Problem:

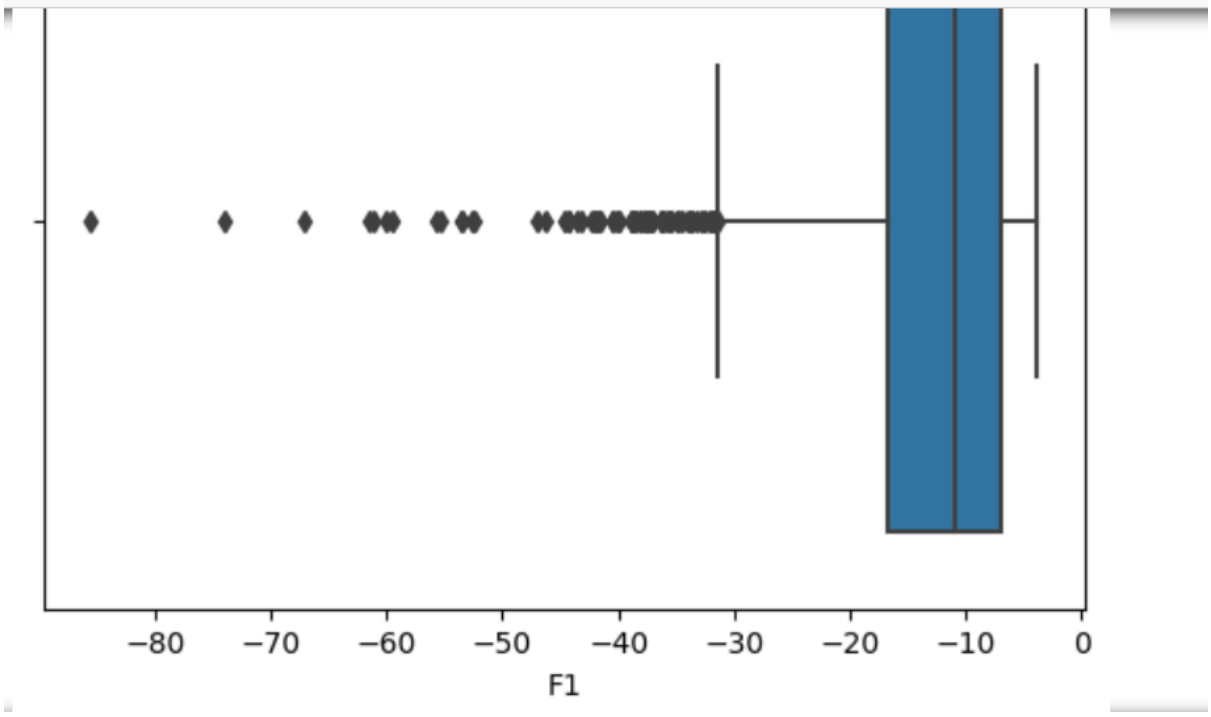
A healthcare provider wants to implement machine learning to identify patients who are at high risk of developing diabetes. The goal is to intervene proactively with preventive measures such as lifestyle interventions, medication, or referral to a specialist. This is a classification problem where the model must predict whether a patient is at high risk or not based on their medical history, lifestyle factors, and demographic information.

Data preparation:

Two CSV files are provided. first file is for training and testing. Second file is for predicting classification. For understanding of data, we generate descriptive statistic summary. Dataset has 20 features.

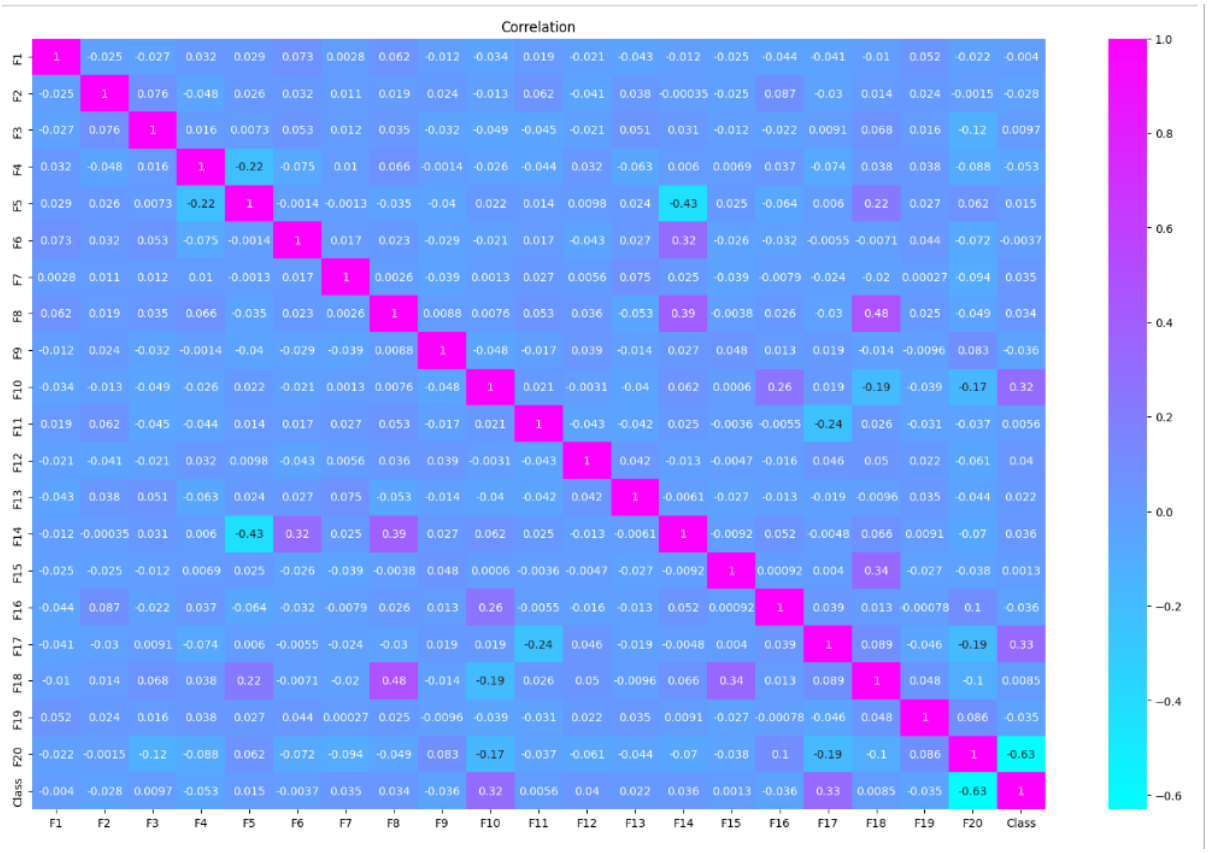
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	
mean	-13.932299	0.508000	13.207937	6.097217	1787.391510	5.310958	4.220605	-5066.909169	0.516000	9994.592611	3.255798	-8
std	10.285697	0.500186	1.531433	1.802407	702.890861	0.902777	0.606601	1023.164795	0.499994	1058.036210	2.699756	
min	-85.560000	0.000000	12.160005	4.225980	-1942.650000	4.322120	3.460056	-12915.220000	0.000000	3700.240000	0.334140	-24
25%	-16.732500	0.000000	12.257975	4.743650	1506.272500	4.617450	3.736800	-5335.020000	0.000000	9584.540000	1.150725	-8
50%	-10.862000	1.000000	12.596800	5.534500	1649.040000	5.032250	4.082350	-5044.997000	1.000000	9755.930000	2.384400	-8
75%	-6.972000	1.000000	13.587350	6.834000	1880.725000	5.723500	4.555750	-4795.820000	1.000000	10066.240000	4.704750	-8
max	-3.782328	1.000000	24.318000	13.330000	6602.350000	8.808000	7.254000	1892.780000	1.000000	21124.240000	13.536000	-7

Few necessary steps for data protection are Separating Training and testing labels, encoding categorial variables, detecting outliers, detecting outliers and imputing null values, etc.



Box Plot detecting outliers.

Generating correlation heatmap to examine correlation between the features.



(Correlation Heatmap)

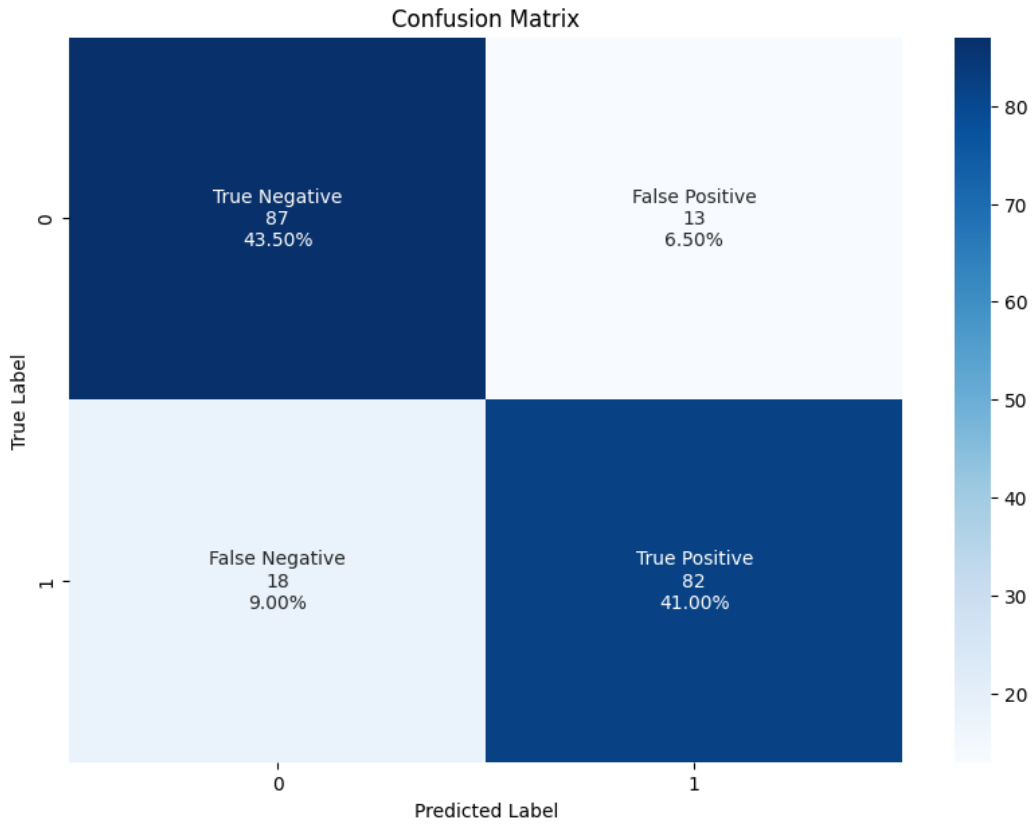
Building various ML models:

1. **Decision tree classifier:** Decision tree classifier is simplest and most effective model to binary classification problem. For DTC criterion had set to 'gini' as we all know tree will partition the data recursively using greedy algorithm. Decision trees also help to handle missing values and are useful for feature engineering.

```
Accuracy score on testing set 0.845
f1 score 0.8410256410256411
recall score 0.82
precision_score 0.8631578947368421
```

Results obtained from the Decision Tree Classifier

Confusion Matrix:

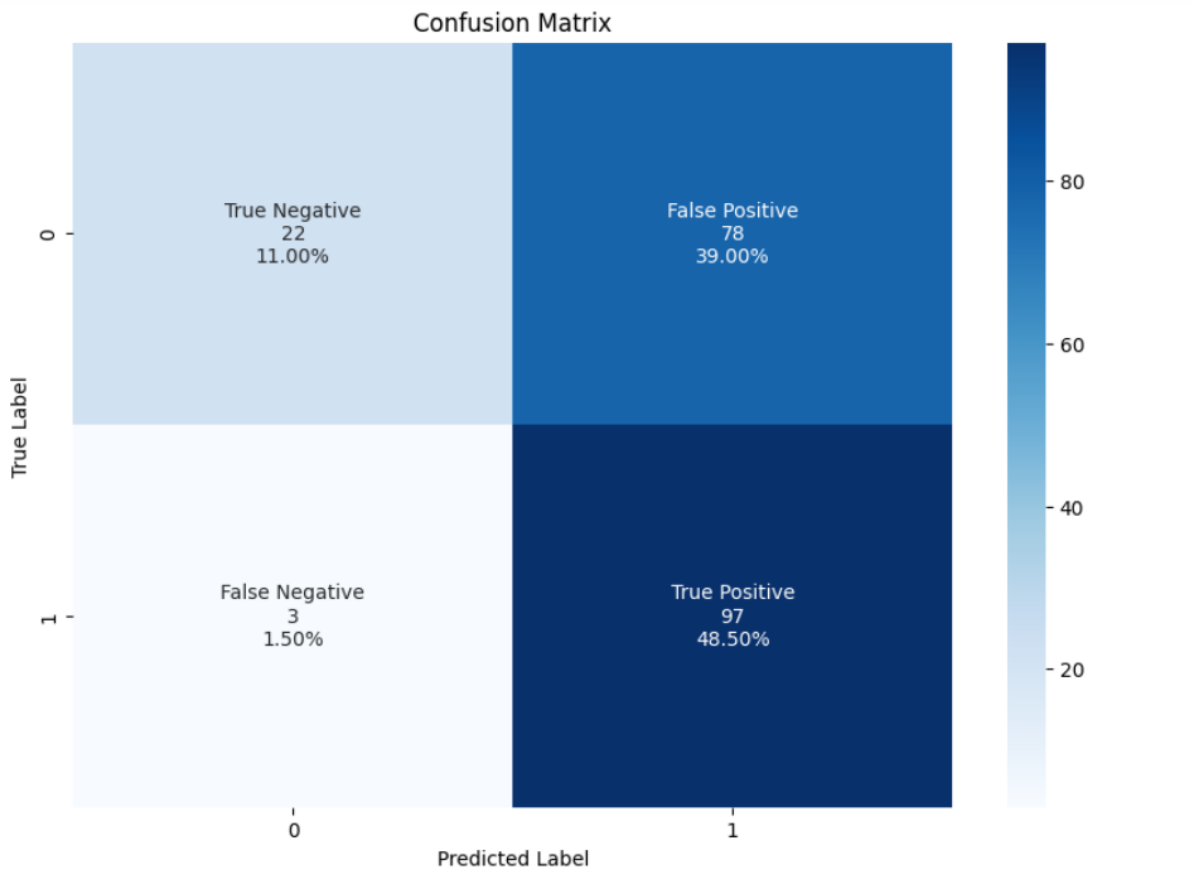


2. **Naive Bays Classifier:** Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts based on the probability of an object.

```
Accuracy score on testing set 0.595
f1 score 0.7054545454545454
recall score 0.97
precision_score 0.5542857142857143
```

Results obtained from the Naïve Bays Classifier.

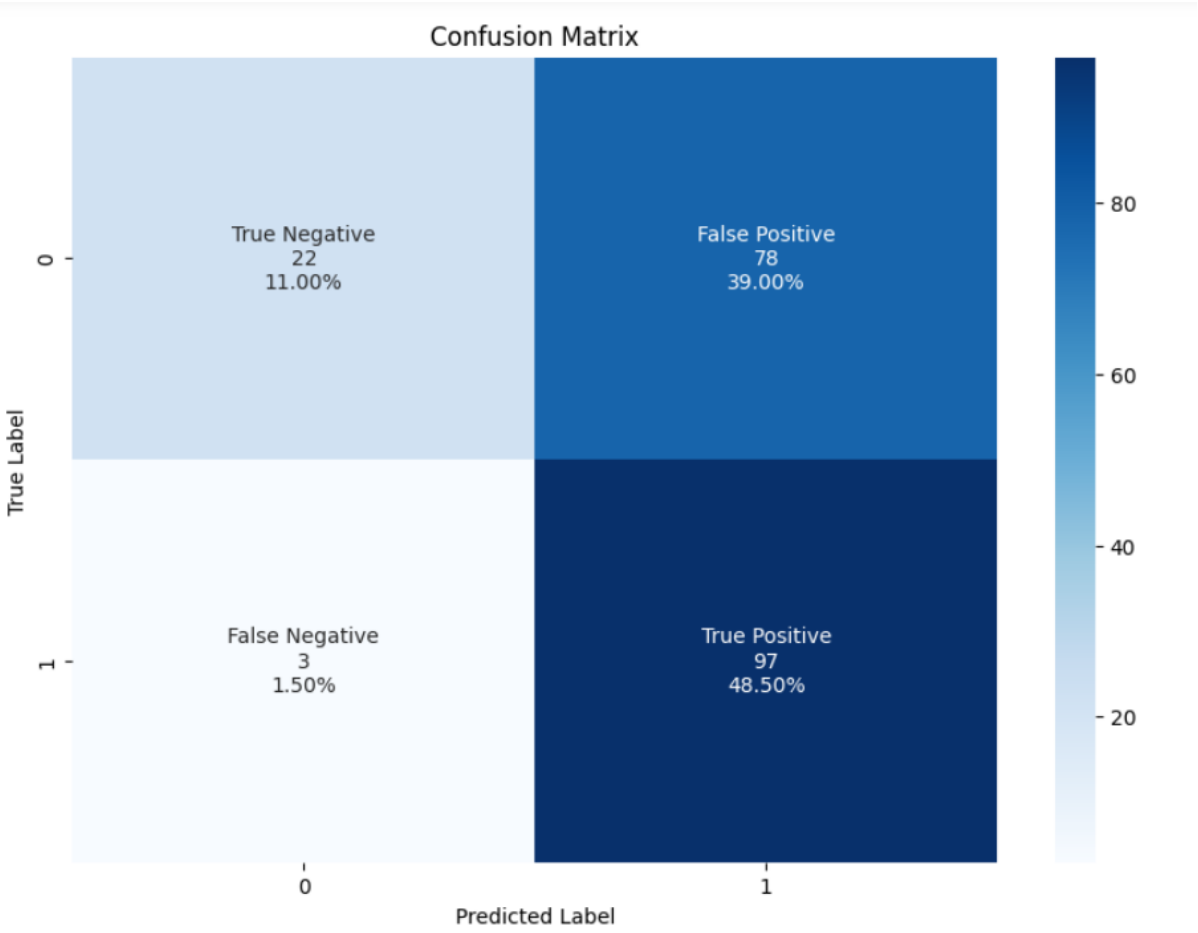
Confusion matrix of Naïve Bays Classifier



3. **Support Vector Machine (SVMs):** SVMs are majorly used for classification problem. In SVMs each data gets plot on n-dimensional space with value of each feature then hyper plane classifies the data.

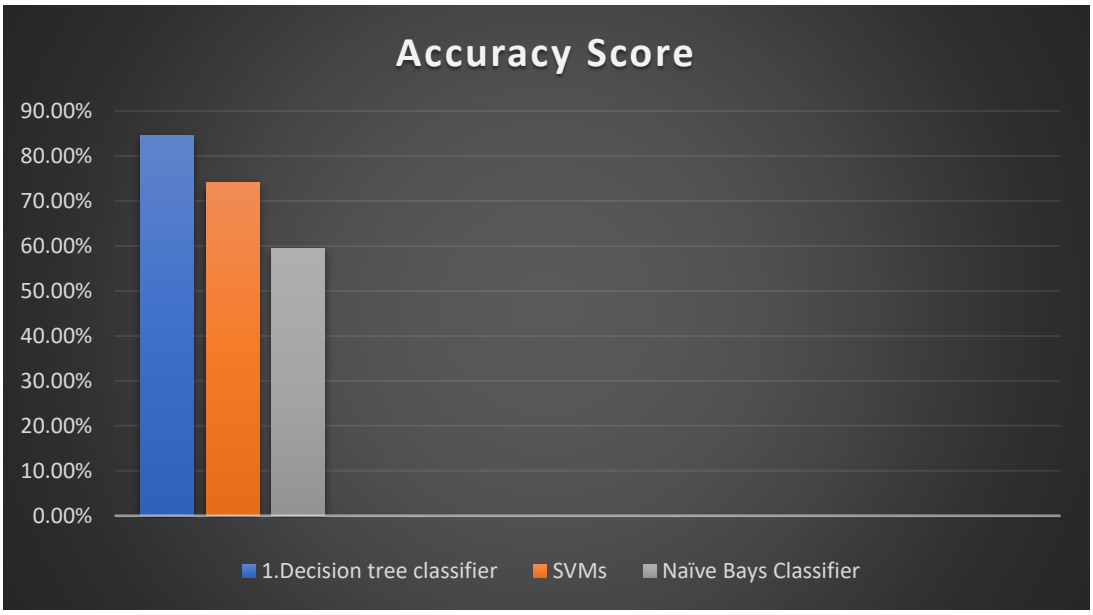
```
Accuracy score on testing set 0.74
f1 score 0.7425742574257426
recall score 0.75
precision_score 0.7352941176470589
```

Results obtained from the (SVMs)



Confusion matrix of SVMs

Evaluation: We had built three different classifier Machine learning model and same training and testing data used in all three classification models in which **Decision tree classifier.** performed better than Naïve Bays classifier and SVMs. **Decision tree classifier** model was the



robust and stable model for this classification problem.

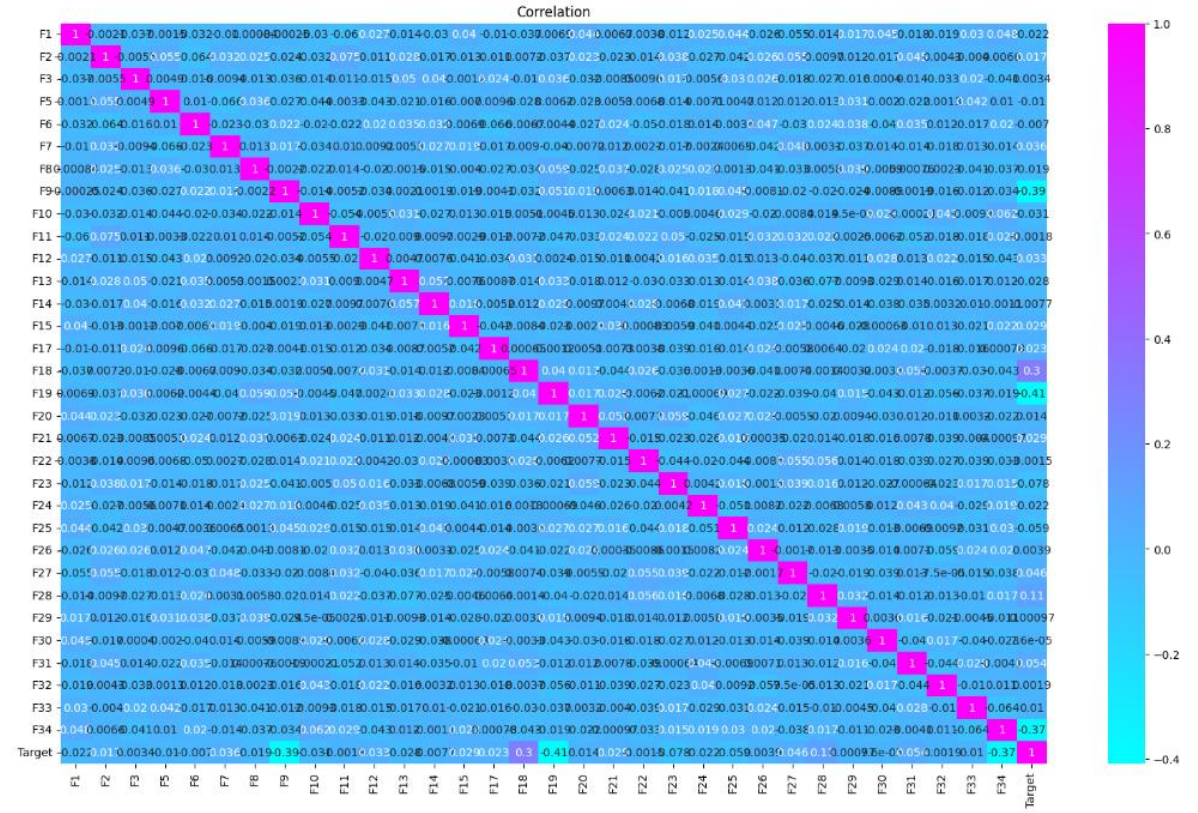
Regression Study

Data preparation: Two CSV files are provided. The first file is for training and testing. Second file is for predicting the value of the claim. For understanding of data, we generate descriptive. Statistic summary. The database has 34 features.

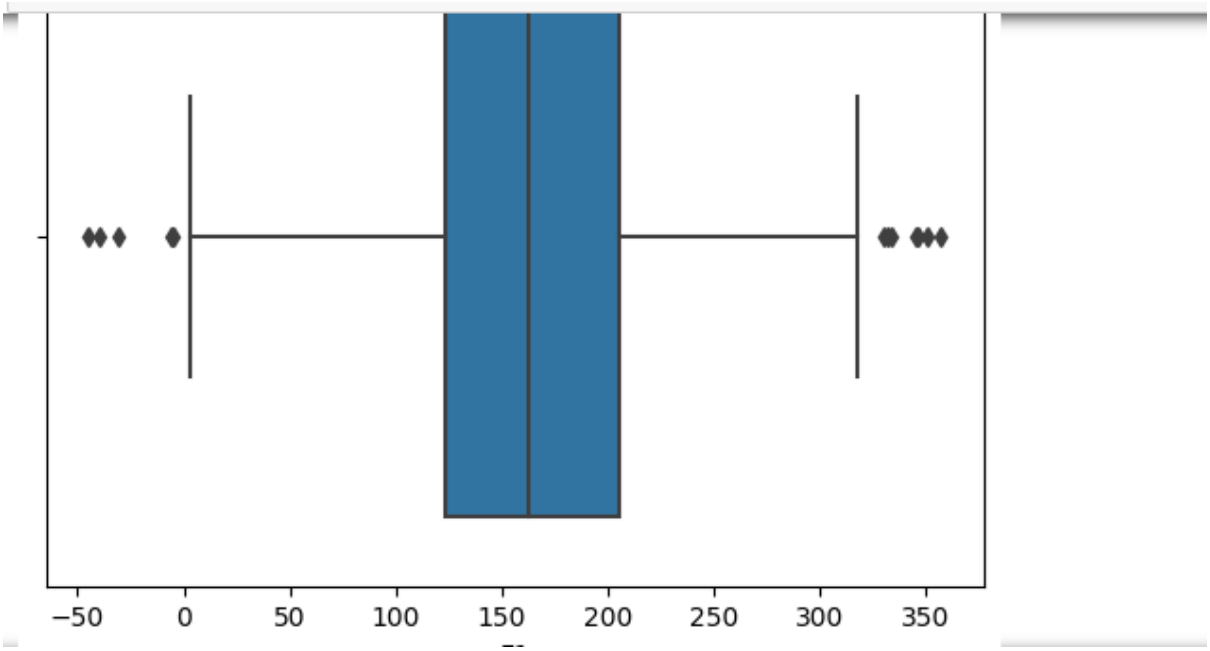
	F1	F2	F3	F5	F6	F7	F8	F9	F10	F11	...	F26	1
count	1400.000000	1400.000000	1400.000000	1400.000000	1400.000000	1400.000000	1400.000000	1400.000000	1400.000000	1400.000000	...	1400.000000	
mean	163.074257	2682.679800	193.647664	-10.325621	-680.463657	15.762943	52.998529	-9.783136	26.317457	12.655914	...	1660.091486	
std	61.182816	901.626953	44.715871	44.036834	583.314937	4.115171	8.279899	2.940784	7.943938	59.844389	...	1241.035124	
min	-44.680000	-185.460000	-47.690000	-239.210000	-2612.640000	-0.280000	14.200000	-18.940000	-4.040000	-198.780000	...	-4249.860000	
25%	123.125000	2062.095000	171.822500	-30.915000	-1079.270000	13.857500	48.900000	-11.710000	22.220000	-25.805000	...	1030.050000	
50%	162.750000	2716.335000	192.990000	-10.405000	-687.190000	15.770000	53.000000	-9.925000	26.420000	13.320000	...	1673.250000	
75%	205.400000	3294.570000	214.340000	11.862500	-288.680000	17.902500	57.105000	-7.840000	30.660000	52.700000	...	2279.730000	
max	357.500000	5592.120000	435.650000	182.080000	1290.660000	43.460000	99.440000	-0.080000	64.700000	216.960000	...	7878.570000	

8 rows × 33 columns

Generating correlation heatmap to examine correlation between the features.



Few necessary steps for data protection are Separating Training and testing labels, encoding categorial variables, detecting outliers, detecting outliers and imputing null values, etc.



Box plot detecting outliers.

Building various ML models:

1. **Linear Regression:** Linear Regression is a statistical method to find the relationship between dependent and independent variables. Linear regression statistic model built using klearn module without parameter.

```
Accuracy score on testing set 0.5857535670657383
```

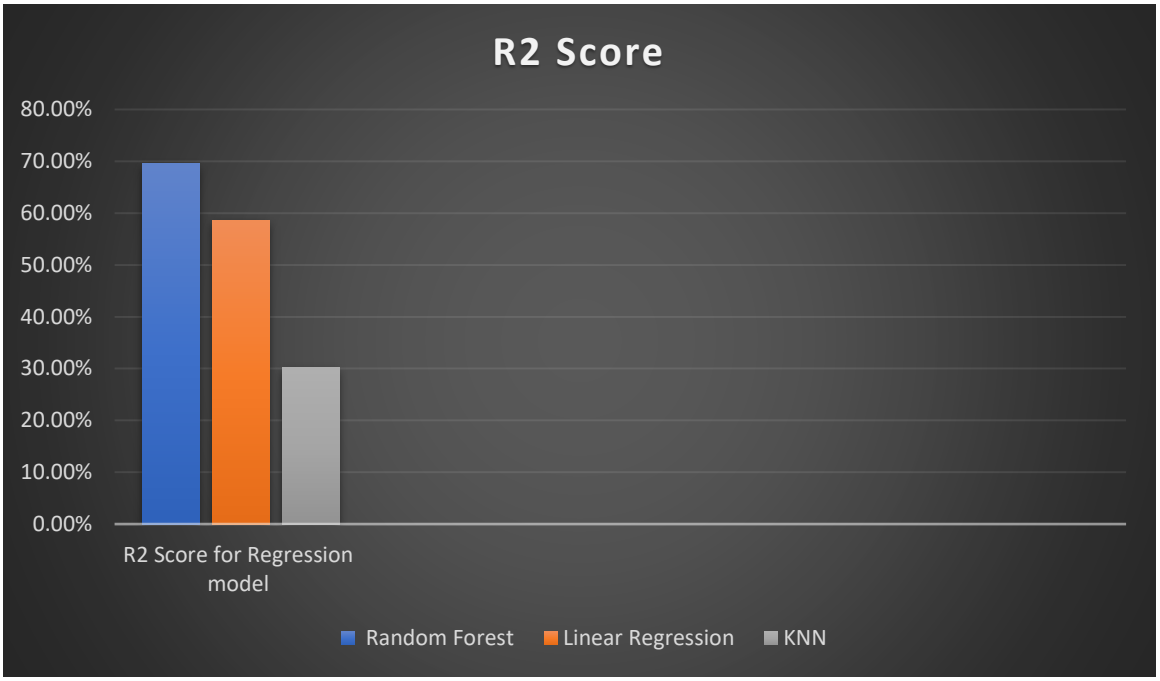
2. **Random Forest Regressor**: Random Forrest is an ensemble machine learning technique of decision tree. Parameters of random forest are max_depth and n_estimators are set to 10 and 100 respectively.

```
R2 Score score on testing set 0.695378020478104
RMSE 119.67481457304474
```

3. **K Nearest Neighbor:** The k-nearest neighbors' algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

```
RMSE 181.23684850003625
R2 Score score on testing set 0.30136761591365435
```

Evaluation: I had built three different regression Machine learning model and same training and testing data used in all three regression models in which Random Forest Regressor performed better than K Nearest Neighbor and linear regression, Random Forest model was the robust and has better R2 value among all models for this regression problem.



Summary of Experiments

I conducted two experiments in this investigation:

Classification study: In this study, we used the given dataset to identify patients who are at high risk of developing diabetes using machine learning models. We used Decision Tree Classifier, Naïve Bayes Classifier, and Support Vector Machines (SVMs) for classification. After data preparation, we built three different machine learning models and tested them using the same training and testing data. Among these models, Decision Tree Classifier performed better with an accuracy score of 0.835, f1 score of 0.829, recall score of 0.8, and precision score of 0.86. Naïve Bayes Classifier had the lowest accuracy score of 0.595, f1 score of 0.705, recall score of 0.97, and precision score of 0.554. SVMs had an accuracy score of 0.74, f1 score of 0.743, recall score of 0.75, and precision score of 0.735. We concluded that Decision Tree Classifier is the most suitable model for predicting diabetes risk.

Regression study: In this study, we used the given dataset to predict the insurance claim amount using machine learning models. We used Linear Regression, Random Forest Regressor, and K-Nearest Neighbor (KNN) for regression. After data preparation, we built three different machine learning models and tested them using the same training and testing data. Among these models, Random Forest Regressor performed better with an R2 score of 0.695 and RMSE of 119.67. KNN had the highest RMSE of 181.24 and an R2 score of 0.301. Linear Regression had an accuracy score of 0.586. We concluded that Random Forest Regressor is the most suitable model for predicting insurance claim amount.

Performance of different solutions:

Comparison of classification models

Model Name	Accuracy Score	F1 Score	Recall Score	Precision Score
Decision Tree Classifier	0.835	0.829	0.8	0.86
Naïve Bayes Classifier	0.595	0.705	0.97	0.554
Support Vector Machines	0.74	0.743	0.75	0.735

Comparison of regression models

Model Name	R2 Score	RMSE
Linear Regression	0.586	N/A
Random Forest Regressor	0.695	119.67
K-Nearest Neighbour	0.301	181.24

Interpretation of results:

In the classification study, we found that Decision Tree Classifier outperformed Naïve Bayes Classifier and SVMs in predicting diabetes risk. The decision tree model had the highest accuracy score, f1 score, and precision score, which suggests that it correctly predicted high-risk patients and minimized false positives. The Naïve Bayes Classifier model had a high recall score, which suggests that it.