CREDIT CARD FRAUD DETECTION SYSTEM: COMPARISON STUDYON VARIOUS MACHINE LEARNING AND DEEP LEARNING TECHNIQUE(S)

SOFT COMPUTING (ITE1015)
in
B.Tech (IT)
by

Devansh Chauhan(19BIT0153)
Yashkumar Patel(19BIT0183)
Sahil Saxena(19BIT0253)

5thSEMESTER , 2021

Under the Guidance of
Prof. L.AGILANDEESWARI
Associate Professor, SITE

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Abstract

Usage of liquid cash is getting neglected day-by-day because ofthe increased usage of credit card payments method. There are a few countries which are completely relying upon digital transactions, omitting paper currency. But that doesn't mean this kind of payment method are completely reliable. As the cybersecurity field is excelling at a rapid pace , accessing abankaccountis not a tough job to do today. Thus , Credit Card fraudcasesare also piling up in every developed and developing country.This project aims to compare all the trending efficient methodsagainst CreditCard Frauds, by training the model using dataset (from Kaggle). The model consists of 8 modules / techniquesnamely –CNN (Convolutional Neural Network) , Decision Tree Algorithm , K-NN (K –Nearest Neighbour) , SVM (Support Vector Machine) , NaïveBayes Classifier , Random Forest Algorithm , ANN(Artificial Neural Network) , Confusion Matrix. The efficiency of the modeldeveloped shall be checked by the metrics like –Accuracy , Precision , Recall (Sensitivity) ,F-score. Later, the model results shall be compared with similar papers/ projects withalready existingand operational model to check and conclude the most efficient method (as per the metrics)among those , discussed here.

Keywords–Comparison, Metrics , Accuracy, Precision.

INTRODUCTION

Payments of Goods and Services are becoming more and more elephantine. Whether you gotoshopping for garments , want to have/ pickquick snacksat your favoritefood chains OR eateries, halting at thegas station for refueling, etc., digital payments have dominated the way we all made payments a few decades back. The first digital paymentmethod dates back to early 1870s , termed as Electronic Fund Transfer(EFT) –a small technological move towardsthe omittanceof Liquid Money , or in other words, Paper money (cash).But now , the worldis living the 21stcentury, which most people often whooped “The Era of Technology”. Almost Every person today has a bank account , which can be accessed via Internet.Butyou won’t be doing an OnlineBanking Transaction for , say , buying a pair of Shoesata footwear store.You will be needing a more convenient way withminimalisticGUI interface. Here comes Credit / Debit Cards.These are made of complex materials of plastic andPVC and silicon plates / coatingsand coded/ programmedwhich can be swiped through hand-held machines / devices , which directly deducts the amount from the bank account for the goods / service. Now let us talkaboutthe handicapped part of Credit Cards.Though people use it heavily , but they are unaware of the upcomingproblems. Many stores use cloning machines which copies all the data of the card , giving access to the person s’bank account –resulting to a huge fraud. Likewise , shopkeepers are also dupped by false credit cards.

Fraudsters never let an opportunity go towaste –whether the whole world is fighting a war , whether any countryis facing economic crisis –fraudsters have only one motivation –getting the money.For thepast decade , credit cardfraudulent cases have rose by 400%(if counted and checkedannually). People with neweraccounts are duped more than the older/ existing ones by 24%.In 2019 ,morethan 1500 data breaches occurredand millions of user recordswere exposed.

Earlier , CardshadMagnetic Strips which contained card holders’information(they are still in use). But now , Smart Card has emerged as a new hope of secured transaction. Thesedo not contain any magnetic strips , rather they are embedded with Integrated Circuits (ICs) –which enables us to just touch / tap the card to the machine without any swiping and password / pin.Despite of so many technological advancements in termsof cybersecurity and secured banking systems , mishaps occur almost every daywith someoneat any moment of time.

Thoughthe chances are good enough to get duped , there are many machine learning algorithms/ techniqueswhich can be put on field to work and subjugate this pest.

Literature Survey

Authors	Year	Methodology and Techniques used	Advantages	Issues	Metrics Used
Vaishnavi Nath Dornadula Geetha S	2019	Sliding-Window Method	Lead the system to adapt to new cardholder's transaction behaviours timely	To design and develop a novel fraud detection method for Streaming Transaction Data, with an objective, to analyse the past transaction details of the customers and extract the behavioural patterns.	Accuracy Precision Mathews Correlational Coeff. (MCC)
Massimiliano Zanin Miguel Romance Santiago Moral Regino Criado	2018	Parenclitic networks Artificial Neural Networks --- (ANN) Multi-Layer Perceptrons (MLP)	Increased efficiencyin detecting frauds in some niches of operations, like medium-sized and on-line transactions.	To detect illegal instances in a real card transaction dataset.	True Positive Rate (TPR) Receiver Operating Characteristic (ROC) False Positive Ratio (FTR)
Dr . Yvan Lucas Dr . Johannes Jurgovsky	2020	Feature engineering techniques Recurrent Neural Networks **(LSTM) – RNN** Graphical Models (Random fore)	□ Making a strong difference for fraudulent transactions , whichare much more rarethan genuine transactions. □ Allowing Fraud detection systems to obtain good performances	To identify fraudulent transactions that have been issue illegitimately on behalf of the rightful card owner.	Precision True Positive Rate (TPR) Accuracy F1 Score Mathews Correlational Coeff. **(MCC)**
Navanshu Khare Saad Yunus Sait	2018	Decision Tree Random Forest (SUP) SVM Model -- Support Vector Machine Logistic Regression	It's known that Random Forest algorithm will perform better with a larger number of training data.	To check the performance of Decision Tree , Random Forest , SVM and Logistic Regression on highly skewed Credit Card Fraud Data.	Accuracy Sensitivity Specificity Precision
Dahee Choi Kyunghoo Lee	2018	Machine Learning Deep Learning Feature Selection Sampling Supervised Algorithms Unsupervised Algorithms HMM -- Hidden Markov Model	□ ML based method has higherdetection than neural. □ Networks at various ratios Neural Networks gets accuracy as high as 95%.	Proposing a process for accurate fraud detection -- the overall process of detecting financial fraud based on ML andcomparing it with ANN approach to detect fraud and process large amount of financial data.	Accuracy F- measure

Samaneh Sorournejad Zahra Zojaji Reza Ebrahimi Atani Amir Hassan Monadjemi	2016	Misuse detection ANN - Artificial Neural Network Supervised Techniques Unsupervised Techniques Negative Selection Algorithm Genetic Algorithm HMM -- Hidden Markov Model SVM	Efficient Accuracy and Maintainability , Improved Capability of Pattern Recognition Improved Working with Noisy Data.	To review the state of the art in credit card fraud detection techniques, datasets and evaluation criteria.	Accuracy Precision / Hit Rate True Positive Rate / Sensitivity False Positive Rate ROC Cost F1 measure
		Bayesian Network Fuzzy Neural Network			
Yashvi Jain NamrataTiwari ShripriyaDubey Sarika Jain	2019	Support VectorMachine (SVM) Artificial Neural Networks (ANN) Bayesian Network K- Nearest Neighbour (KNN) Hidden Markov Model Fuzzy Logic Based System Decision Trees	Creation of hybrid of various techniques that arealready used in fraud detection to cancel out them limitations and get enhanced performance.	To introduce the concept of fraud related to credit card(s) and them various types and their solutions - valid against modern tricks and Hacks.	Accuracy Detection Rate False Alarm Rate Sensitivity Specificity Cost

Sonal Mehndiratta Kamal Gupta	2019	Data Mining Machine Learning Classification ANN - Artificial Neural Network Genetic Algorithm (GA) HMM -- Hidden Markov Model K- Nearest Neighbour (KNN) Naive Bayes	□ Majority of voting methods achieve good accuracy rates in order to detect thefraud in the credit cards. □ Proposed random forests provide good results on the small dataset. □ There is a large moving window, higher number of attributes and number of link types available which can be searched by CD and SD algorithms.	To avoid the leakage of information to the attacker , which if fails willlead to huge amounts of loss to the Credit card company.	Precision Sensitivity Accuracy Balanced Classification Rate
Kaithekuzhical Leena Kurien Dr . Ajeet Chikkamannur	2019	Machine Learning Artificial Intelligence (AI) Deep Learning	□ Efficient Over Sampling , Under-sampling of data for accuracy. □ Thorough Analysis of the features and sub-sample ratios for imbalanced Datasets.	Probability of fraudulent transactions in prevalence and context of credit card usage.	Feature Co-Relation Quartile True positive(TP) True Negative(TN) False positive(FP) False negative(FN) Precision F1 Score Recall
		Support Vector Machine Genetic Algorithm Bayesian Network	□ By using KNN, it acts as a better classifier for credit card detection.	Solution to the fraud.	Specificity Sensitivity

Apapan Pumsirirat Liu Yan	2018	Creating an Auto Encoder using Deep Learning TensorFlow (google library)	□ Accurately achieve credit carddetection with a large dataset. □ Guarantee thatAE and RBM can make more accurate AUC for receiver operator characteristics.	□ To focus on fraud cases that cannot be detected basedon previous history or supervised learning. □ To create a model of deep Auto-encoder and Restricted Boltzmann Machine (RBM) that can reconstruct normal transactions to find anomalies from normal patterns.	MSE RMSE
Daniyal Baig	2020	Artificial Neural Network	□ Gives the highest accuracy ascompared to other logistic regressions.	To identify the Credit card fraudand provide a reasonable	Accuracy Precision
Ishu Trivedi Monika Mrigya Mridushi	2016	Artificial Neural network Genetic Algorithms Neural Network Bayesian Network	□ This method proves accurate infinding out the fraudulent transactions and minimizing the number of false alerts. □ probability of detecting fraud in avery short span of time after the transactions has been made.	The main aim isto detect the fraudulent transaction andto develop a method of generating test data.	Time Speed Accuracy Rate of Error
Yong Fang Yunyun Zhang Cheng Huang	2019	Light-GBM model Imbalanced data Random Forest Algorithm Gradient Boosting Machine Algorithm (GBM)	The result indicatesthat the new modelis fast, has a lesser error rate and is more accurate.	The goal is to compare the older system with the new system and see their differences	Accuracy Efficiency Recall Rate Speed

S. Abinayaa H. Sangeetha R.A. Karthikeyan K. Saran Sriram D. Piyush	2020	Random Forest Algorithm Machine Learning	By using machine Learning alongside Random Forest Algorithm we can get a better result in the detection of fraud.	To find the most common methods of fraud alongside their detection methods and algorithms.	Sensitivity Precision Accuracy
Nishant Sharma	2019	Random Forest Algorithm AdaBoost Classifier XGBoost Classifier LightGBM Classifier	By this test we can conclude that XGBoost Classifier provides the highest accuracy on credit card fraud detection.	To investigate and check the performance of mentioned techniques onhighly skewed credit card fraud data.	Accuracy Performance
Wael Khalifa Mohamed Ismail RoushdyAbdel-Badeeh M. Salem Hossam Eldin Hossam Eldin Mohammed Abd El-Hamid	2019	Machine Learning Data Mining Support Vector Machine Bayesian Network	By this research survey, we can figure out which technique is the best to detect credit card fraud and which technique can be inour system.	The goal of this research is to survey procedure of various discoverytechniques dependent on Visa.	Accuracy Speed Precision Exactness Cost
Surbhi Gupta Nitima Malsa Vimal Gupta	2017	Artificial Immune System (AIS) Neural Network Genetic Algorithms Decision Tree Support Vector Machine (SVM)	□ By this survey,you can find the differences between each technique mentioned. □ New classifierscan be formed by combining the different techniques.	The goal is to compare different techniques by either testing them separately or by combining and forming new classifiers which can be used to improve the detection of credit card fraud.	Speed of detection Accuracy Cost
Shiv Shankar Singh	2019	Data Mining Logistic RegressionDecision tree Support Vector Machine (SVM)	This survey makes sure that more businesses are aware of the different techniques available to catch credit card fraud.	This presented paper focuses onfraud activities that cannot be detected manually by carrying out research and examine the results of logisticregression, decision tree and support vector machine	Precision Cost Exact outcome Rapidly train machines

Suraj Patil Varsha Nemade Piyush Kumar Son	2018	Logistic regression Designing analytical model for Fraud prediction Decision Tree Random Forest	In this paper we have proposed a robust framework to process large volume of data, thefunctionality of framework can be extended to extractreal time data fromdifferent desperatesources.	The main challenge for today's CCFD system is how to improve fraud detection accuracy with growing numberof transactions done by user persecond.	Precision FDR FOR LR Sensitivity Miss Rate
			These analytical models are run on credit card dataset and accuracy of analytical model is evaluated with helpof confusion matrix".		Fallout Specificity F1 Score
ALTYEB ALTAHER TAHASHARAF JAMEEL MALEBARY	2020	Light-GBM algorithm K-fold cross-validation (CV) Gradient-based oneside Sampling (GOSS) Decision tree	The results reveal that the proposed algorithmis superior to other classifiers. The results also highlight the importance and value of adopting an efficient parameter optimization strategy for enhancing the predictive performance of the proposed approach.	To demonstrate the effectivenessof our proposed Light-GBM for detecting fraud in credit card transactions, experiments were performed using two real- world public credit card transaction data sets consisting fraudulent transactions and legitimate ones.	Accuracy Precision Recall F1-score
Niloofar Yousefi Marie Alaghband Ivan Garibay	2019	Machine learning Supervised learningK- means algorithm ANN Decision Tree (DT) Support-vector machine(SVM)	During this survey, we noticed that supervised learning techniqueshave been used more frequently than unsupervised methods. To be more specific, the most commonly used fraud detection techniques are LR,ANN, DT, SVM andNB.	To drive the future research agenda for the community in order to develop more accurate, reliable and scalable models of credit card fraud detection.	Accuracy Precision Stable

Shimin LEI Ke XU YiZhe HUANG Xinye SHA	2020	CNN XGboost based model	This Survey presents an XGboost-based financial system to detect transaction fraud.	To introduce a new set of features based on analyzing the periodic behavior of the time of a transaction usingthe von Mises distribution.	Accuracy Auc-Roc score True positive rate (TPR) False positive rate (FPR)
Mehak Mahajan Sandeep Sharma	2019	Supervised learning SVM Logistic Regression Naive Bayes Neural NetworkK- NN Decision tree	□ In this paper wehave discussed various techniquesof data mining through which we can detect the fraud. □ Various techniques like Hidden Markov Model, K-mean clustering algorithm, K-nearest neighbor, Decision Tree, Fusion approach due using dumpsterShafer, Bayesian Network, Neural Network, SVM and Logistic Regression are used.	In this paper we have research about the various detecting techniques to identify and detect the fraud through varied techniques of data mining.	Accuracy Efficiency Precision
Ali Yeşilkanat Barış Bayram Bilge Köroğlu Seçil Arslan	2020	Gradient Boosting Tree (GBT) XGBoost	In this work, for the real-time detection of credit card fraud,a new approach is proposed for training dataset construction to make usable and valuable of different types of attributes of a transaction by combining numerical, hand- crafted numerical,categorical and textual features.	In this survey, new strategy for training dataset generation employing the sliding window approach in a given time frameto adapt to the changes on the trends of fraudulent transactions.	False-Positive Rate (FPR) Recall Precision Area Under Curve (AUC)

Alejandro Correa Bahnsen Djamila Aouada Aleksandar Stojanovic Björn Ottersten	2016	Neural networks Bayesian learning Association rules Hybrid models Support vector machines Peer group analysis Random Forest	□ In this paper, we address the cost-sensitivity and the features pre- processing to achieve improved fraud detection and savings. □ In this paper, we proposed a new cost-based measureto evaluate credit card fraud detection models, taking into account the different financial costs incurred by the fraud detection process.	In this paper we expand the transaction aggregation strategy, and propose to create a new set of features based on analyzing the periodic behavior of the time of a transaction usingthe von Mises distribution.	Accuracy Recall Precision F1-Score
Yaodong Han Shun Yao Tie Wen Zhenyu Tian Changyu WangZheyuan Gu	2020	Logistic Regression Decision Tree Random Forest Naïve Bayes Boosted Tree AdaBoost Neural Network Support vector machine (SVM)	□ In this article, we build machine learning models on a synthetically generated dataset about credit card applications and evaluated their performance. □ As for building more models, we can also try ensemble model, such as voting or stacking classifiers, to further improve performance.	Traditional approaches, such as expert system, suffers from the incapability to handle complex problems and tremendous amount of data,while the recentdevelopment ofvarious machine learning techniques brings new solutions.	Accuracy Precision
		K-Nearest Neighbor			
Ronish Shakya	2018	Bagging Boosting Logistic regression Random Forest XGBoost	□ We implemented the algorithmic approaches such asbagging and boosting to tackle the class imbalance problem. □ Besides these models, we chose logistic regression model to compare with other models. □ Then, we analyzed all threemodels with and without using resampling techniques.	In this survey to tackle this credit card fraud problem, data- level approach, where different resampling methods such as under-sampling, oversampling, and hybrid strategies, have been implemented along with an algorithmic approach where ensemble models such as bagging and boosting have been applied to a highly skewed dataset containing 284807 transactions.	True Positive (TP) False Positive (FP) True Negative (TN) False Negative (FN)

Issues In Existing System

1.Enormous Data is processed every day and the model build is not that fast enough to respond to the scam in time.

2.Imbalanced Data i.e most of the transactions *(99.8%)*are not fraudulent which makes it really hard for detecting the fraudulent ones

3.Data availability as the data is mostly private.

4.Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported.

5.Adaptive techniques used against the model by the scammers.

Motivation & Objective

Motivation:

Usage of liquid cash is getting neglected day-by-day because of the increased usage of credit card payments method. There are a few countries which are completely relying upon digital transactions , omitting paper currency. But that doesn’t mean this kind of payment method are completely reliable. As the cybersecurity field is excelling at a rapid pace , accessing a bank account is not a tough job to do today. Thus , Credit Card fraud cases are also piling up in every developed and developing country.

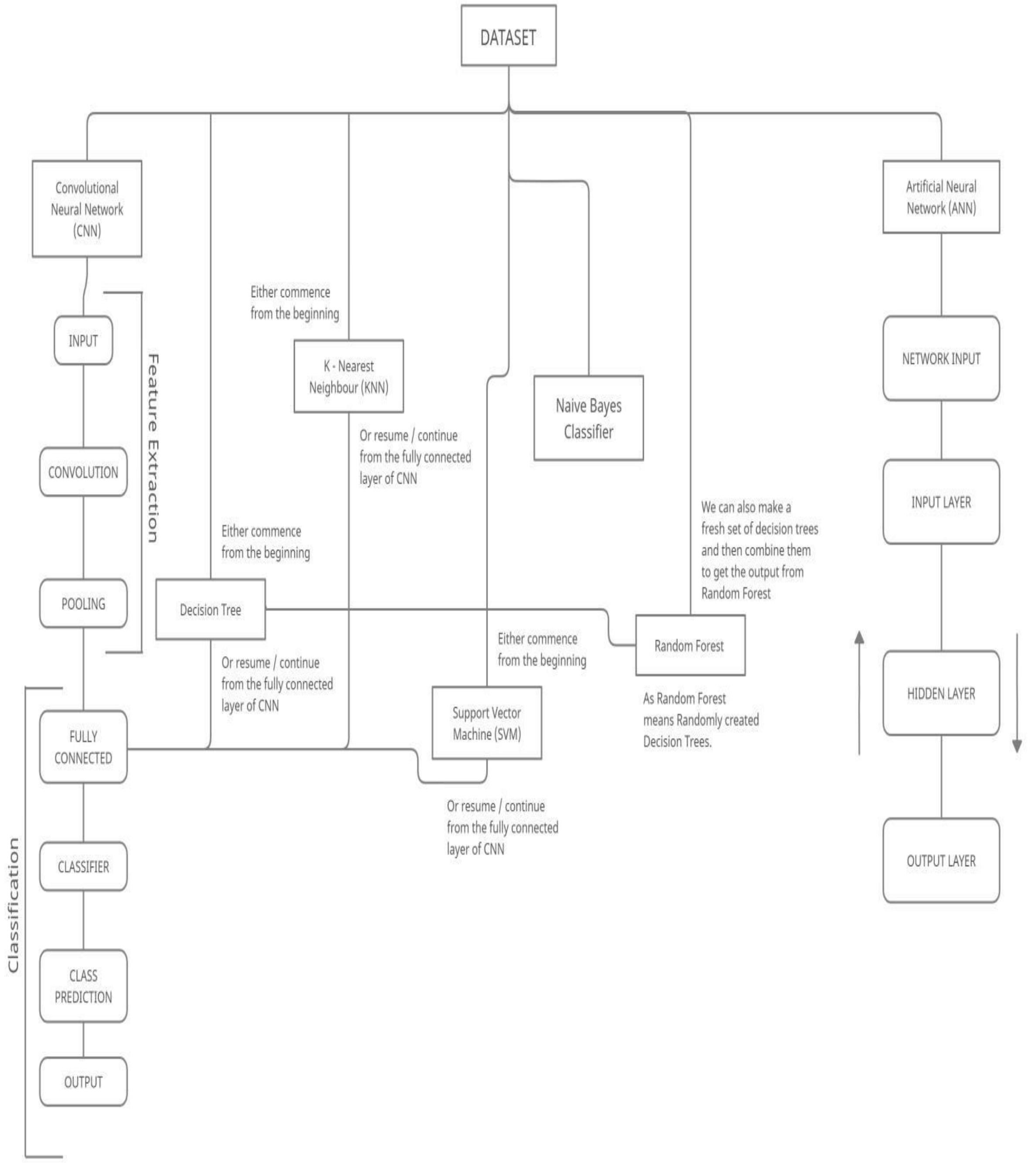
Objective:

This project aims to compare all the trending efficient methods against Credit Card Frauds , by training the model using dataset (from Kaggle). The model consists of 4 modules / techniques namely –CNN (Convolutional Neural Network) , Decision Tree Algorithm , K-NN (K – Nearest Neighbour) and XG booster . The efficiency of the model developed shall be checked by the metrics like –Accuracy , Precision , Recall (Sensitivity) , F-score.

Problem Statement

This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not.

ARCHITECTURE OF THE PROPOSED SYSTEM


DATASET
Convolutional Neural Network (CNN)
INPUT
CONVOLUTION
POOLING
Feature Extraction
FULLY CONNECTED
CLASSIFIER
CLASS PREDICTION
OUTPUT
Classification
Either commence from the beginning
K - Nearest Neighbour (KNN)
Or resume / continue from the fully connected layer of CNN
Either commence from the beginning
Decision Tree
Or resume / continue from the fully connected layer of CNN
Naive Bayes Classifier
Either commence from the beginning
Support Vector Machine (SVM)
Or resume / continue from the fully connected layer of CNN
We can also make a fresh set of decision trees and then combine them to get the output from Random Forest
Random Forest
As Random Forest means Randomly created Decision Trees.
Artificial Neural Network (ANN)
NETWORK INPUT
INPUT LAYER
HIDDEN LAYER
OUTPUT LAYER

Modules & Description

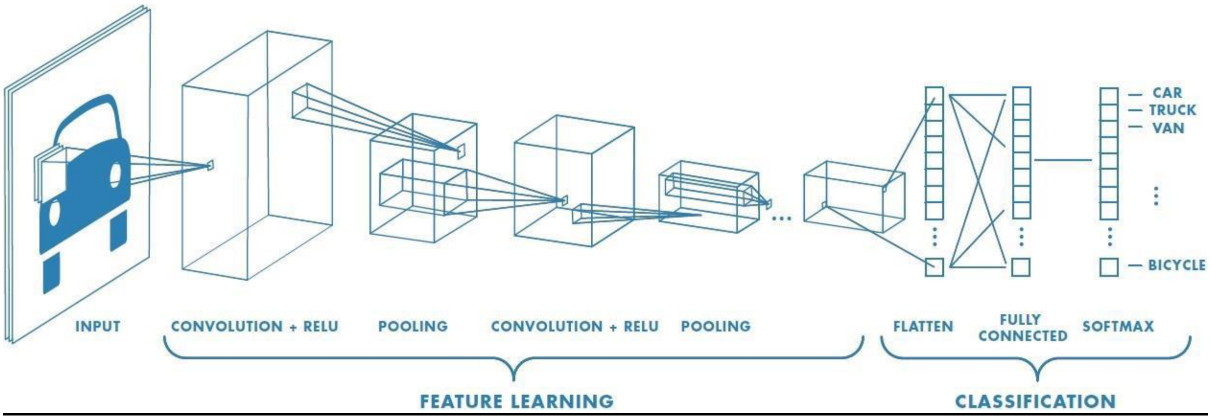
1. CONVOLUTIONAL NEURAL NETWORK :

Being a Deep Learning Algorithm , it can take input image, assign importance to various aspects / objects in the image and be able to differentiate one from the other. The architecture of a CNN (also called ConVet) , is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex (The visual cortex of the brain is the area of the cerebral cortex that processes visual information).

We propose to get familiar with a Decision tree, which explains the particular justification every forecast made by the CNN at the semantic level. i.e., the Decision tree expands feature representations in high conv-layers of the CNN into Primary Concepts of item parts. Thusly, the Decision Tree tells individuals which article parts enact which channels for the forecast and the amount they add to the metrics' score. 36

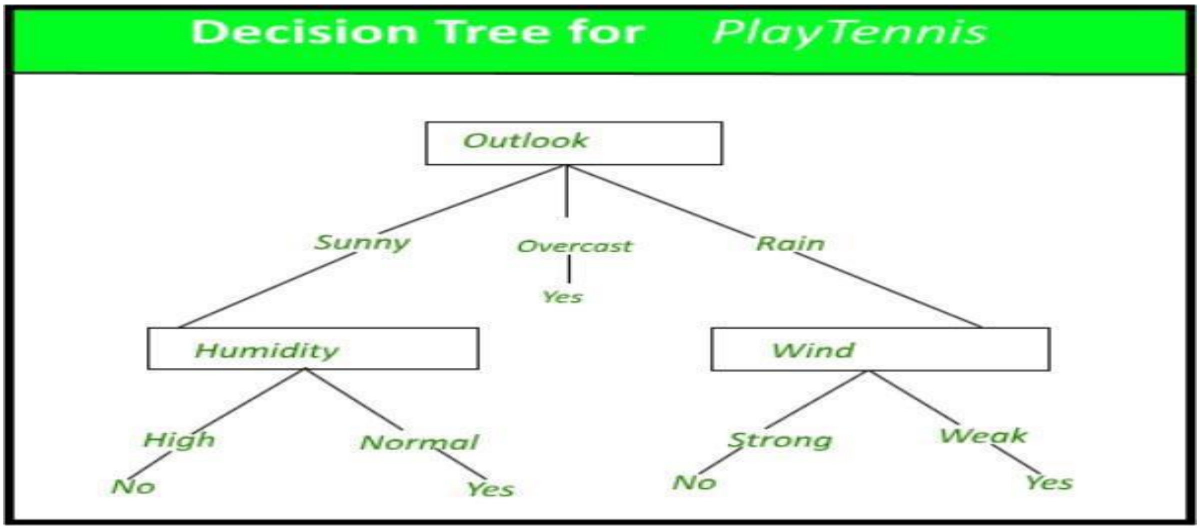
The CNN-KNN model uses advantages of both methods (CNN and KNN). The advantages of CNN are sparse connectivity among the neurons between successive layers and weights sharing between layers. The KNN classifies the nearest data samples as a class based on similar measures. This CNN - KNN model extracted the salient features automatically and reduce the laborious and time consumption.

Studies shows that using a CNN - SVM model increases the accuracy of the obtained results , rather than interpreting from simple / traditional CNN models. Thus , SVM is used along with CNN to increase Accuracy.



2. DECISION TREE ALGORITHM :

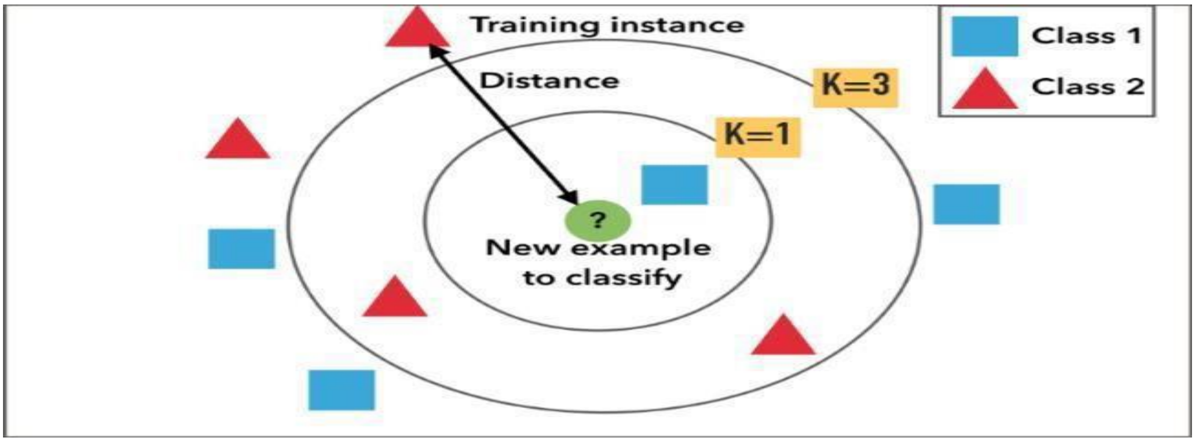
It’s a Supervised Learning Algorithm.A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It’s mostly is used here to sort the information received and sorts and classifies the data into different groups.



3. K-NEAREST NEIGHBOUR :

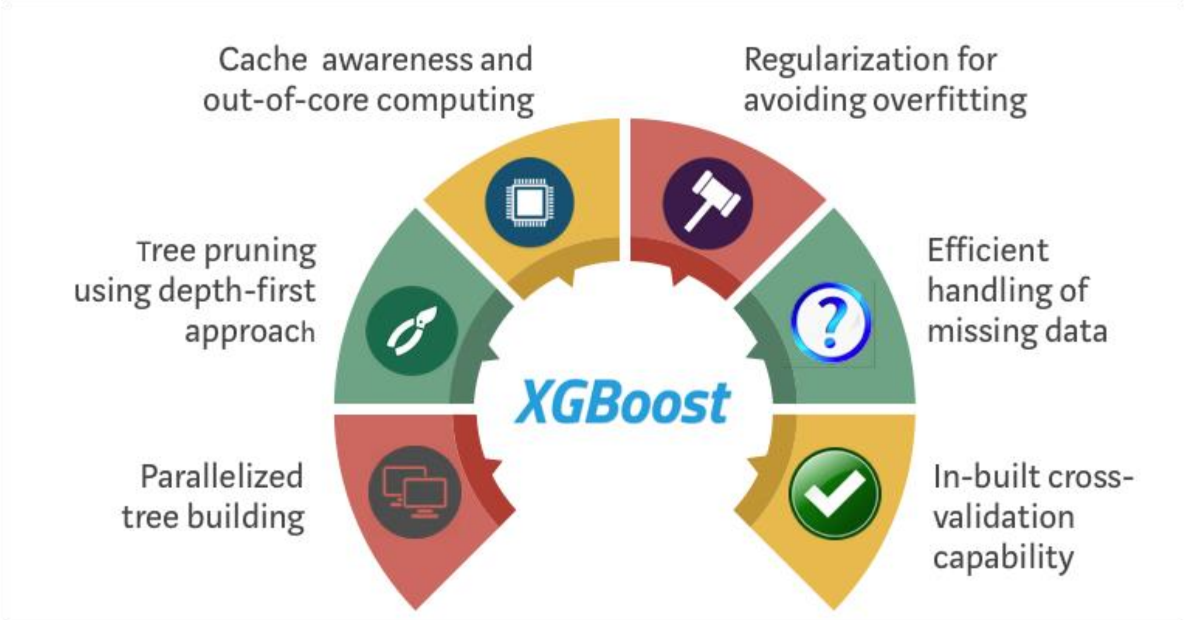
K-Nearest Neighbor is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms suchas GMM, which assume a Gaussian distribution of the given data).



4. XG Booster

XGBoost is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. XGBoost models majorly dominate in many KaggleCompetitions.


Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

Evaluation Matrix

Accuracy: It’s the ratio of correctly predicted observation to the total observations.

Accuracy = TP+TN / TP+FP+FN+TN

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP / TP+FP

Recall (Sensitivity): Recall is the ratio of correctly predicted positive observations to the all observations in actual class -yes.

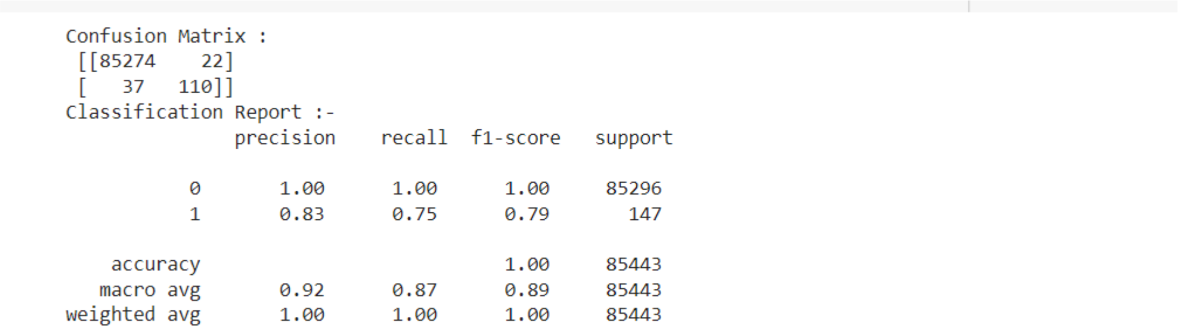
Recall = TP / TP+FN

F1 Score: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

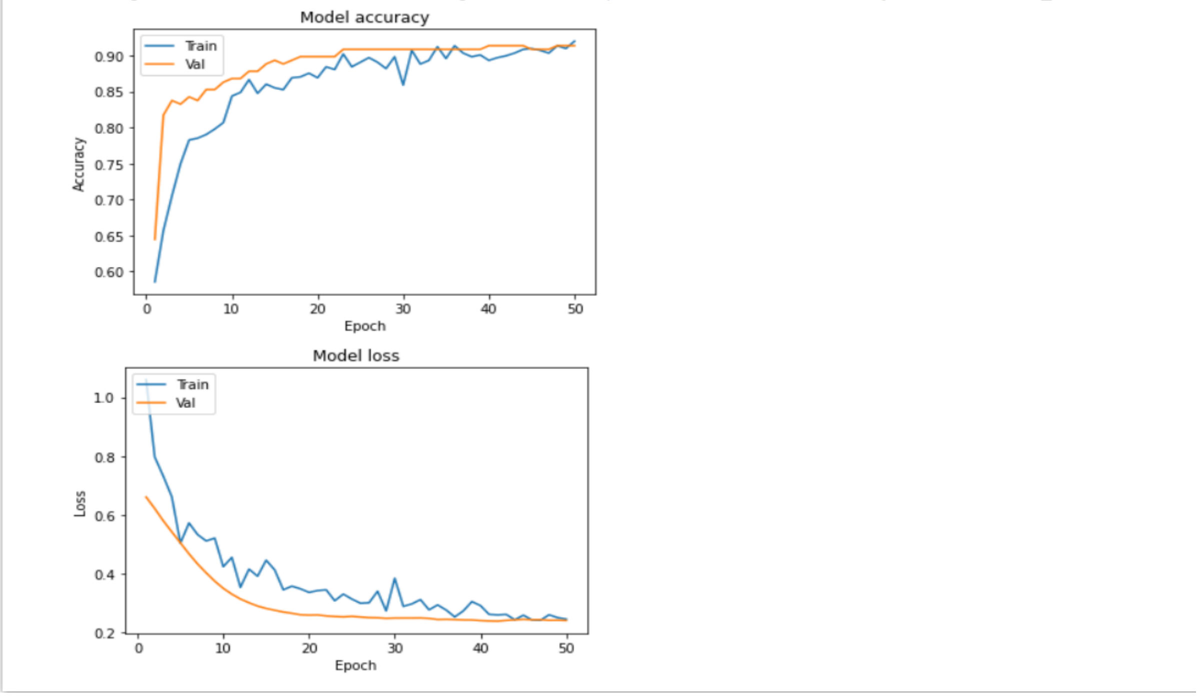
Results

Decision Tree

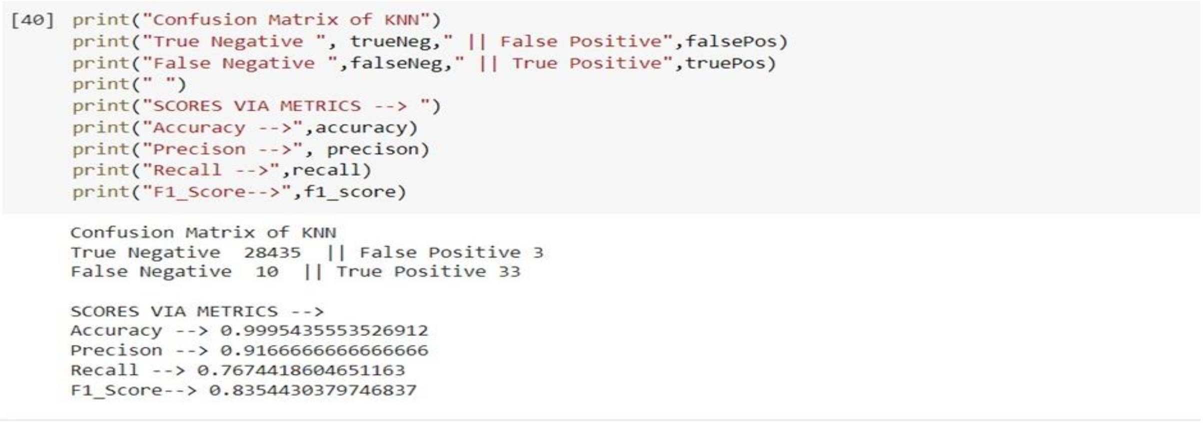


```
Confusion Matrix :
 [[85274    22]
 [   37   110]]
Classification Report :-
               precision    recall  f1-score   support

           0       1.00      1.00      1.00     85296
           1       0.83      0.75      0.79       147

    accuracy                           1.00     85443
   macro avg       0.92      0.87      0.89     85443
weighted avg       1.00      1.00      1.00     85443
```

CNN



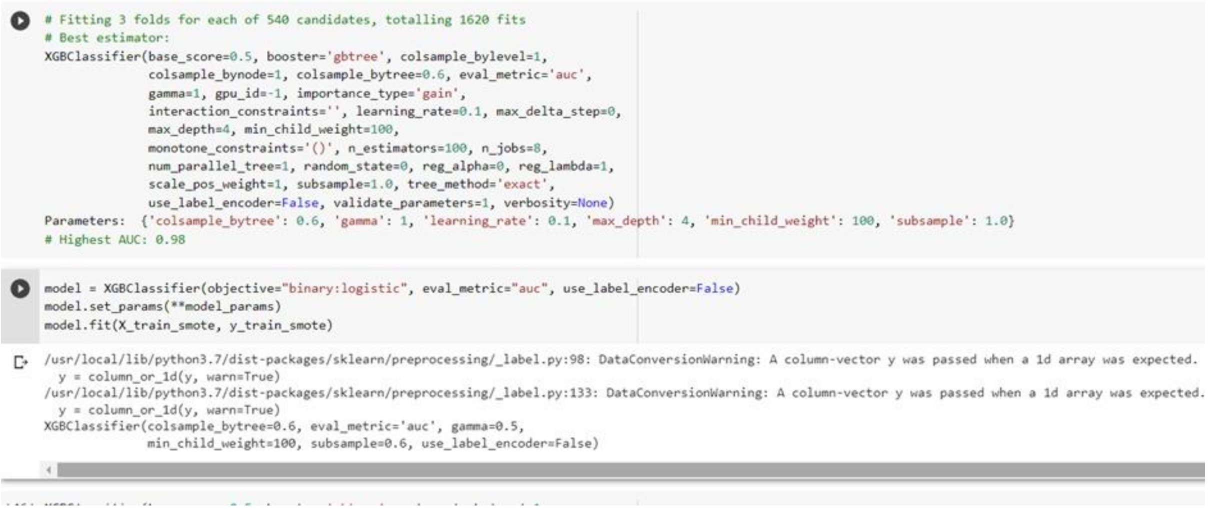
KNN



```
[40] print("Confusion Matrix of KNN")
     print("True Negative ", trueNeg," || False Positive",falsePos)
     print("False Negative ",falseNeg," || True Positive",truePos)
     print(" ")
     print("SCORES VIA METRICS --> ")
     print("Accuracy -->",accuracy)
     print("Precison -->", precison)
     print("Recall -->",recall)
     print("F1_Score-->",f1_score)
```

```
Confusion Matrix of KNN
True Negative  28435  || False Positive 3
False Negative  10  || True Positive 33

SCORES VIA METRICS -->
Accuracy --> 0.9995435553526912
Precison --> 0.9166666666666666
Recall --> 0.7674418604651163
F1_Score--> 0.8354430379746837
```

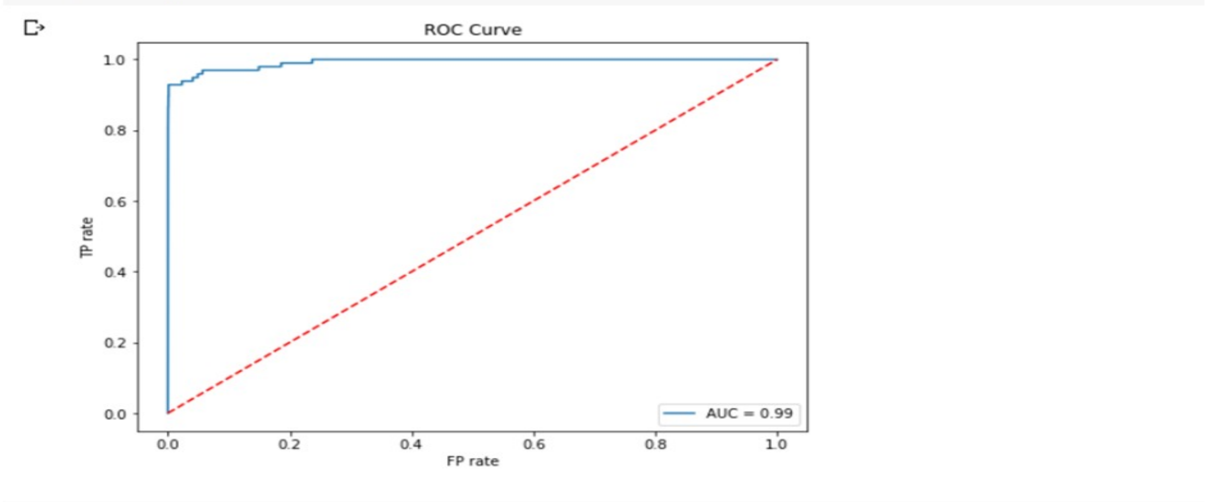
XG Boost



```
# Fitting 3 folds for each of 540 candidates, totalling 1620 fits
# Best estimator:
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=0.6, eval_metric='auc',
              gamma=1, gpu_id=-1, importance_type='gain',
              interaction_constraints='', learning_rate=0.1, max_delta_step=0,
              max_depth=4, min_child_weight=100,
              monotone_constraints='()', n_estimators=100, n_jobs=8,
              num_parallel_tree=1, random_state=0, reg_alpha=0, reg_lambda=1,
              scale_pos_weight=1, subsample=1.0, tree_method='exact',
              use_label_encoder=False, validate_parameters=1, verbosity=None)
Parameters:  {'colsample_bytree': 0.6, 'gamma': 1, 'learning_rate': 0.1, 'max_depth': 4, 'min_child_weight': 100, 'subsample': 1.0}
# Highest AUC: 0.98
```

```
model = XGBClassifier(objective="binary:logistic", eval_metric="auc", use_label_encoder=False)
model.set_params(**model_params)
model.fit(X_train_smote, y_train_smote)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/preprocessing/_label.py:98: DataConversionWarning: A column-vector y was passed when a 1d array was expected.
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.7/dist-packages/sklearn/preprocessing/_label.py:133: DataConversionWarning: A column-vector y was passed when a 1d array was expected.
  y = column_or_1d(y, warn=True)
XGBClassifier(colsample_bytree=0.6, eval_metric='auc', gamma=0.5,
              min_child_weight=100, subsample=0.6, use_label_encoder=False)
```



Comparative Study

Method	Accuracy	Precision	F1score	Recall
CNN	93.7%	-	-	-
Decision Tree	99.93%	83%	75%	79%
KNN	99.95%	91.6%	76.7%	83.5%
XG boost	99.99%		-	

Novelty

Most of the other projects and research papers have used KNN, CNN, ANN, SVM, Decision Tree and Random Forest models but what we have done is we have taken an another model that is XGBoost, SMOTE, and threshold moving. The accuracy of XGB is 99.99% which is greater then other modules that we have used that is KNN ,CNN And Decision Tree.

Conclusion

This paper compared various machine learning and deep learning techniques with respect to Evaluation Metrics , which are –Accuracy , Precision , F1 Score , Recall. Among all the discussed models in our project , we conclude that –XGB has the highest Accuracy (99.99%) and CNN (Convolutional Neural Network) has the lowest Accuracy (93.65%). In case of Precision , KNN dominates above all (91.6%) while Decision Tree comes at the bottom (83.00%). For Recall (Sensitivity) , KNN dominates (83.5%) , while Decision Tree drops down (79.00%). Last but not the least , for F1 Score , KNN dominates (76.7%) while Decision Tree drops (75.00%).

CODE’S LINK:

CNN:
https://colab.research.google.com/drive/1FwuuGyt_p1Yj10ehpJGEUJsjwaIpFSZC?usp=sharing

KNN:
https://colab.research.google.com/drive/13HHCBrkSOHUsvPvWaj9jrjSrRCL1I1W8?usp=sharing

Decision Tree

https://colab.research.google.com/drive/1_w0FrizZBsHfglE9jWgYYw88nqsPR3cL?usp=sharing

XG booster
https://colab.research.google.com/drive/1RRcGnp6skUOQJy03UyAUy0NVam_pNZ0T?usp=sharing

DATA SET:
https://www.kaggle.com/mlg-ulb/creditcardfraud

References

[1] Vaishnavi Nath Dornadula , S Geetha (2019) , Credit Card Fraud Detection using Machine Learning Algorithms

[2] Massimiliano Zanin , Miguel Romance , Santiago Moral , Regino Criado (2018) , Credit Card Fraud Detection through Parenclitic Network Analysis

[3] Yvan Lucas, Johannes Jurgovsky (2020) , Credit card fraud detection using machine learning: A survey

[4] Navanshu Khare , Saad Yunus Sait (2018) , Credit Card Fraud Detection Using Machine Learning Models and Collating MachineLearning Models

[5] S P Maniraj, Aditya Saini, Shadab Ahmed, Swarna Deep Sarkar (2019) , Credit Card Fraud Detection using Machine Learning and Data Science

[6] Dahee Choi , Kyungho Lee (2018) , An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment:A Survey and Implementation

[7] Samaneh Sorournejad , Zahra Zojaji , Reza Ebrahimi Atani , Amir Hassan Monadjemi (2016) , A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective

[8] Yashvi Jain , NamrataTiwari , Shripriya Dubey , Sarika Jain (2019) , A Comparative Analysis of Various Credit Card Fraud Detection Techniques

[9] Sonal Mehndiratta , Mr. Kamal Gupta (2019) , Credit Card Fraud Detection Techniques: A Review

[10] Kaithekuzhical Leena Kurien , Dr. Ajeet Chikkamannur (2019) , DETECTION AND PREDICTION OF CREDIT CARD FRAUD TRANSACTIONS USING MACHINE LEARNING

[11] Apapan Pumsirirat , Liu Yan (2018) , Credit Card Fraud Detection using Deep Learning based on AutoEncoder and RestrictedBoltzmann Machine

[12] Daniyal Baig (2020) , Credit Card Fraud Detection Using Supervised Learning Algorithms.

[13] S. Abinayaa , H. Sangeetha, R. A. Karthikeyan , K. Saran Sriram, D. Piyush (2020) , Credit Card Fraud Detection and Prevention using Machine Learning

[14] Nishant Sharma (2019) , CREDIT CARD FRAUD DETECTION PREDICTIVE MODELING 91

[15] Wael Khalifa , Mohamed Ismail Roushdy , Abdel-Badeeh M. Salem , Hossam Eldin , Hossam Eldin Mohammed Abd El-Hamid (2019) , Machine Learning T Machine Learning Techniques for Credit Card Fraud Detection and Detection

[16] Surbhi Gupta , Mrs. Nitima Malsa , Mr. Vimal Gupta (2017) , Credit Card Fraud Detection & Prevention –A Survey

[17] Shiv Shankar Singh (2019) , Electronic Credit Card Fraud Detection System by Collaboration of Machine Learning Models

[18] Ishu Trivedi , Monika , Mrigya Mridushi (2016) , Credit Card Fraud Detection

[19] Yong Fang , Yunyun , Zhang Cheng Huang (2019) , Credit Card Fraud Detection Based on Machine Learning

[20] Aman Gulat , Prakash Dubey , MdFuzailC , Jasmine Norman , Mangayarkarasi R (2017) , Credit card fraud detection using neural network and geolocation

[21] N.Geetha , T.Kavipriya (2017) , Study on Credit Card Fraud Detection Using Data Mining Techniques

[22] Suraj Patil ,Varsha Nemade , PiyushKumar Son (2018) , Predictive Modelling for Credit Card Fraud Detection Using Data Analytics

[23] Altyeb Altaher Taha , Sharaf Jameel Malebary (2020) , An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine

[24] Niloofar Yousefi , Marie Alaghband,Ivan Garibay (2019) , A Comprehensive Survey on Machine Learning Techniques and User Authentication Approaches for Credit Card Fraud Detection

[25] Shimin LEI , Ke XU , YiZhe HUANG , Xinye SHA (2020) , An XGboost based system for financial fraud detection

[26] Mehak Mahajan , Sandeep Sharma (2019) , Detect Frauds in Credit Card using Data Mining Techniques

[27] Ali Yeşilkanat , Barış Bayram , Bilge Köroğlu,Seçil Arslan (2020) , An Adaptive Approach on Credit Card Fraud Detection Using Transaction Aggregation and Word Embeddings

[28] Alejandro Correa Bahnsen,Djamila Aouada,Aleksandar Stojanovic,Björn Ottersten,(2016),Feature engineering strategies for credit card fraud detection

[29] Yaodong Han , Shun YaoTie Wen , Zhenyu Tian , Changyu Wang , Zheyuan Gu (2020) , Detection and Analysis of Credit Card Application Fraud Using Machine Learning Algorithms

[30] Ronish Shakya (2018) , Application of Machine Learning T Application of Machine Learning Techniques in Credit Card Fraud Detection