

Credit Card Fraud Detection

*COEN 240
Machine Learning*

Report by:

*Deep Savla (W1652329)
Jash Hemant Shah (W1650179)*

Abstract:

Our project is aiming to use algorithms like Support Vector Machine, the K-Nearest Neighbors (KNN) and Logistic regression, that are highly utilized machine learning algorithms for classification of data in a binary manner. We use them for training a model which will help in distinguishing between credit card transactions that are either fraudulent or legal. Moreover, we also will be evaluating how accurate all these three models are.

Introduction:

In today's time and age, the major chunk of transactions are carried out electronically, which has led to the drastic increase in the importance of making sure that a particular transaction is legitimate or a fraudulent one. And in this process, credit cards play an important role to make this system trustworthy, where we could provide protection against such fraudulent activities. So when we talk about Credit card frauds, they are a type of identity theft in which the criminal makes unauthorized purchases or takes cash in advance using someone else's name. And so, detection of these kinds of frauds become a task of utmost importance such that we could minimize the financial losses for both parties, the card issuer and also the person who holds that card.

Concepts:

Mentioned below are the three machine learning algorithms that we have used and compared to get to the most efficient algorithm:

1. Logistic Regression
2. Support Vector machine
3. K-Nearest Neighbors

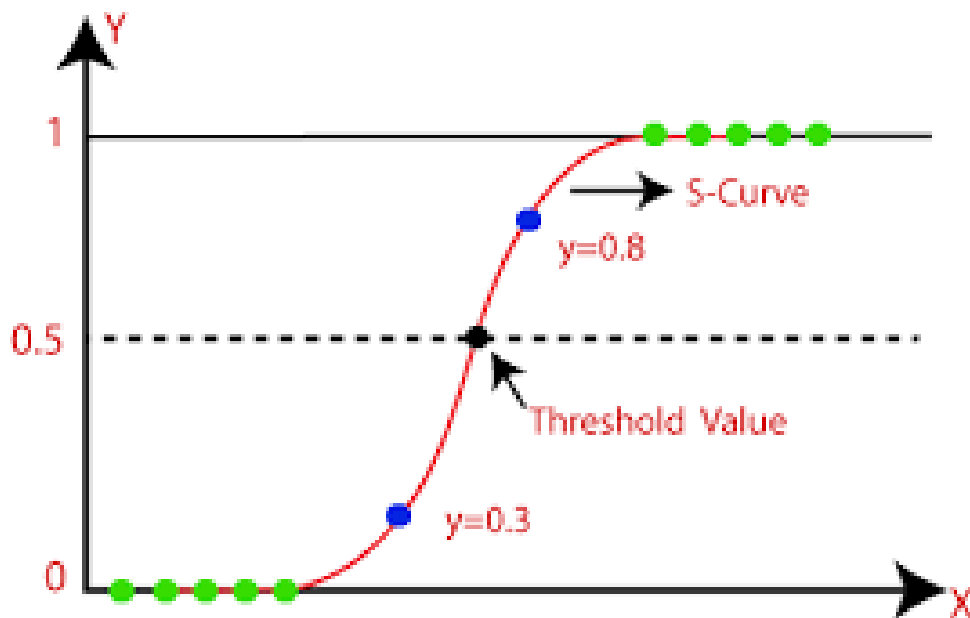
Logistic regression:

Logistic regression is a machine learning algorithm that is highly used for binary classification problems and is based on statistical modeling. It comes in really handy when the output variable needs to be categorical and the aim is to find the probability for the occurrence of an event. Logistic regression model works by modeling the relationship between the two entities, one is the independent variable and the other is a specific outcome's probability.

The main concept within logistic regression, is the concept of the logistic function which is also called a sigmoid function. It is used to map the linear combination of the input variables to the binary values of 0 and 1. This value is basically talking about the probability of an event occurring and it lies between 0 and 1. It is this logistic function, due to which it is possible to model even nonlinear relationships that exist between the independent variable and their outcomes.

Using the logistic regression model, we estimate the coefficients that are associated with the independent variable, this indicates

the strength and also the direction of each of their influences on the final outcome. So one of the most common ways to derive these coefficients is to use the maximum likelihood estimation, which basically finds the values that would maximize the probability of observing the given data.



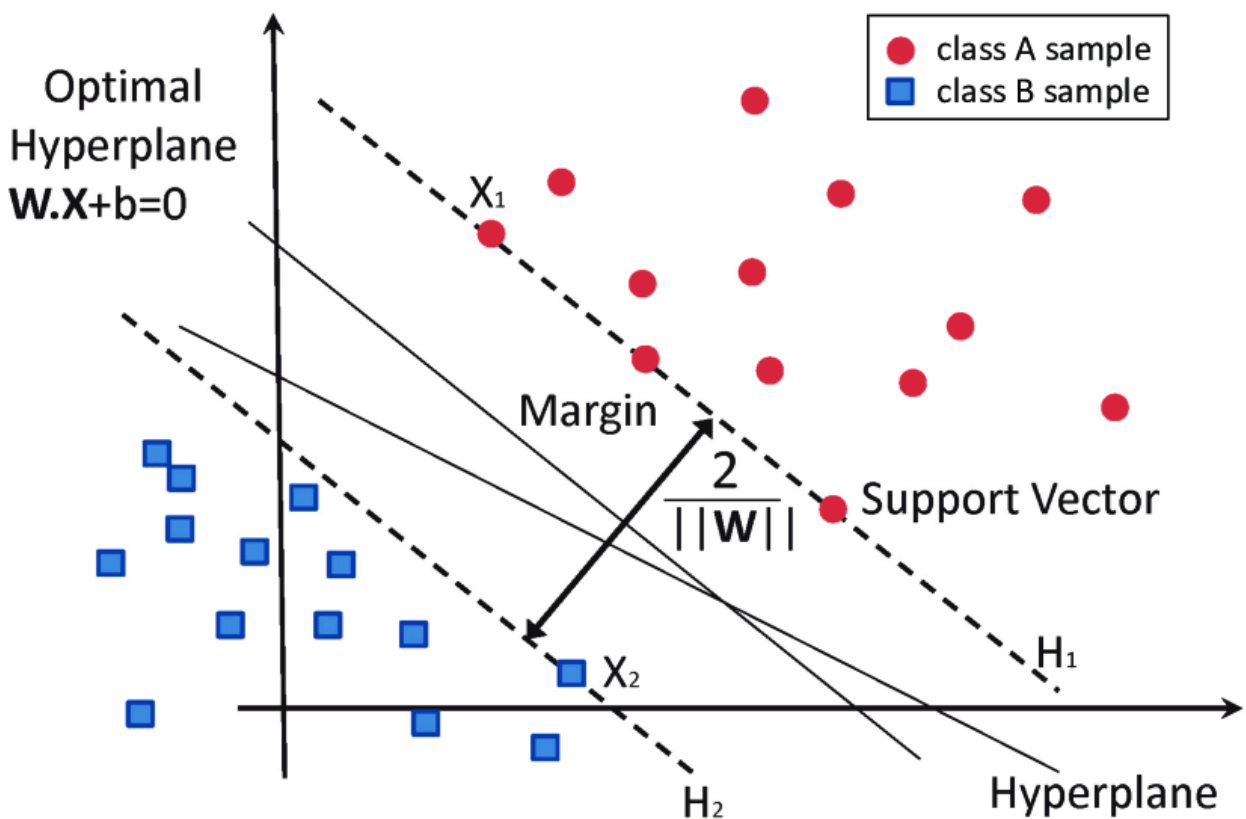
Support Vector Machine:

One of the other powerful and highly used ML algorithms is Support Vector Machine and it is used for both classification related tasks as well as regression tasks. It tries to find an optimal hyperplane that's efficient in separating different classes of data by maximizing the margin that lies between them. The selection of the hyperplane is done in such a way that it accurately classifies the data as well as is efficient in generalizing and managing unseen instances.

Data points in a SVM are represented in the vector format, and

the algorithm aims to look for a hyperplane which maximizes the distance between the closest points from each class, and these are known as support vectors. This approach is called the margin maximization approach and it allows for the SVM to handle complex decision boundaries.

So the ability of being able to handle high dimensional data efficiently when the total number of the features it has exceeds the total number of its samples is one of the major advantages of SVMs. Due to this, it also has less chances of overfitting as rather than just fitting data precisely it focuses more on maximizing the margin. And also, SVMs not only have the ability to handle datasets that are linearly separable but also the ones that are non linearly separable using various kernel tricks. (mine)

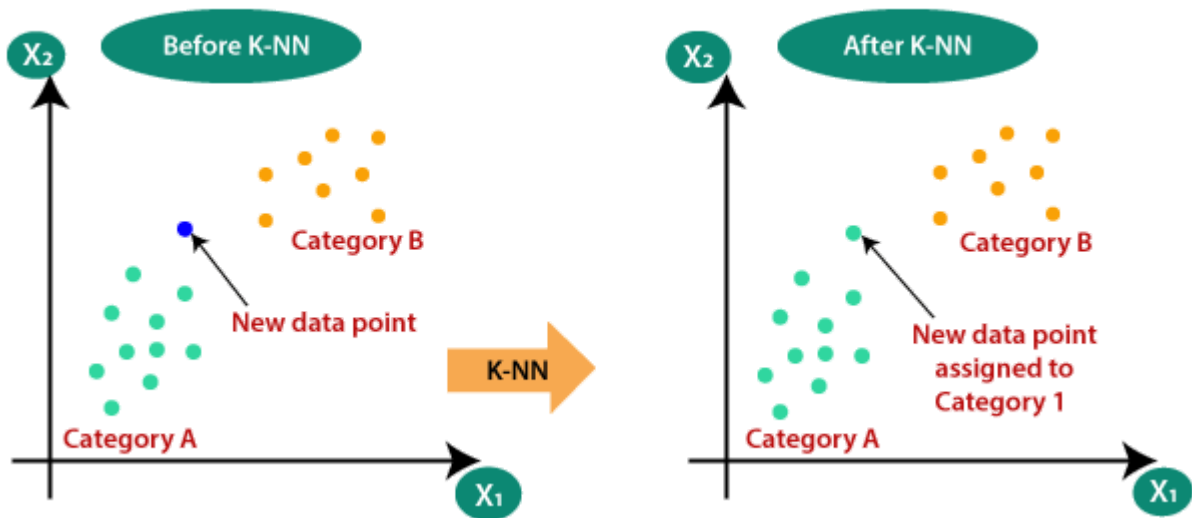


K-Nearest Neighbors:

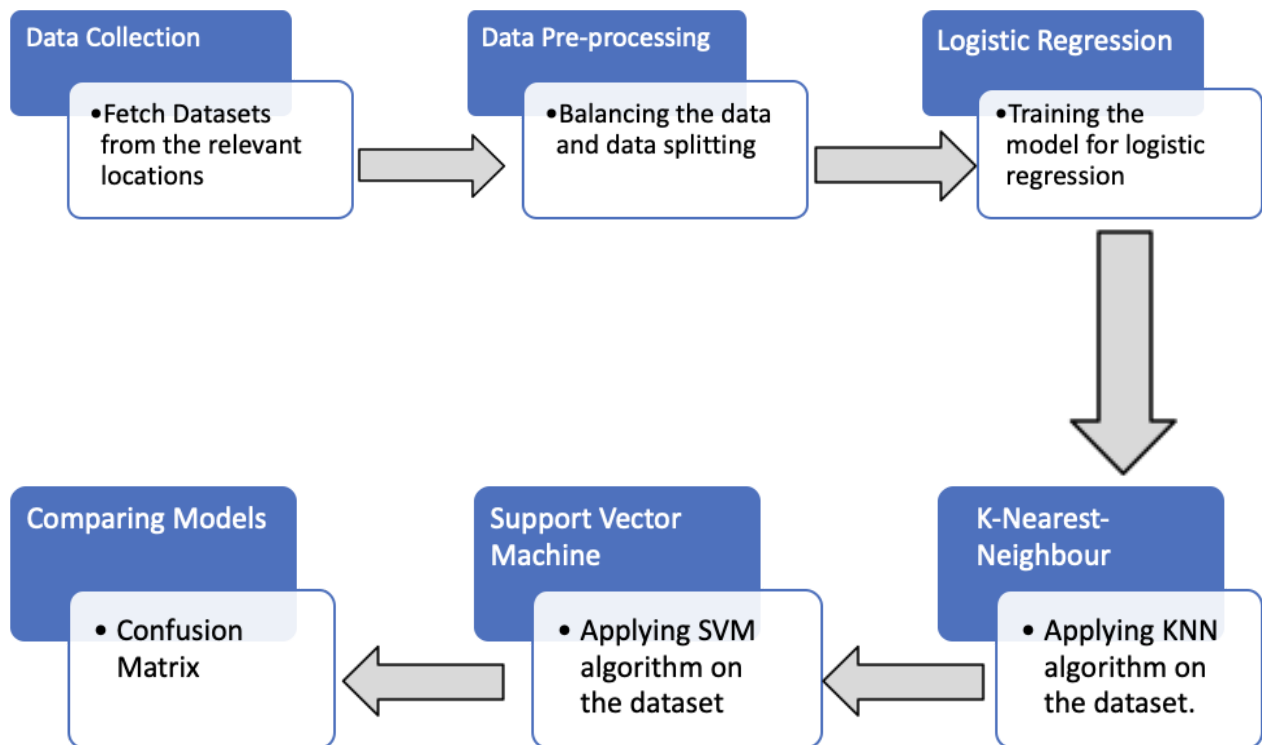
K-Nearest Neighbors is a highly intuitive and commonly used ML algorithm which comes in handy to solve a lot of classification and regression based problems. The underlying idea behind KNN is, how to be able to classify a new data point by drawing a comparison with its K nearest neighbors in a specific feature space. The assumption that it makes in this process is that multiple instances that are similar also tend to have the same label as well.

Value of K is something that will decide the number of neighbors we consider for classification while using the KNN algorithm. So the way the KNN works is that it calculates the distance between the new point that is being considered and all the other training instances to finally classify that new data point. And whatever the majority votes of its K nearest neighbors would be for that new data point would ultimately become the class label for it.

KNN is an algorithm that will not make any assumption for the underlying data distribution. This enables it to be robust even if the data is noisy and it even to data that is noisy and still act in a versatile manner across datasets of all different types. Not only that, KNN is also capable of handling both binary and multi class classification problems, as well as regression related problems.



Approach:



Dataset:

	Time	V1	V2	V3	V4	V5	...	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	...	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	...	0.125895	-0.006983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	...	-0.139097	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	...	-0.221929	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	...	0.502292	0.219422	0.215153	69.99	0

Data Contains :

The dataset that we have worked with contains some crucial parameters that have helped us identify and understand fraudulent activities.

Firstly let's look at TIME, we have the 'Time' parameter, It takes the time in seconds between the current transaction that is being considered and the first transaction that is in the dataset. Knowing this time-based information helps us understand and have insights about the patterns of fraudulent activities and also about its timing.

Secondly, the 'Transaction amount' is a parameter that allows us to understand what the monetary value associated with every transaction is and that inturn helps us identify transaction amounts which could be unusual or abnormal and which may indicate some fraud.

Coming to the 'Class' parameter, it is a very important parameter as it labels different transactions in the manner of either '0' for normal transactions or '1' for fraudulent transactions. Due to this labeling, we have some information that is the ground truth to train our logistic regression model accurately.

Additionally, we also have the 'V1-V28' features, which are nothing but the set of hidden features that capture a lot of underlying patterns and

also some major characteristics of transactions, and this they do to both normal and fraudulent transactions.

Workflow:

1. Data Collection: At the beginning of the machine learning process, it is important to identify the necessary data. Collecting data is a critical step in solving supervised machine learning problems. It is essential to consider different methods of gathering data for training your model. Data collection is a fundamental part of our workflow because we need a substantial amount of patient data to effectively train our machine learning model.

2. Data Exploration: Data exploration is the initial phase of analyzing the data. We use techniques such as data visualization and statistical methods to understand various characteristics of the dataset, such as its size, number of entries, and accuracy. By visually examining the data and identifying relationships between different variables, outliers, data distribution, and dataset structure, we can gain deeper insights into the raw data. Automated data exploration tools can be employed to facilitate this process, using techniques like bar graphs and scatter plots to display the data.

3. Model for linear regression: To compare the performance of other classification models, we will utilize linear regression as the foundational machine learning model. This model employs a binary classification approach to determine whether a person has heart disease or not. It uses the sigmoid function, which takes a real input (x) and predicts the probability of an outcome between 0 and 1. When the predicted probability ($P(x)$) is greater than 50%, the output (Y) is assigned a value of 1. The parameters of the logistic regression model are estimated using the maximum likelihood method.

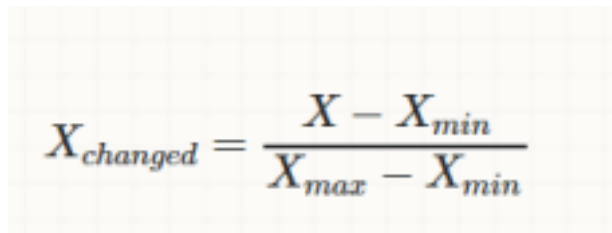
4. K Nearest neighbor: The K-nearest neighbors algorithm,

commonly known as KNN, is a non-parametric supervised learning classifier that predicts or classifies data points based on their proximity to other data points. The distance between data points is calculated using methods like Euclidean, Manhattan, or Minkowski distances. In our KNN classification, we use the Euclidean distance formula. For our project, we have chosen a value of $k = 9$, which determines the number of nearest neighbors taken into account for classification.

5. Support Vector Machine: SVM is a classification algorithm used to separate datasets and find a maximum margin hyperplane (MMH). Support vectors are the data points closest to the hyperplane and are used to define the decision boundary.

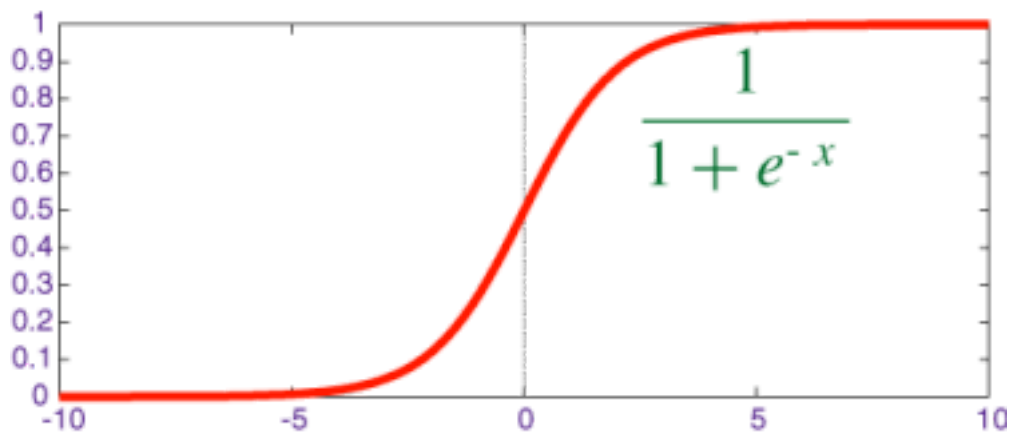
Experiment:

To conduct the comparisons, we selected the logistic regression model as the baseline. We utilized the Sklearn library to create this logistic regression model.


$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Next, our dataset was divided into training and testing data sets, following a 20:80 split ratio. We constructed the sigmoid function for logistic regression. To assess the performance of the model, we examined its test accuracy using tools such as the confusion matrix and Sklearn's classification report, which provides metrics like f1-score, precision, and recall.

In the process, we incorporated concepts such as forward and backward propagation, the sigmoid function, and gradient descent. These ideas helped enhance the logistic regression model's predictive capabilities.



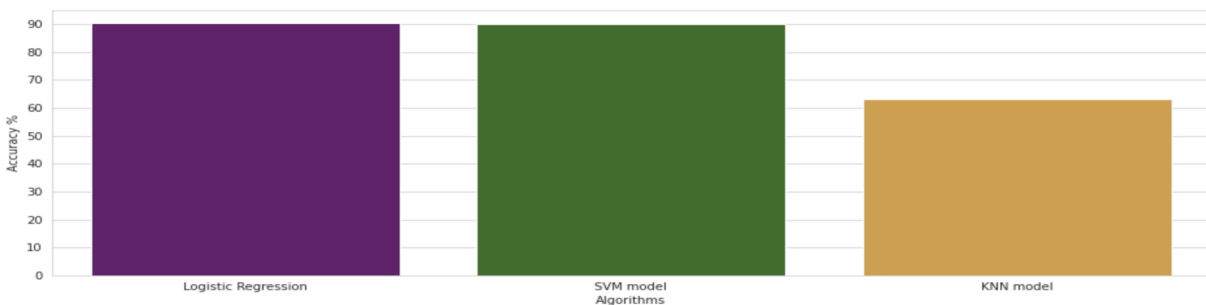
The cost function is

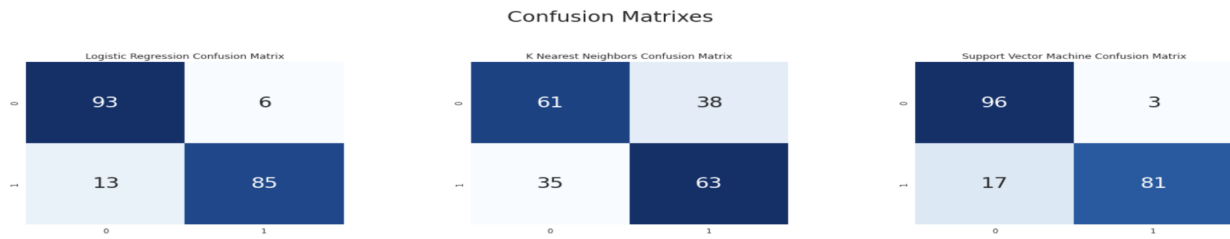
$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))]$$

The test accuracy for Logistic Regression using sklearn function is 90.53%

The accuracies for SVM and KNN were found to be 88.95% and 63.68% on comparison.

Comparing the models:





We are utilizing the confusion matrix as a tool to evaluate the performance of machine learning classification problems with multiple classes. It provides insights into four potential outcomes by comparing predicted and actual values. These outcomes are true positive, true negative, false positive, and false negative. By analyzing the confusion matrix, we determined that the Logistic Regression model achieved the best accuracy in our project.

Conclusion:

After the comparisons it can be concluded that among the three classification algorithms, Logistic Regression stands out as the most effective. It offers optimal accuracy when it comes to predicting the fraudulent nature of a transaction..

References:

Dataset: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

<https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4>

<https://www.javatpoint.com/logistic-regression-in-machine-learning>

<https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/>

<https://medium.com/@kushaldps1996/a-complete-guide-to-support-vector-machines-svms-501e71aec19e>

<https://www.ibm.com/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables.>