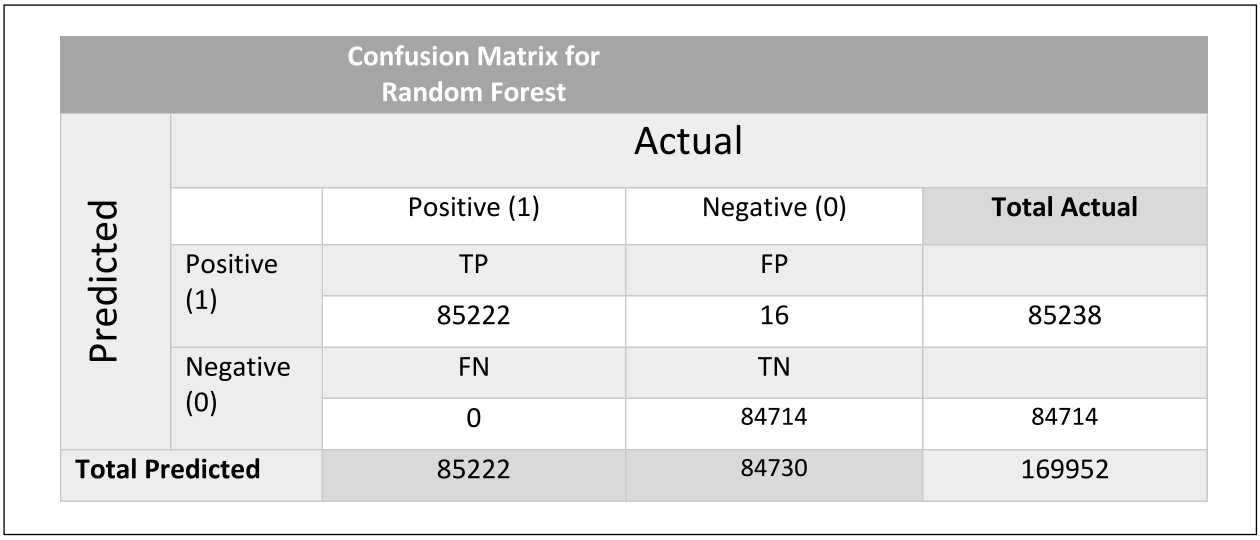
1. Numerical Results

In this paper, we have applied six supervised machine learning techniques termed as Random Forest, XGBoost, Decision Tree, K-nearest Neighbours, Naïve Bayes approach, and Logistic Regression for the Credit Card Fraud detection dataset. We have performed these algorithms on a dataset comprising of European Cardholders in which the total number of samples is **284807** and the total number of attributes is **31**. Last column represents a class label where positive class represents Fraudulent transaction and negative class represents Legit/Non-fraudulent transaction. In our experiments, **397216** samples are chosen as training data and remaining **170236** samples for test data. Here, True positive (TP) represents number of samples with the Fraudulent transactions predicted as those having Fraudulent transactions, False positive (FP) represents number of samples with the Legit Transactions predicted as those having Fraudulent transactions. True negative (TN) represents number of samples with the Legit Transactions predicted as those having Legit Transactions, and False negative (FN) represents number of samples with the Fraudulent transactions predicted as those having Legit transactions. Here, we have used the following quality measures to check the performance of machine learning techniques:

- Accuracy = (TP + TN)/(TP + TN + FP + FN)
- Recall (Sensitivity or true positive rate) = TP/(TP + FN)
- Specificity (True negative rate) = TN/(TN + FP)
- Precision = TP/(TP + FP)
- Negative predicted value (NPV) = TN/(TN + FN)
- False positive rate (FP rate) = FP/(FP + TN)
- Rate of misclassification (RMC) = (FP + FN)/(TP + TN + FP + FN)
- F1-measure = 2 * (precision * recall)/(precision + recall)
- G-mean = sqrt(precision*recall)

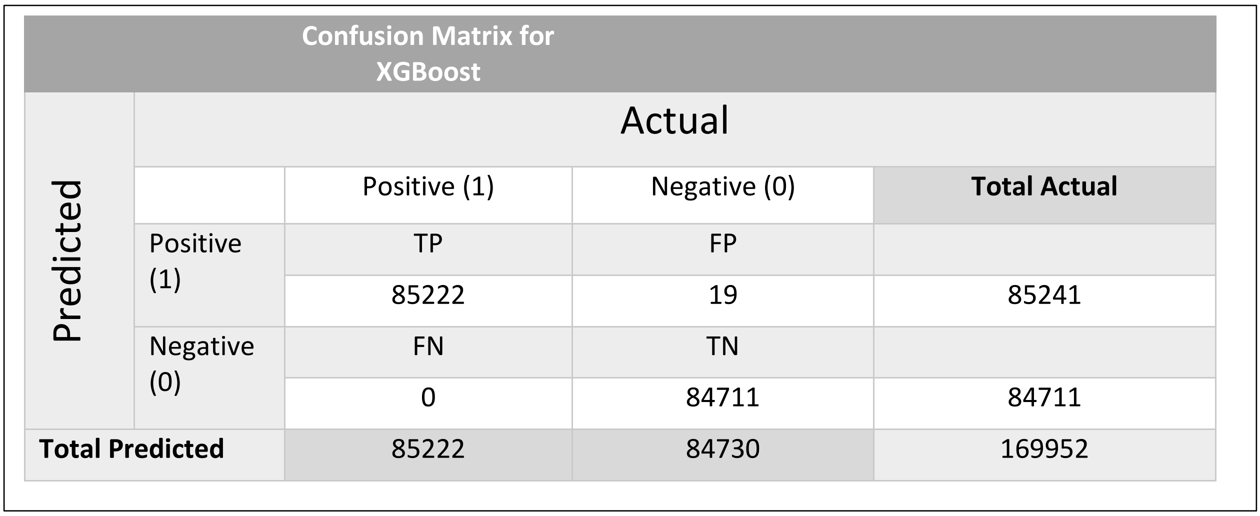
The confusion matrix of prediction results for Random Forest, XGBoost, Decision Tree, K-nearest Neighbours, Naïve Bayes approach, and Logistic Regression are tabulated in Tables 1, 2, 3, 4, 5 and 6. One can observe from these tables that Random Forest, XGBoost, Decision Tree, K-nearest Neighbours are having the highest number of true positive (with the Fraudulent transactions predicted as those having Fraudulent transactions) and Random Forest is having the highest number of true negatives (the Legit Transactions predicted as those having Legit Transactions). Further, Random Forest, XGBoost, Decision Tree, K-nearest Neighbours are having the lowest number of false negative and Random Forest is having the lowest number of false positive. We have also drawn the classification results of these methods in Fig. 1. We have computed the value of TP, FP, TN and FN for Random Forest, XGBoost, Decision Tree, K-nearest Neighbours, Naïve Bayes approach, and Logistic Regression and depicted in Table 6. Further, we have computed quality measures termed as accuracy, recall, true positive rate, precision, negative predicted value, false positive rate, rate of misclassification, F1-measure, G-mean and ROC_AUC_Score based on predicted result by using Random Forest, XGBoost, Decision Tree, K-nearest Neighbours, Naïve Bayes approach, and Logistic Regression and depicted in Table 7 and also shown in Fig. 2. Here, one can conclude from this Table 7 that Random Forest has performed better among all six algorithms, whereas Naïve Bayes is having the lowest accuracy. In terms of F1-measure, Random Forest performed better than other compared methods.

1. Random Forest



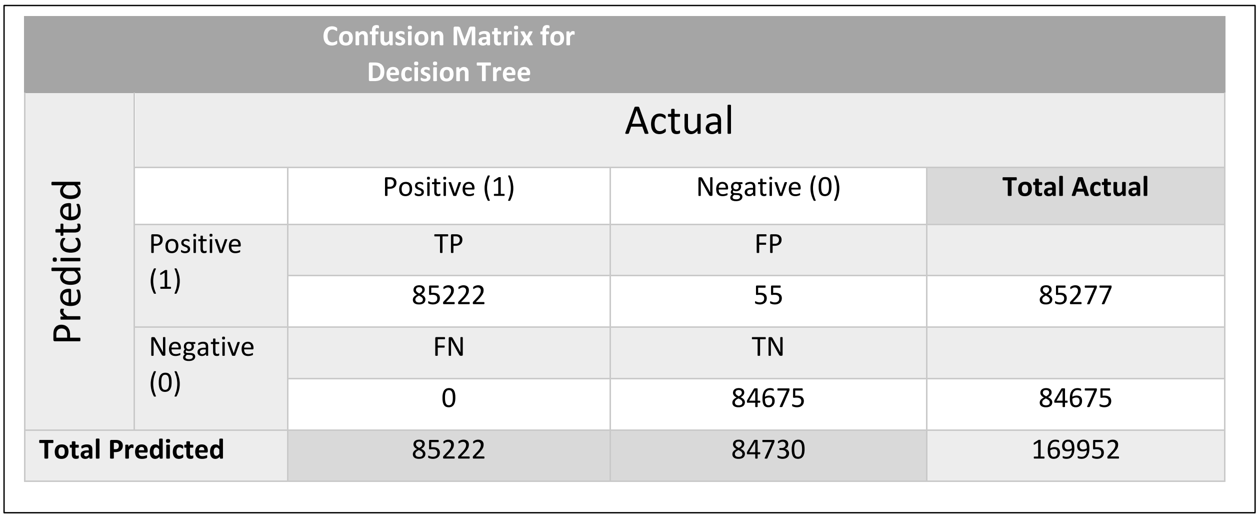
Confusion Matrix for Random Forest				
		Actual		
Predicted		Positive (1)	Negative (0)	**Total Actual**
	Positive (1)	TP	FP	
		85222	16	85238
	Negative (0)	FN	TN	
		0	84714	84714
Total Predicted		85222	84730	169952

2. XGBoost

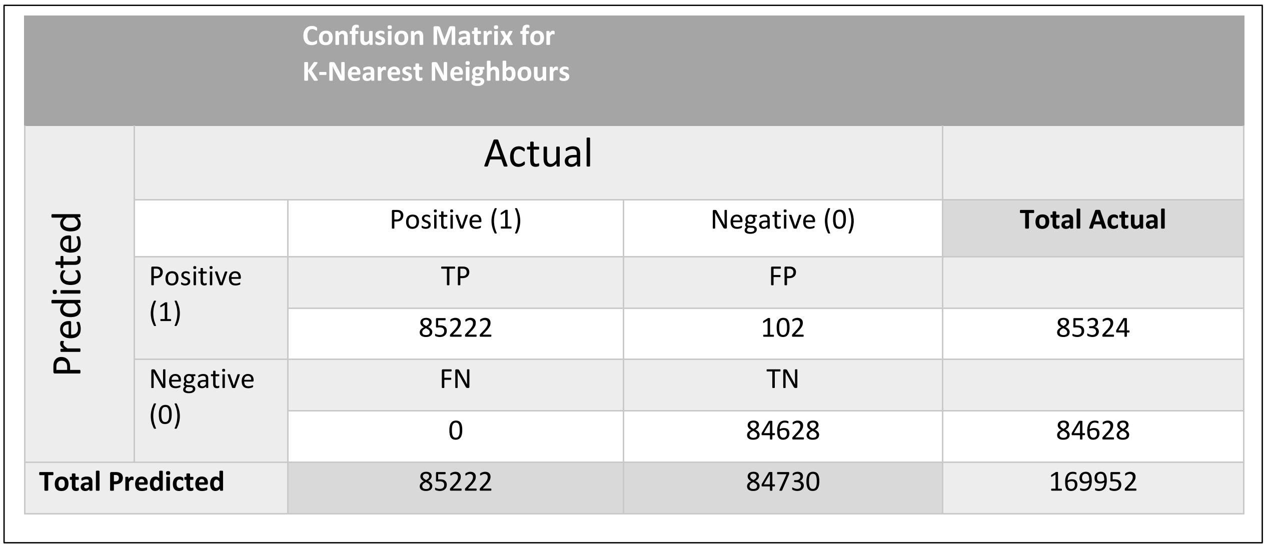


Confusion Matrix for XGBoost				
		Actual		
Predicted		Positive (1)	Negative (0)	**Total Actual**
	Positive (1)	TP	FP	
		85222	19	85241
	Negative (0)	FN	TN	
		0	84711	84711
Total Predicted		85222	84730	169952

3. Decision Tree

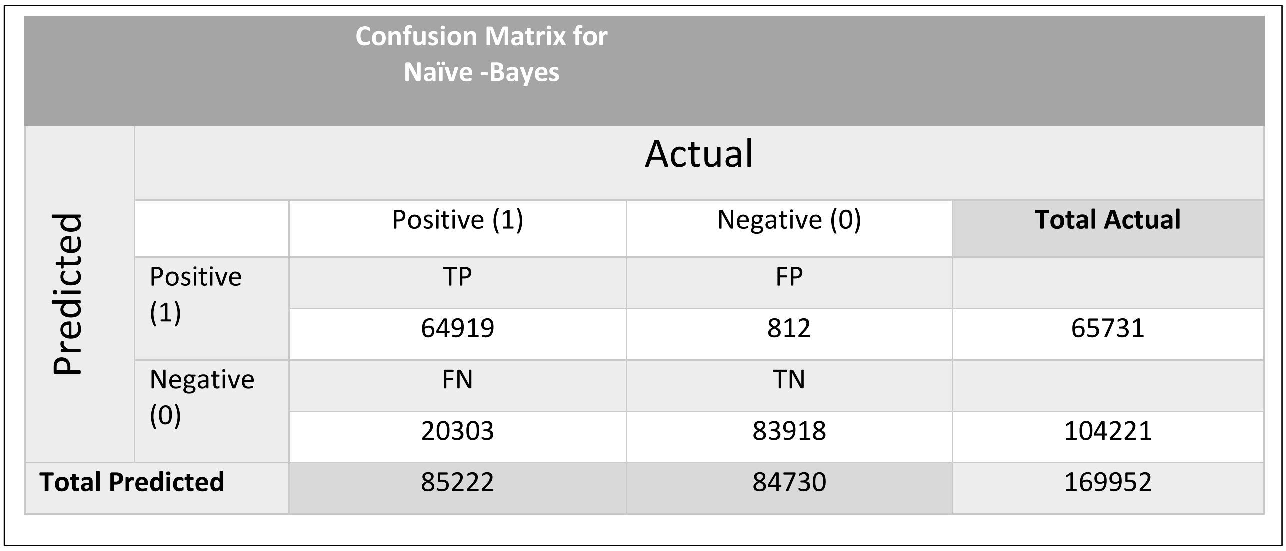


Confusion Matrix for Decision Tree				
		Actual		
Predicted		Positive (1)	Negative (0)	**Total Actual**
	Positive (1)	TP	FP	
		85222	55	85277
	Negative (0)	FN	TN	
		0	84675	84675
Total Predicted		85222	84730	169952



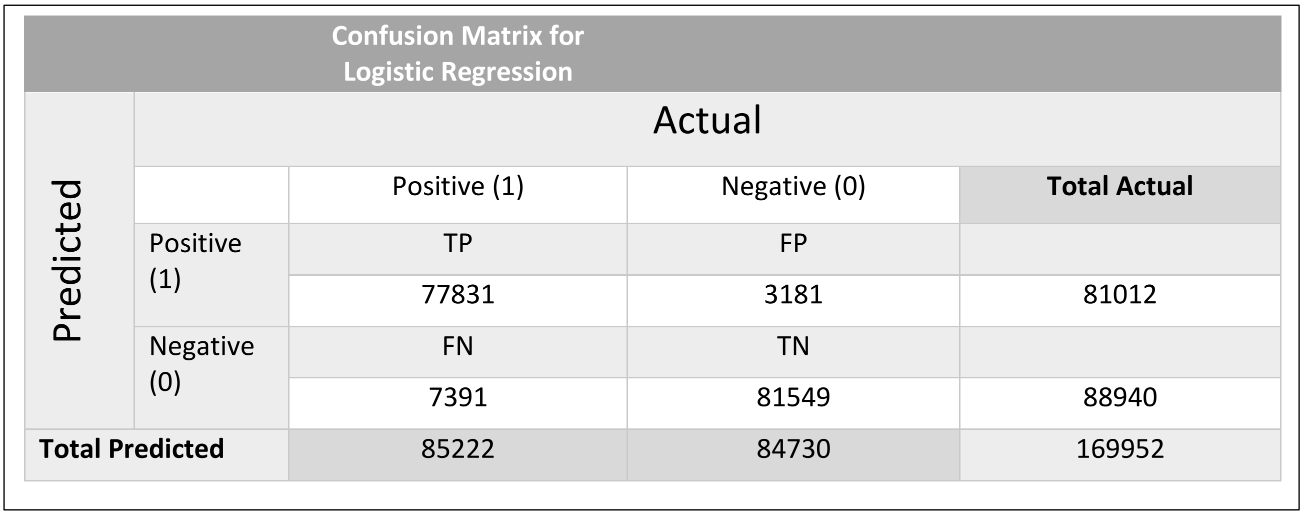
Confusion Matrix for K-Nearest Neighbours				
Predicted		Actual		
		Positive (1)	Negative (0)	**Total Actual**
	Positive (1)	TP	FP	
		85222	102	85324
	Negative (0)	FN	TN	
		0	84628	84628
Total Predicted		85222	84730	169952

5. Naïve -Bayes



Confusion Matrix for Naïve -Bayes				
Predicted		Actual		
		Positive (1)	Negative (0)	**Total Actual**
	Positive (1)	TP	FP	
		64919	812	65731
	Negative (0)	FN	TN	
		20303	83918	104221
Total Predicted		85222	84730	169952

6. Logistic Regression



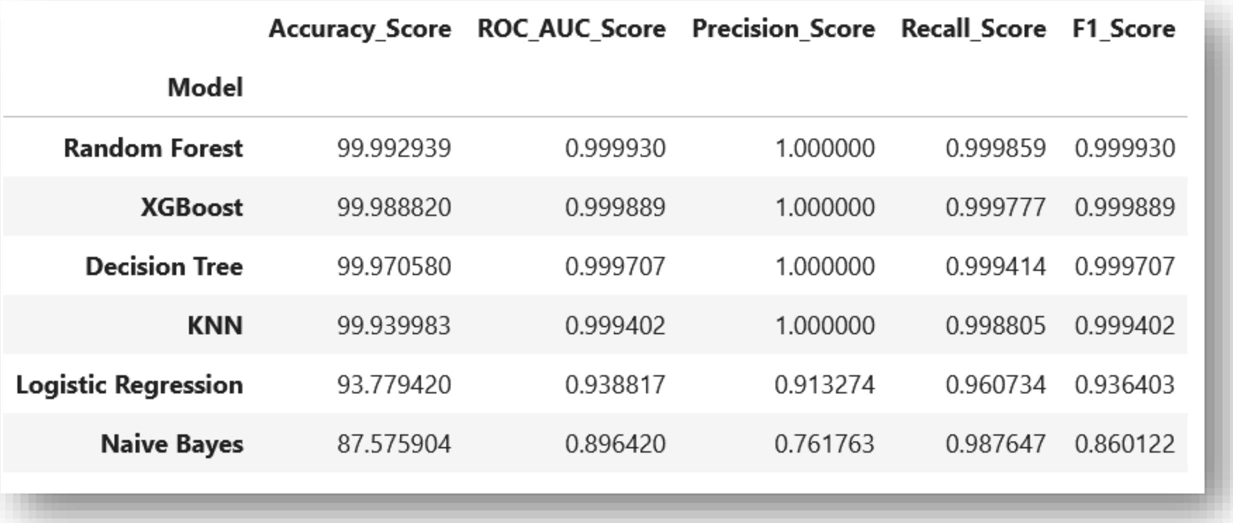
Confusion Matrix for Logistic Regression				
Predicted		Actual		
		Positive (1)	Negative (0)	**Total Actual**
	Positive (1)	TP	FP	
		77831	3181	81012
	Negative (0)	FN	TN	
		7391	81549	88940
Total Predicted		85222	84730	169952

Table 6- Values of TP, FP, TN, and FN for Random Forest, XGBoost, Decision Tree, K-nearest Neighbours, Naïve Bayes approach, and Logistic Regression

	Random Forest	XGBoost	Decision tree	KNN	Naïve Bayes	Logistic Regression
TP	85222	85222	85222	85222	64919	77831
FP	16	19	55	102	812	3181
TN	84714	84711	84675	84628	83918	81549
FN	0	0	0	0	20303	7391

Table 7- Classification performance measure indices of Random Forest, XGBoost, Decision Tree, K-nearest Neighbours, Naïve Bayes approach, and Logistic Regression

	Random Forest	XGBoost	Decision tree	KNN	Naïve Bayes	Logistic Regression
Accuracy	0.99990586	0.99988820	0.99967638	0.99939983	0.87575904	0.93779420
Recall (Sensitivity or true positive rate)	1.00000000	1.00000000	1.00000000	1.00000000	0.76176339	0.91327357
Specificity (True negative rate)	0.99981116	0.99977576	0.99935088	0.99879618	0.99041662	0.96245722
Precision	0.99981229	0.99977710	0.99935504	0.99880456	0.98764662	0.96073421
Negative predicted value (NPV)	1.00000000	1.00000000	1.00000000	1.00000000	0.80519281	0.91689903
False positive rate (FP rate)	0.00018884	0.00022424	0.00064912	0.00120382	0.00958338	0.03754278
Rate of misclassification (RMC)	0.00009414	0.00011180	0.00032362	0.00060017	0.12424096	0.06220580
F1-measure	0.99990614	0.99988854	0.99967742	0.99940192	0.86012202	0.93640290
G-mean	0.99990614	0.99988855	0.99967747	0.99940210	0.86738287	0.93670335
ROC_AUC_Score	0.99993	0.999889	0.999707	0.999402	0.89642	0.938817



	Accuracy_Score	ROC_AUC_Score	Precision_Score	Recall_Score	F1_Score
Model					
Random Forest	99.992939	0.999930	1.000000	0.999859	0.999930
XGBoost	99.988820	0.999889	1.000000	0.999777	0.999889
Decision Tree	99.970580	0.999707	1.000000	0.999414	0.999707
KNN	99.939983	0.999402	1.000000	0.998805	0.999402
Logistic Regression	93.779420	0.938817	0.913274	0.960734	0.936403
Naive Bayes	87.575904	0.896420	0.761763	0.987647	0.860122