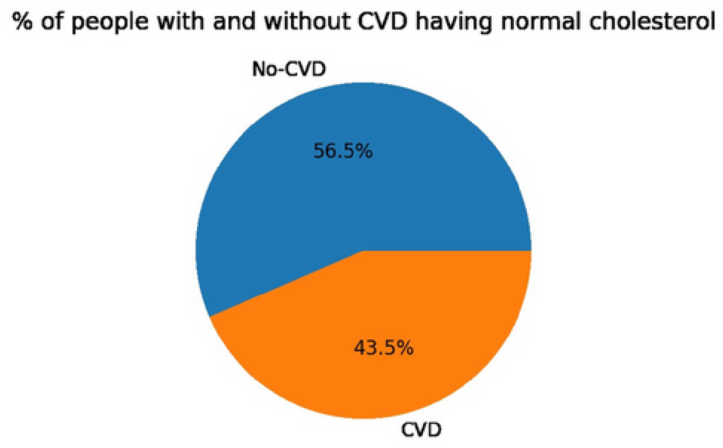
Histograms were drawn to analyze the distribution of participants, with or without CVD, in specific ranges of the various numerical features. The insights obtained are as follows:-

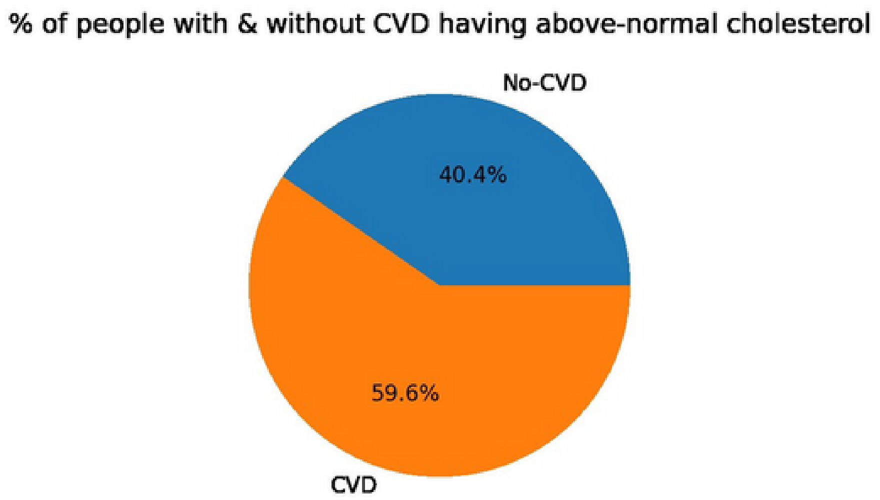
- As the age bracket increases, the number of people with CVD is higher than those without CVD, especially for people beyond 22500 days (61 years). **(See Figure 1)**

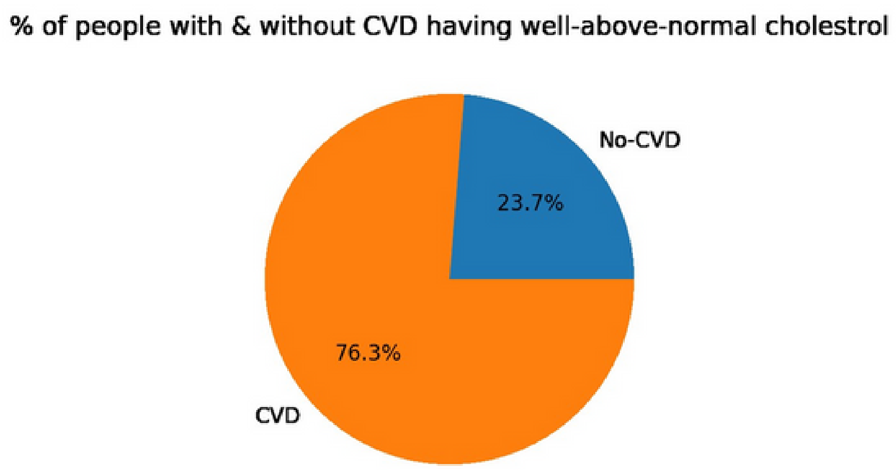
- As the weight bracket increases, the proportion of people having CVD is significantly higher as compared to those not having CVD, especially in the 75 kg and beyond range **(See Figure 2).**

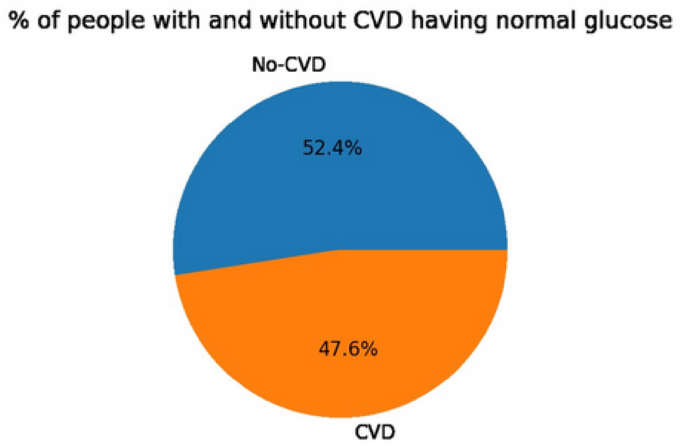
- In populations with higher systolic, diastolic, and pulse pressures, cardiovascular disease (CVD) frequency is higher. Specifically, systolic BP >=140 mmHg, diastolic BP>=90 mmHg, and pulse pressure >=50 mmHg are associated with increased CVD prevalence. **(See Figures 3,4,5,6)**.

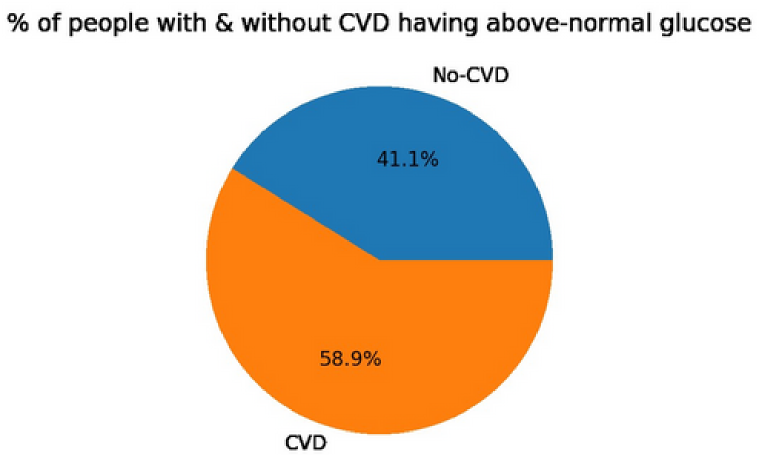
Pie Charts

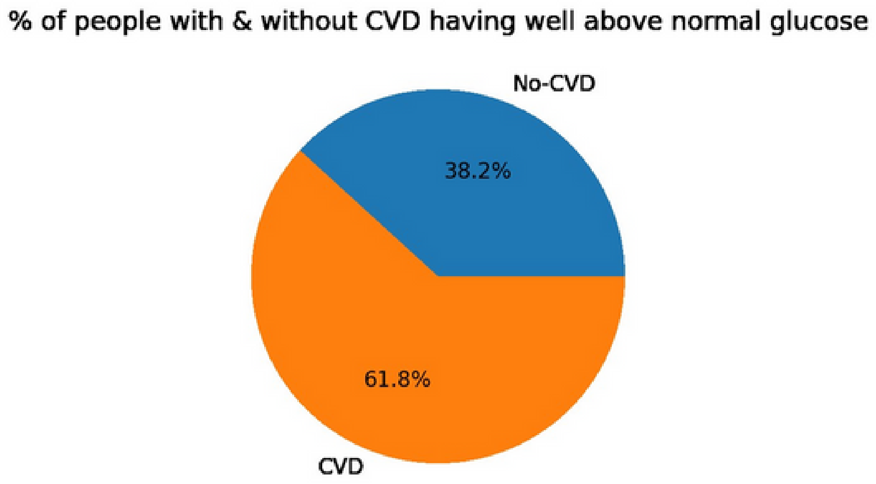


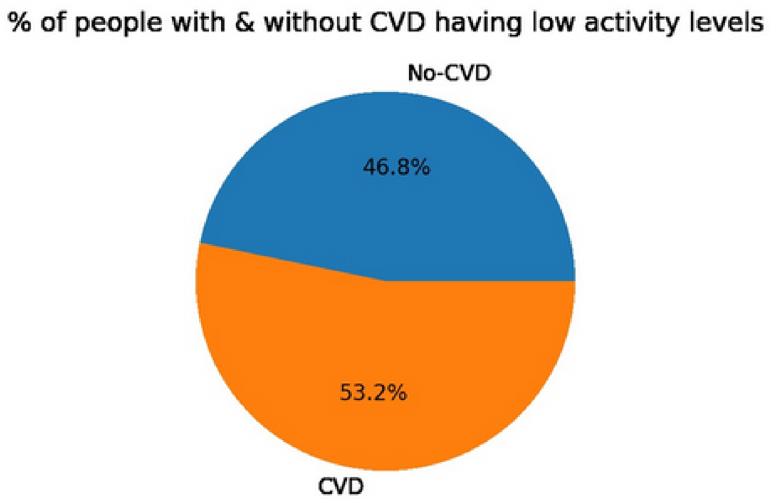


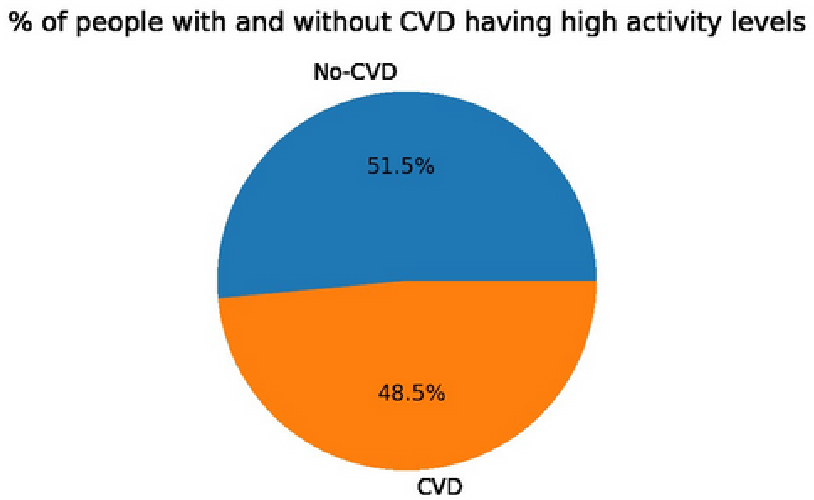












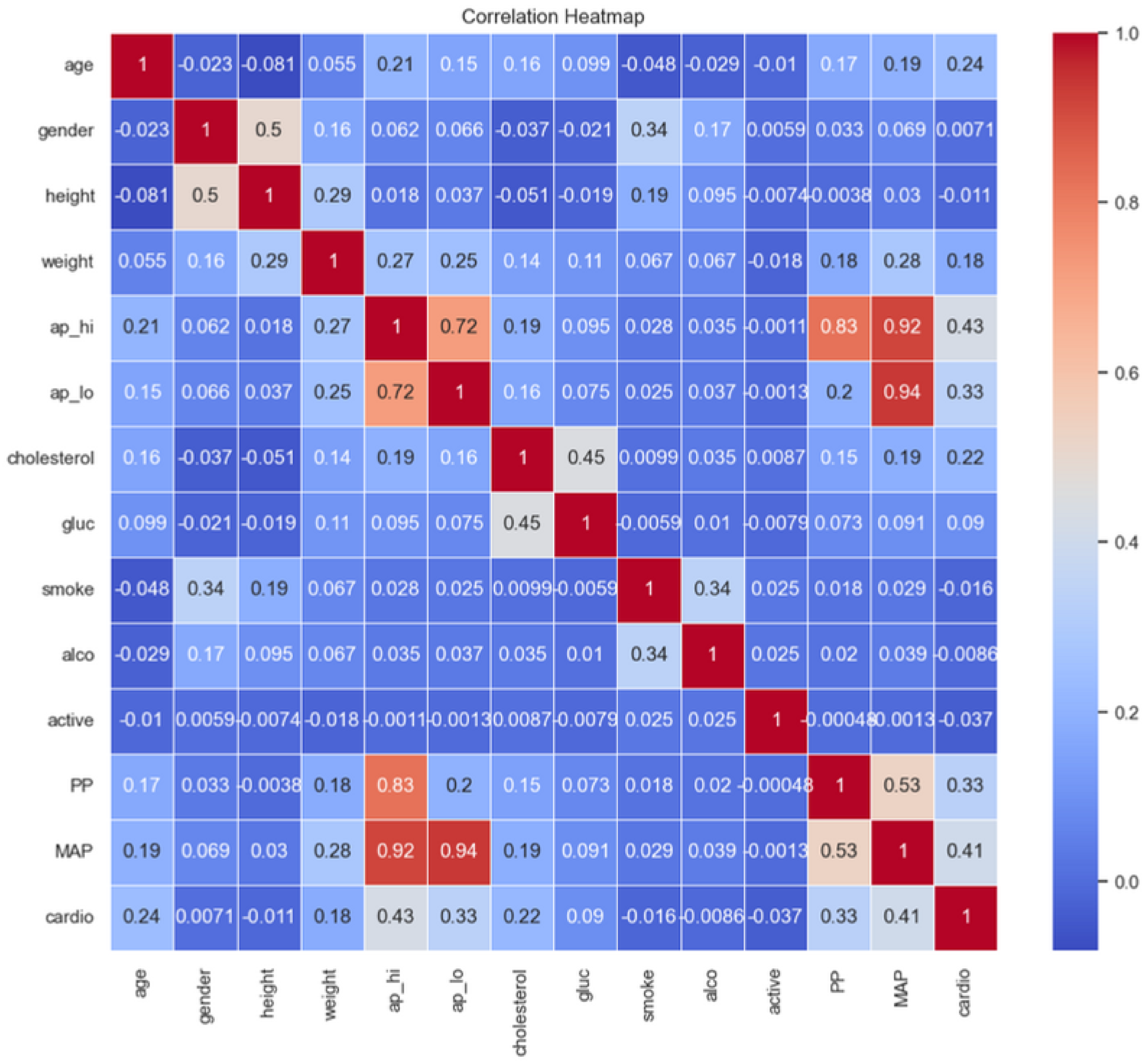
Inferences

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

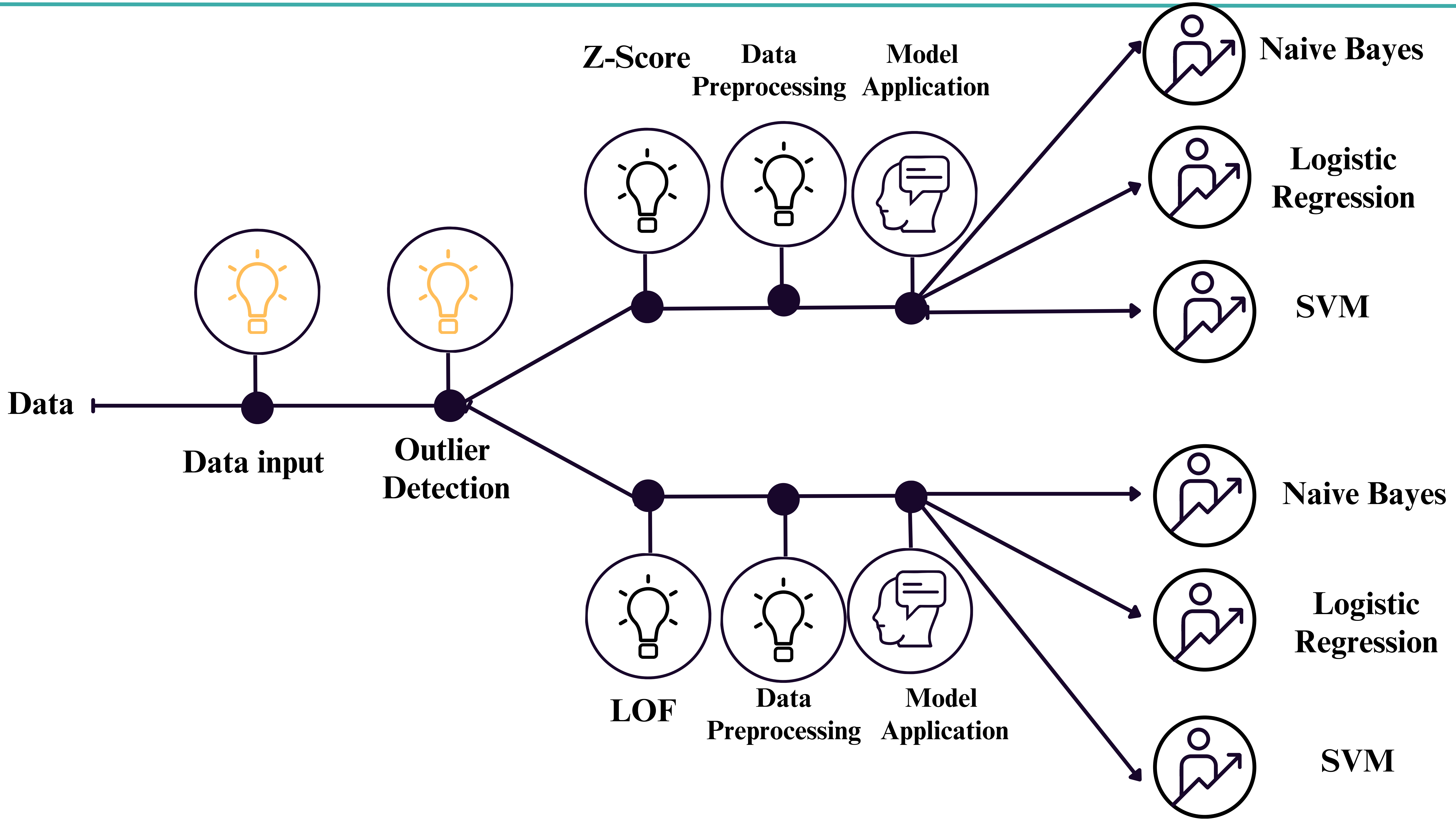


- Ap_lo and Ap_hi have a strong correlation(close to 1) as greater diastolic pressure means a greater systolic pressure.
- PP and MAP both have a strong correlation with ap_hi and ap_lo.
- Ap_hi and map have a moderate correlation (around 0.5) with CVD as higher blood pressure generally means a greater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

Outline

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

Flow Chart



Z-Score
Data Preprocessing
Model Application
Naive Bayes
Logistic Regression
SVM
Data
Data input
Outlier Detection
Naive Bayes
Logistic Regression
LOF
Data Preprocessing
Model Application
SVM

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

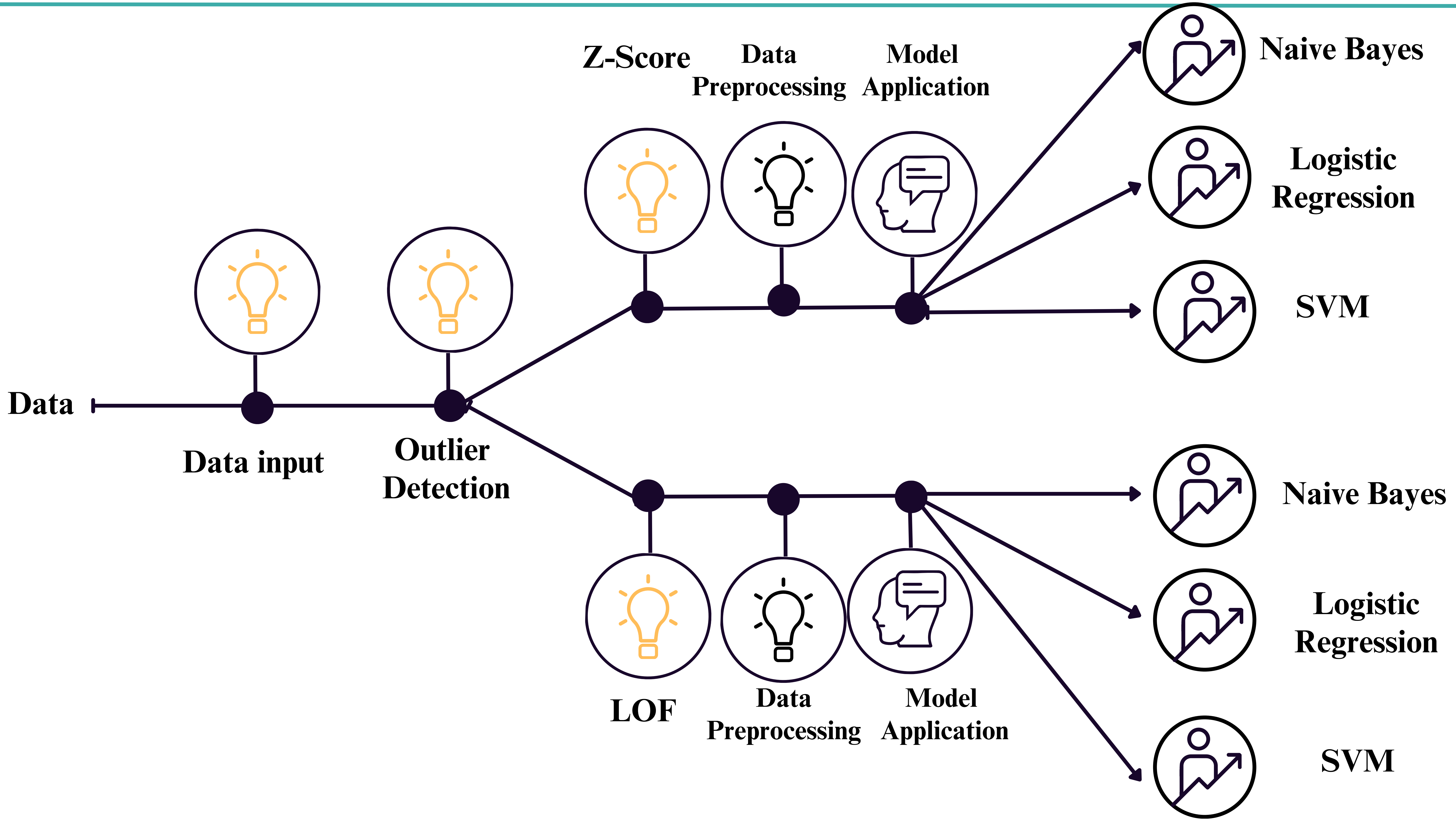
68,727 → 65,048

LOF

Applied Local Outlier Factor on the Manually Capped Data

68,727 → 54,981

Flow Chart



Data
Data input
Outlier Detection
Z-Score
Data Preprocessing
Model Application
Naive Bayes
Logistic Regression
SVM
LOF
Data Preprocessing
Model Application
Naive Bayes
Logistic Regression
SVM

Z-Score

Formula

$$Z = \frac{\overset{\text{Score}}{x} - \overset{\text{Mean}}{\mu}}{\underset{\text{SD}}{\sigma}}$$

Lower bound=-2.75

Upper bound=2.75

Final size: 65,048

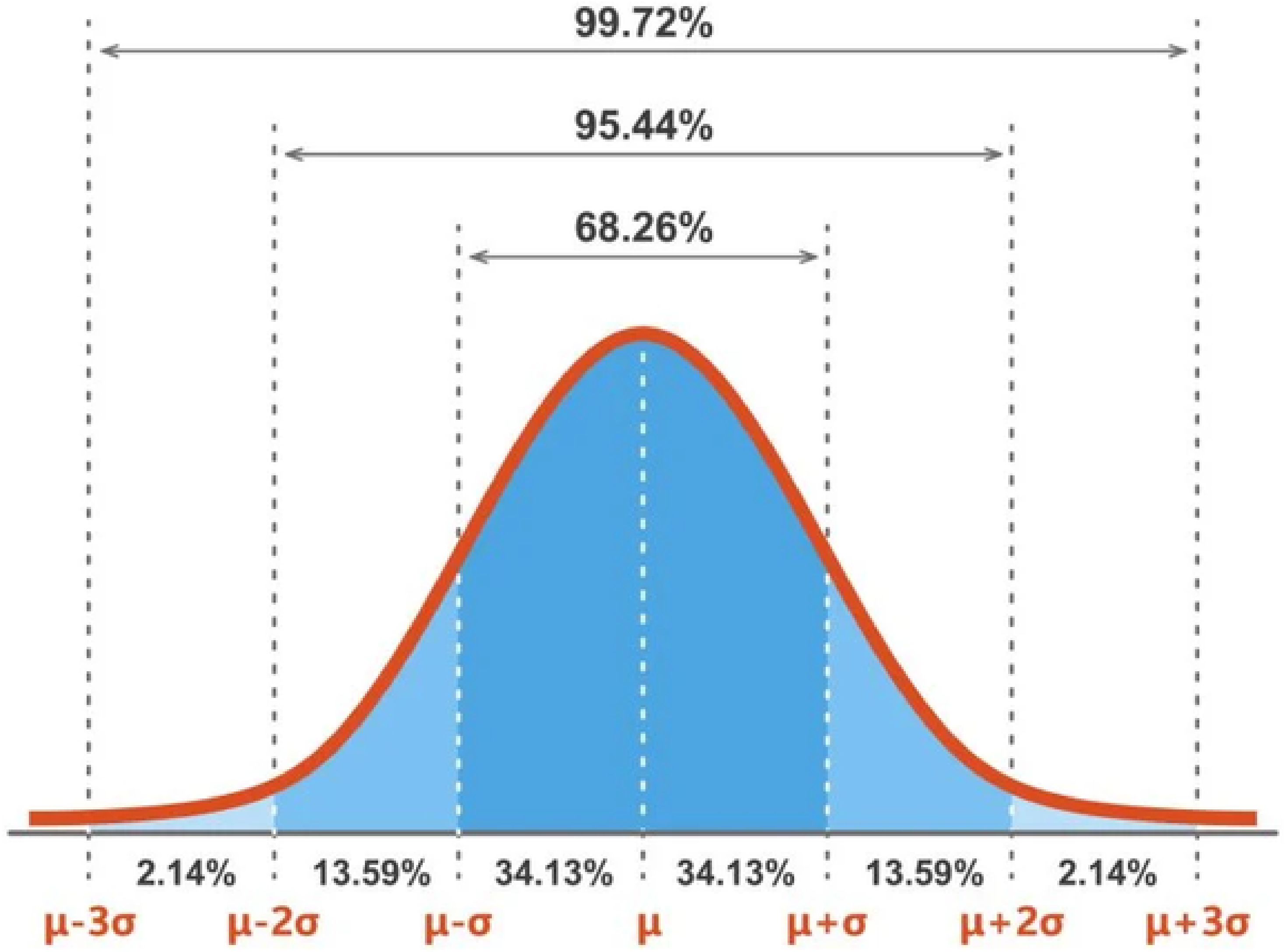


Image Source : https://www.simplypsychology.org/z-score.html

Local Outlier Factor

Formula

For a given Data set

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

Local Outlier Factor for each data point is given by

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

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

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

www.mlpoint.com

Final size: 54,981

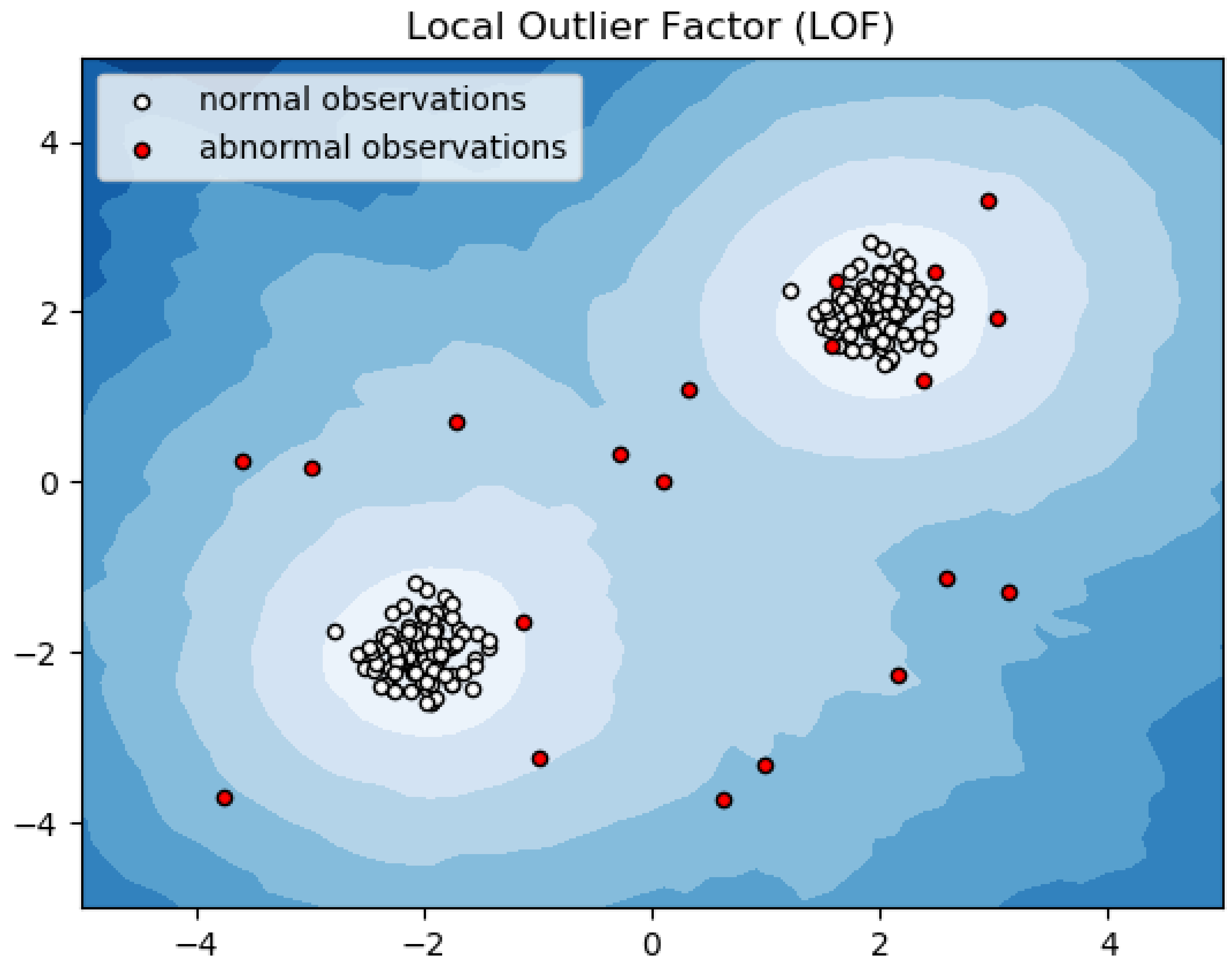
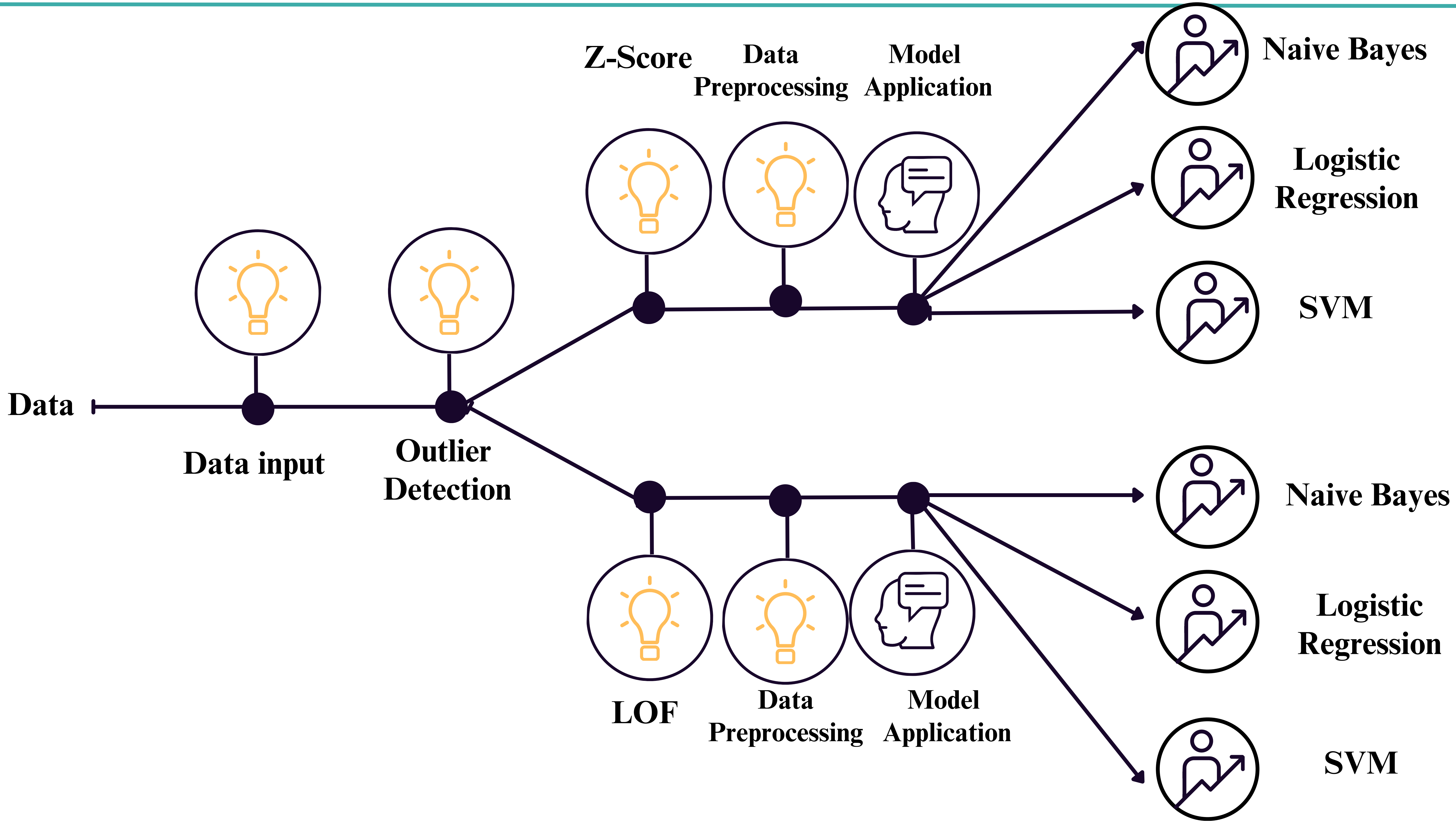


Image Source: https://scikit-learn.org/0.19/auto_examples/neighbors/plot_lof.html

Flow Chart



Z-Score
Data Preprocessing
Model Application
Naive Bayes
Logistic Regression
SVM
Data
Data input
Outlier Detection
Naive Bayes
Logistic Regression
LOF
Data Preprocessing
Model Application
SVM

Data Pre-Processing

Train-Test-Split

70:30

Data Standardization

StandardScaler()

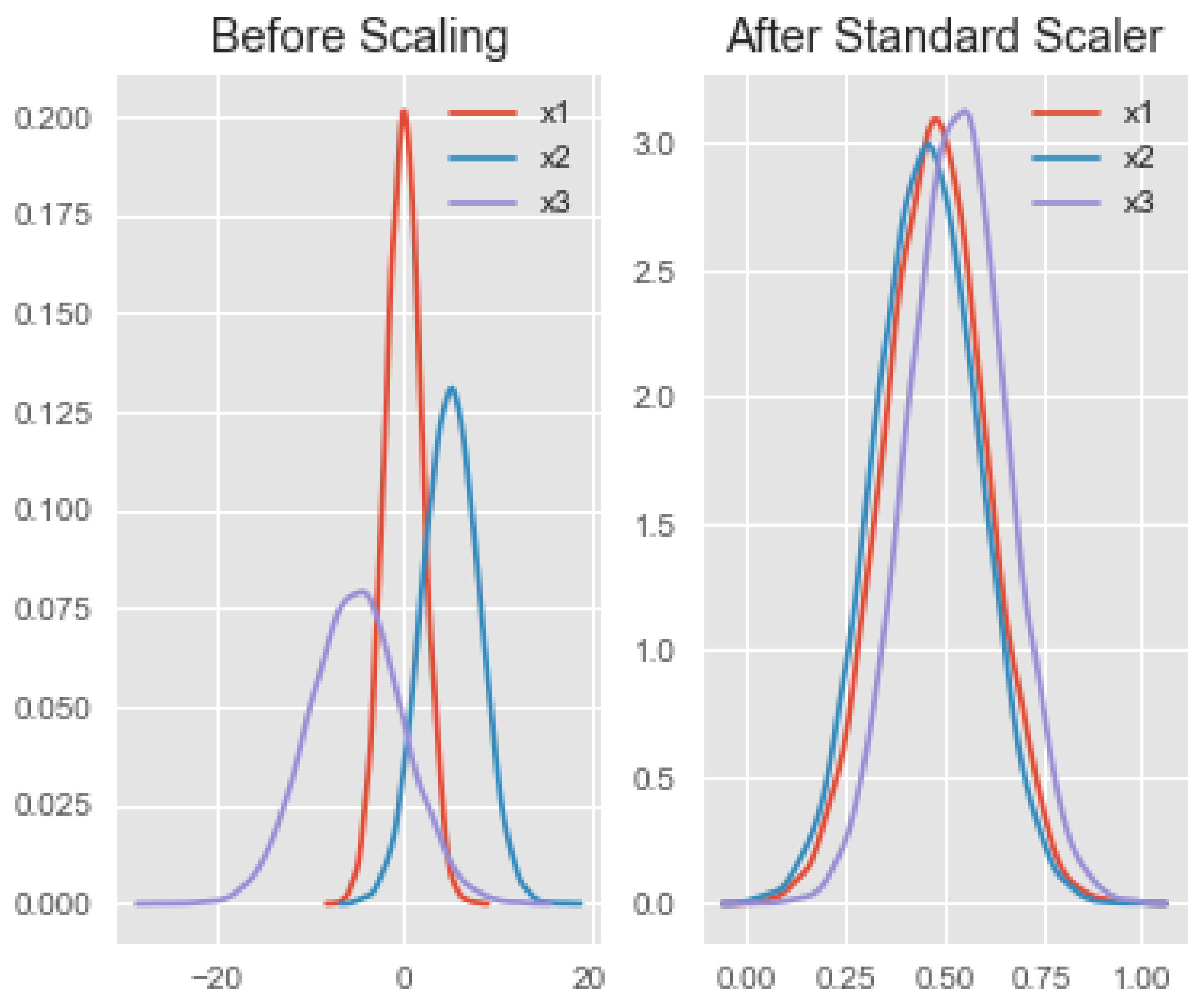
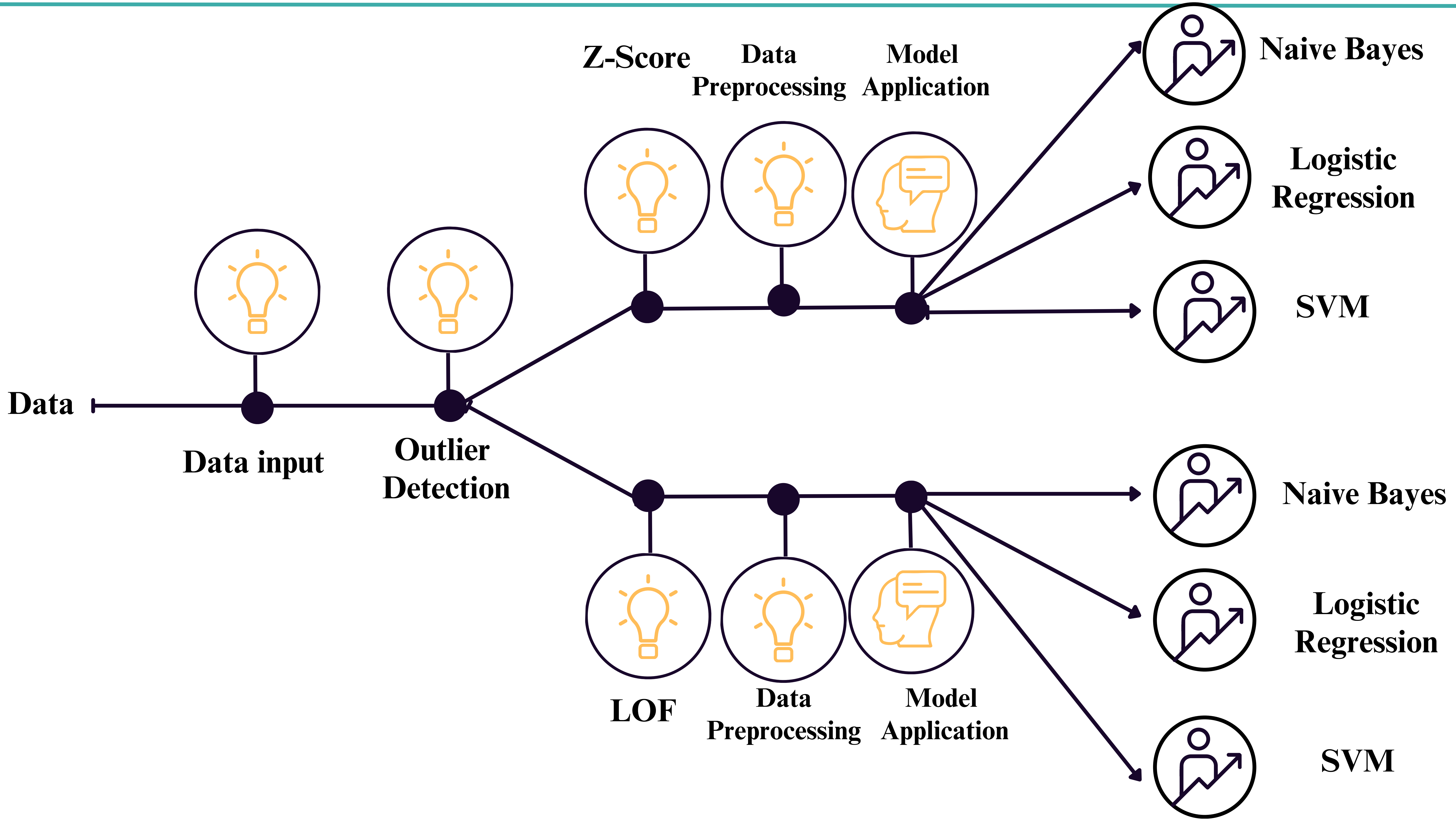


Image Source: https://stackoverflow.com/questions/39120942/difference-between-standardscaler-and-normalizer-in-sklearn-preprocessing

Flow Chart



Data
Data input
Outlier Detection
Z-Score
Data Preprocessing
Model Application
Naive Bayes
Logistic Regression
SVM
LOF
Data Preprocessing
Model Application
Naive Bayes
Logistic Regression
SVM

Models and Methodologies

Classification Problem

Naive Bayes Classifier

Logistic Regression Classifier

Support Vector Machine

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)

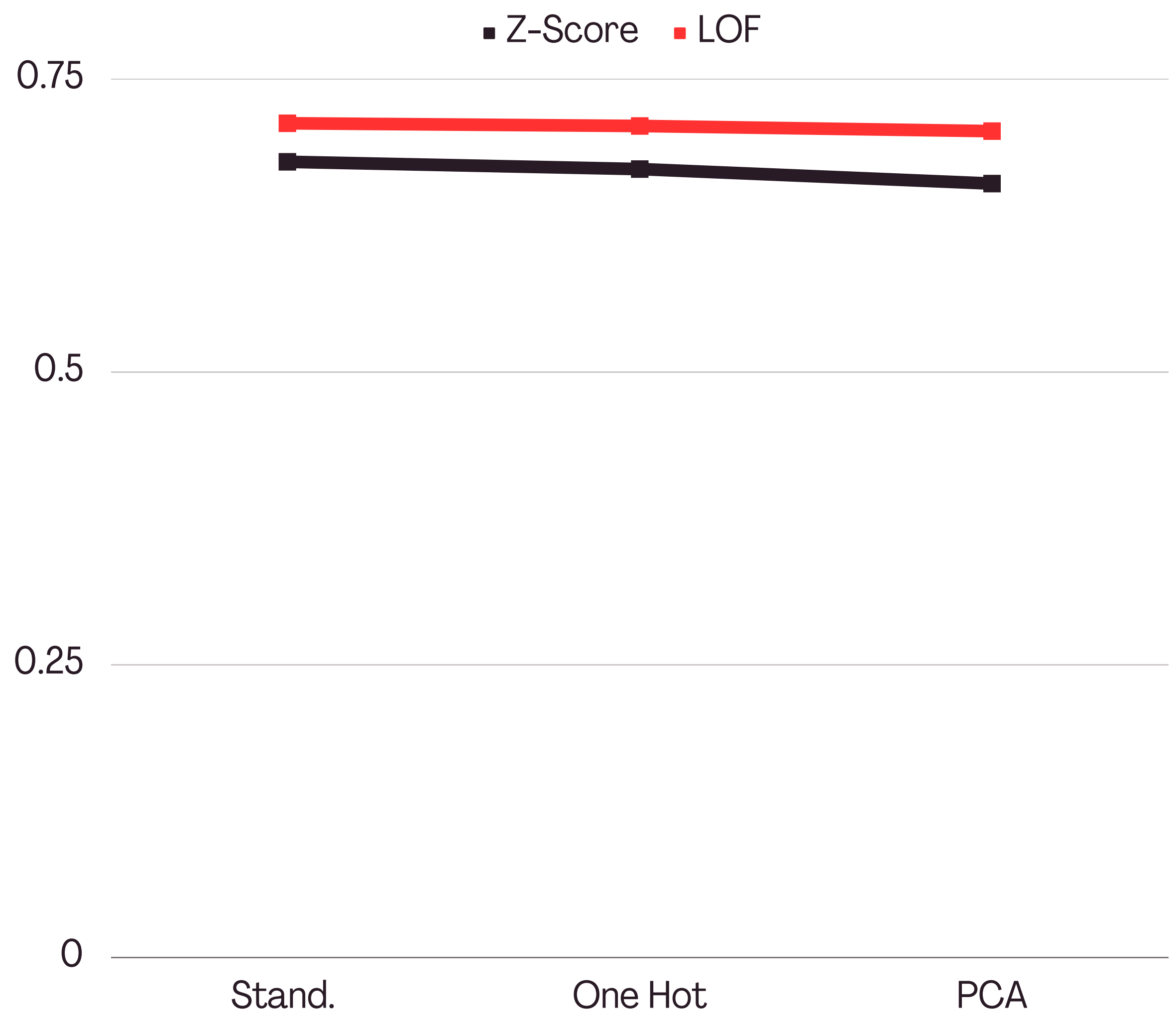
Naive Bayes

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)

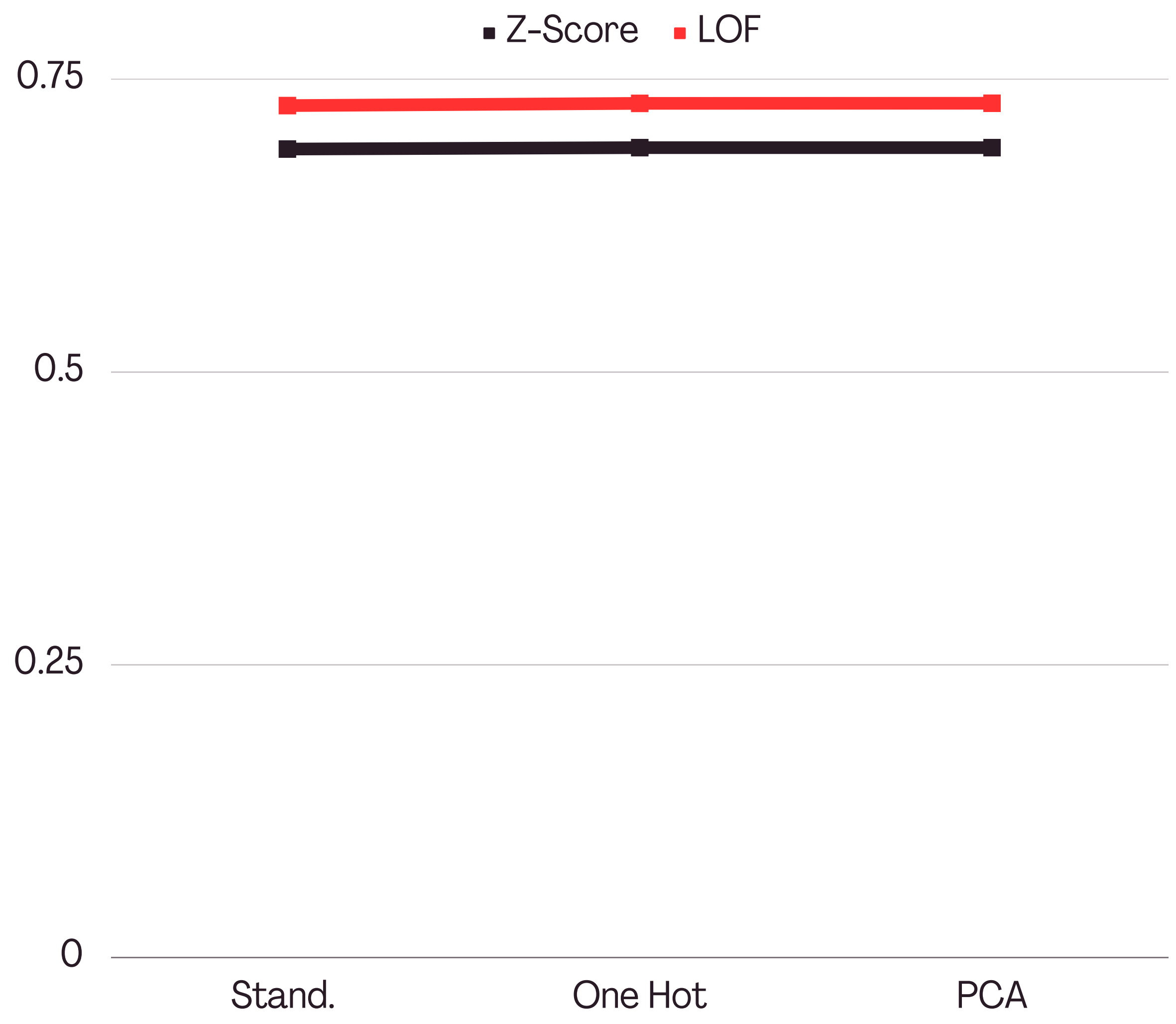
Logistic Regression

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

Models (Stats)

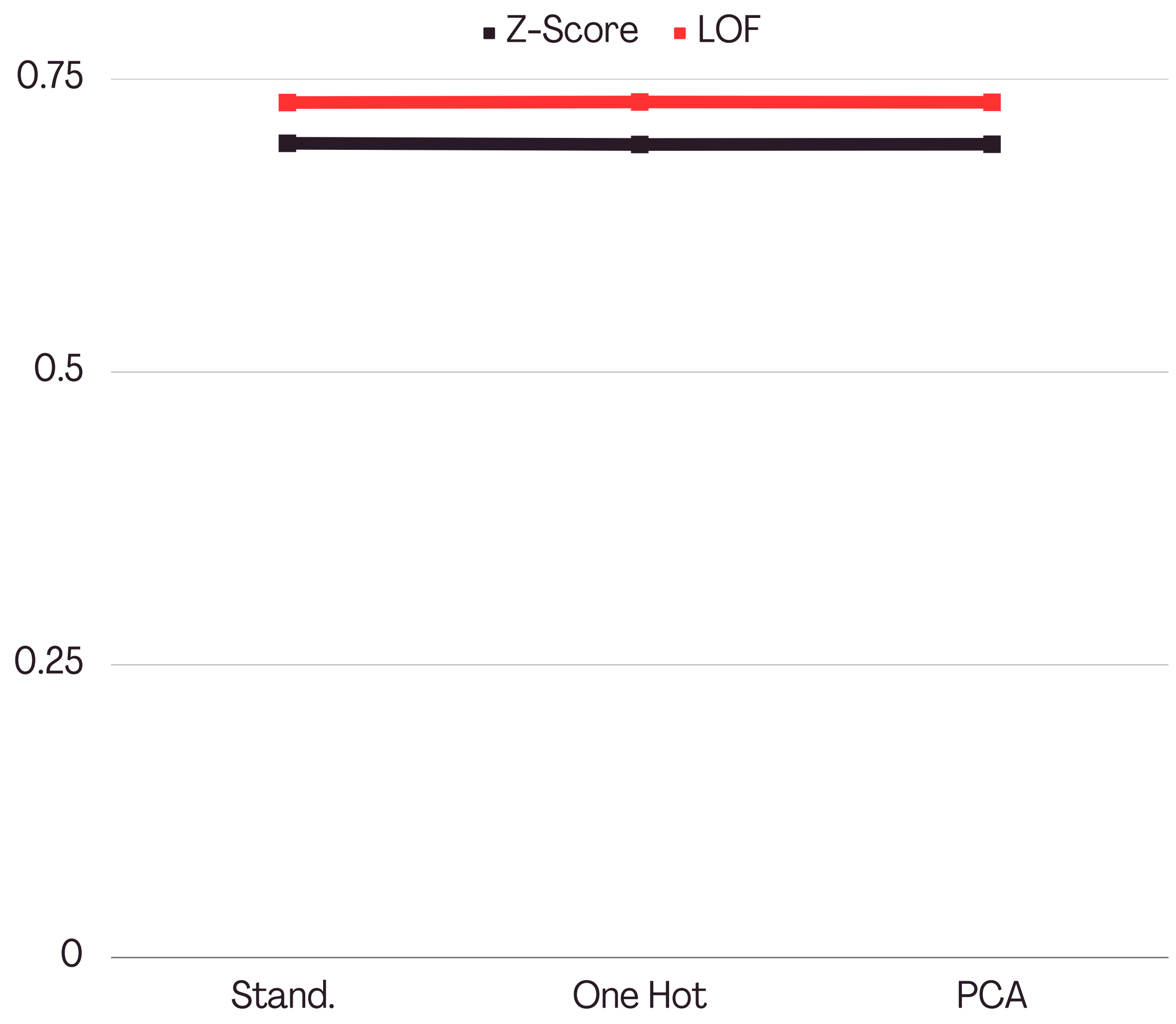
Support Vector Machine

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



F1-Score

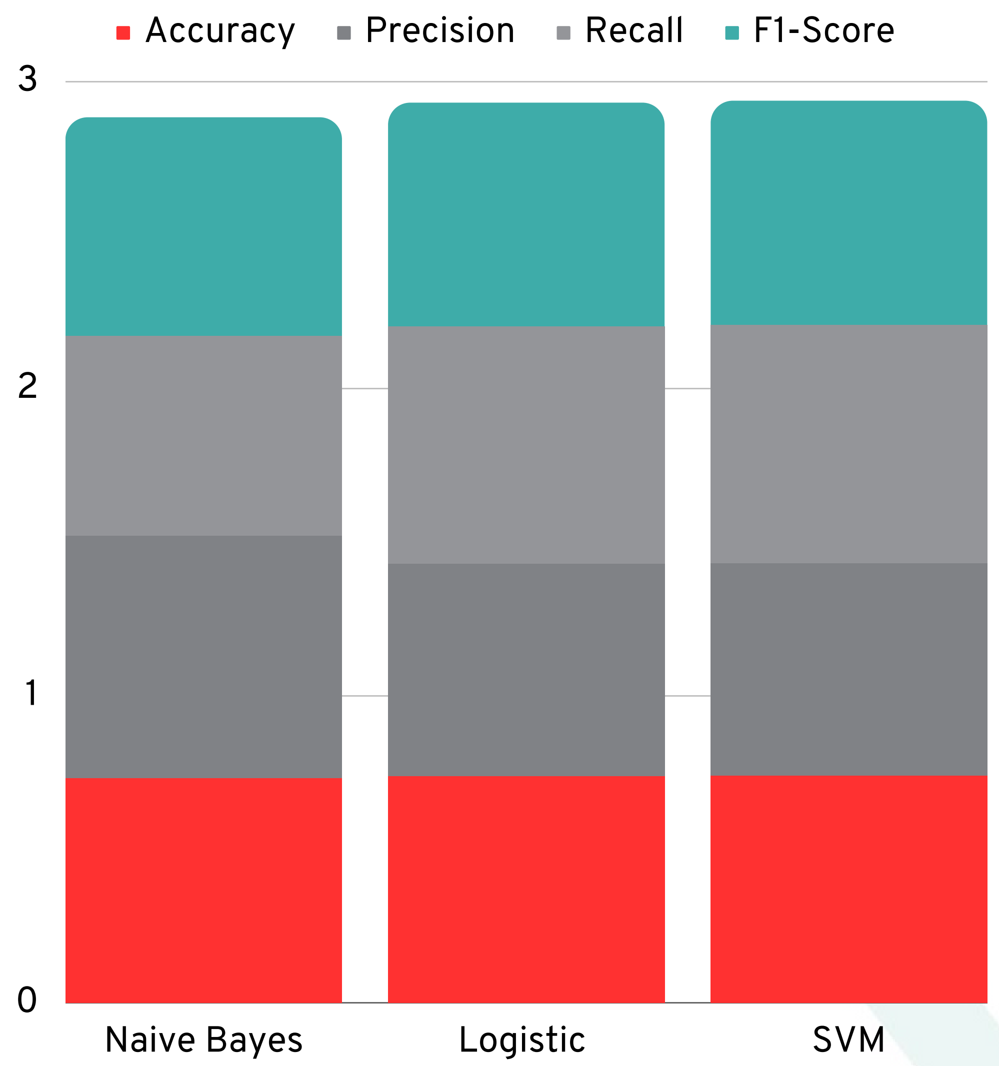
Models (Stats)

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
	Standard Data	0.7209	0.7191	0.7570	0.7592	0.6159	0.6148	0.6792	0.6794
Gaussian Naive Bayes	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
	Standard Data	0.7244	0.7206	0.6480	0.6433	0.7444	0.7448	0.6928	0.6903
Logistic Regression	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
	Standard Data	0.7312	0.7281	0.6391	0.6338	0.7620	0.7642	0.6952	0.6929
Support Vector Machines	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

Table 2. Metrics on the dataset cleaned using 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
	Standard Data	0.7275	0.7313	0.7763	0.7889	0.6403	0.6493	0.7018	0.7123
Gaussian Naive Bayes	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
	Standard Data	0.7311	0.7350	0.6810	0.6902	0.7575	0.7688	0.7172	0.7274
Logistic Regression	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
	Standard Data	0.7370	0.7388	0.6821	0.6892	0.7669	0.7760	0.7220	0.7300
Support Vector Machines	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

Table 3. Metrics on the dataset cleaned using Local Outlier Factor (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, higher recall is preferred.
- Best Recall observed in case of Support Vector Machines **(0.7754)** on one-hot encoded data with PCA applied, cleaned using LOF.
- Unequal weights are given to misclassification. Thus, the F1 Score is a better metric as compared to Accuracy.
- F1 Score for all models trained on data cleaned using LOF, better than the data cleaned using Z-Score.
- Best performance was achieved by Support Vector Machines (SVM) with an F1 score of **0.7305** on the testing data after one hot encoding was applied on the dataset. This was closely followed by Logistic Regression, which was able to achieve an F1 score o**f 0.7293** on the testing data

Conclusion

- From the above analysis, it can be concluded that out of Naive Bayes, Logistic Regression, and Support Vector Machines, the best performance (considering the F1 score as the overall evaluation metric) was achieved by Support Vector Machines (0.7305) followed by Logistic Regression (0.7294) and Naive Bayes (0.7274). This makes SVM the best model if only model performance is the concern.

- If training times are also considered, then logistic regression is the best model since it offers an almost comparable F1 score as SVM but has much lower training times.

- Further improvement of the F1 score can be achieved by using other classification models like Decision Trees and Random Forests. The existing models can also further be improved by hyper-parameter tuning via K-Fold Cross Validation, which will be taken up after the midsem evaluation

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
K-Modes + 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
- Slides

Karan Gupta

2021258

Contributions:
- Data Cleaning
- Multivariate Exploratory Data Analysis
- Naive Bayes
- Report Writing
- Slides.

Shivesh Gulati

2021286

Contributions:
- Literature Review (RP-2)
- Univariate Exploratory Data Analysis
- Logistic Regression
- Report Writing
- Slides.

Vishal Singh

2021575

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