**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Advanced Data Analytics  Big Data Storage and Processing |
| **Assessment Title:** | Integrated CA1 Sem 2 MSc in Data Analytics |
| **Lecturer Name:** | David McQuaid  Muhammad Iqbal |
| **Student Full Name:** | Maria Dominguez Alvarenga |
| **Student Number:** | 2019008 |
| **Assessment Due Date:** | 6th October 2023 |
| **Date of Submission:** | 6th October 2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Big Data using Deep Learing for the Credit Card Fraud Detection System applying Artificial Neural Networks and Random Forest classifier.**

Maria Dominguez  
 CCT College Dublin, Ireland  
2019008@student.cct.ie

*Abstract*—*Two high-focus topics in the data science domain are Deep Learning and Bid Data. Since Private and Public companies have been gathering huge amount of specific information regarding issues such as marketing, fraud detection, medical issues among others, the management* *of Big Data has become very important since provides valuable data. Therefore, big data platforms are used nowadays for storing huge amounts of data such as Hadoop and Apache Spark. Considering the previous, since the e-commerce transactions have become more common and helped to increase credit card transactions, which could be either online or standard transactions. Due to this, those fraudulent credit card transactions represent significant losses for many companies worldwide every year. Therefore, to tackle this issue, the development of appropriate and effective credit card fraud detection models has become critical to counteract this raising issue and therefore, minimize those financial losses. However, since most credit card datasets are highly unbalanced as the amount of legitimate transactions are valid, this becomes a challenge. This academic paper proposes the use of deep learning Artificial Neural Network (ANN) and Random Forest (RF) classifiers. The application of ANN model produced better results in the detection of credit card transactions result of fraud in comparison with RF results.*

Keywords—Big Data, Deep Learning, Credit Card, Artificial Neural Network, Machine Learning, Random Forest.

This template, modified in MS Word 2007 and saved as a “Word 97-2003 Document” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) Ease of Use

# Introduction.

The automatization of the extraction of the abstractions from the data is the main concept in Deep learning algorithms [1]. As pointed out by Tang et al. [2] Big data is considered to come in volume, variety, velocity, variety and value. Also, big data has been formed as a result of the appear and accumulation of variety of data such as unstructured, structured and semi unstructured format. Nevertheless, as outlined by Jan et al. [3] different sizes of input data and formats, quick recovery of information, quality of data, data storage, etc are challenges that Big Data Analytics faces for machine learning and data analysis. Similarly, as stated by Gheisari et al. [4] as Big Data are extremely huge data sets that need to be analysed in order to find trends and patterns, is when Deep learning techniques can be used to find useful and abstract patterns within the data. Therefore, as outlined by Jan et al. [3] Big Data Analytics main task is to be able to extract valuable patterns that are used to make decisions and predictions, which comes from a huge amount of data. Similarly, as outlined by Vangumalli [5], it means that deep learning extract valuable data out from big data [3].

As Big Data is being used in the domain of Fraud detection, the research question will be prepared on the topic Credit Card Fraud Detection using the dataset Credit Card fraud detection from Kaggle. [6]. Nowadays people are using credit card more often, either on normal or online transactions, this is the reason why credit card frauds are currently popular. Asha and Kumar [7] mention that credit card frauds can occur in any different organization such as banks, industry, automobile industry among others. Hence, this study is aimed to produce a deep learning model Artificial Neural Network to efficiently predict the fraud in Credit Card transactions, which was going to be compared in performance with Random Forest predictor for a classification problem. Since the dataset is big data, it will be used Hadoop solely for the storing of data, but Apache Spark is going to be used to store and read the data.

# Deep Learning in Data Analytics.

Research conducted by Najafabadi et al [1] mentions that a huge amount of unsupervised data is used by Deep Learning algorithms to extract complex patterns from the data [3] The extraction of features and pattern identification from big volume of unsupervised complex data without the intervention of humans, makes the concept of deep learning a critical tool for Big Data analysis. Artificial Intelligence is the field in which main objective is to emulate the ability from the human brain to learn, analyse, observe, make decisions, mainly to the particularly complex problems. Thus, deep learning algorithms are inspired by Artificial Intelligence [1]. Khan et al [8] outlines that many questions which are associated with Big Data are answered by Deep Learning since those algorithms automatically learn structures and patterns hidden in the raw data through machine learning techniques.

It is shown by Najafabadi [1] that deep learning has produced outstanding results in different machine algorithms and constitute a promising technology when analysing issues related which computer vision, speech recognition and processing of natural language. Chen and Lin mention that [9] successfully application of Deep Learning techniques has been observed in industry products that make the most of the large volume of digital data. Moreover, multinationals companies like Apple, Google and Facebook, who gather and analyse on daily basis enormous amounts of data, are the ones that have been interested in deep learning related project being pushed forward as they are including deep learning in their services and products.

Nevertheless, as outlined by Aru et al. [37] deep learning approaches has shown some problems with theory of Mind (ToM) as the task that have been used to research ToM through deep learning systems have been limited. It has been recommended on the same research to do more research Theory of Mind in environments that are open-ended. The same study also recommended to use the tools from AI interpretability to study the relationship between the aspects of ToM and different networks components.

Likewise, as depicted by Chollet [38] deep learning model has limitations as they do not have any understanding in any human sense about their inputs. Humans have their own way to understand sounds, language, and images, which is supported in their humans’ sensorimotor experience. Hence machine learning models cannot understand in any human relatable way their inputs.

Similarly, on research conducted by Thomson et al. [39] outlines the computational limits of deep learning as this constrain would force deep learning towards a less computationally demanding approaches of improvement, which will push machine learning to be adapted to techniques that are more efficiently in terms of computing power than deep learning.

# Big Data in Data Analytics.

Data that surpasses the typical processing, storage and computing capacity of conventional data analysis techniques and conventional databases are the concepts how Big Data refers [1] As Khan et at el. states [8] the Big data unique features such as varieties, volume, high speed and complexity have challenged the traditional statistical and data mining techniques that were mainly developed for small datasets. However, as pointed out by Azeroual and Fabre [31] big

data it became to be a topic of public debate when it comes to common global needs. As big data consist of a large volume of complex semi-structure, structure and unstructured data which is defined beyond the conventional databased processing capabilities. Hence, the processed of big data nowadays is critical in the forecasting, decision making, product development, business analysis among other domains [31]. Big Data is also facing challenges and problems as companies need to invest in new IT systems. The variety as well as the huge amount of information set challenges for companies. As the same author went on [31], it is crucial the creation of transparency in the data sources, databases and also the variety of data to be able to validate, manage and analyse data effectively.

Nevertheless, as pointed out by Puligadda and Chapala [30] one challenges that Big data Analytics faces is dealing with fast moving and streaming data, which is really useful when dealing with fraud detection when monitoring task. Therefore, deep learning needs to be adapted to be able to handle streaming data. As well as that, there is a need to produce algorithms that are capable of handling continuous input data from large amount of data.

However, big data could be expensive and complex to work with as it needs investment for businesses in storage solutions, cybersecurity, and analytics tools. Hence, there is a lack in data scientists and big data specialists [40]. Also, another key point are the security and privacy concerns [41] as data in high volumes comes with the risk of being vulnerable to data breaches, cyberattacks and security threats. Hence, individuals and small businesses are left at risk of those security breaches as methodologies and tools to secure data could be expensive. Furthermore, the main drawback that faces working with data is dealing with data quality issues [40, 41] as Data scientists needs to ensure that the data that is being used is pertinent, in the right format and accurate before this big data can be analysed. If data quality issues are not addressed by business, it can be discovered that the results from their analytics could be useless.

## Research Question:

The application of Deep Learning Artificial Neural Networks could produce better predictions for the detection of Credit Card Fraud transactions in comparison with traditional machine learning algorithms like Random Forest classifier.

## Problem Statement:

Fraud in Credit Card Transactions has increased due to online purchasing nowadays. Hence, the application of more effective algorithms that help to prevent the fraud by identifying correctly the illegitimate transactions is required specially when dealing with large datasets.

## Research Objectives.

* The main purpose of this academic paper is to analyse the application of deep learning techniques that have been used to predict credit card fraud transactions, specifically Artificial Neural Networks algorithms, which is going to be compared with Random Forest machine learning model to check their performance on Big Data.
* Subsequently, the data is going to be storage in Hadoop, but also stored and read through Apache Spark.

## Research Methodologies:

Case studies, research papers, websites and lectures from Big Storage and Advance Data Analytics modules.

# Literature Review:

Different researchers have provided studies that aim to prevent fraudsters using the credit card information before any transaction was approved. As pointed out by Asha and Kumar case study [7] the implementation of supervised Neural Networks was applied and compared with other machine learning models to measure their performance.

As same author goes on [7] when the stealer uses the other person’s credit card without the authorization of such person and stealing crucial information as password, PIN and other personal credentials, is when the credit card fraud takes place. Owing to this, though the implementation of deep learning is when it can be identified whether the upcoming transaction is genuine or a result of fraud.

Kumar and Iqbal [10] carried out a survey which portraits different techniques applied in the detection of MasterCard fraud through the application of traditional machine learning models such as K-nearest Neighbour, Support Vector Machine, Neural Network among others to measure their performance with the metrics obtained.

However, on a different study conducted by Benchaji et al. [11] a new credit card fraud detection system was proposed based on Long Short-Term Memory (LSTM) networks to be able to predict accurately the fraud detection transactions.

A study carried out by Figuerola [12] points out the issue with the fraud credit card datasets since there are highly unbalanced due to genuine number of transactions recorded in comparison with the fraud ones. As a result of this, the classifier is biased towards the legitimate transactions, which are the popular class which yields a reduced performance for the transactions produced by fraud.

Taken on board the use of imbalanced datasets, a method to identify credit card fraud was proposed in study carried out by Asha and Kumar [7] using Neural Network deep learning, which was compared with Support Vector Machine and K-Nearest neighbour. The study showed that Artificial Neural Network produced an accuracy of 0.9992%, precision of 0.81% and recall of 0.76 %, which were followed by KNN 0.9982% and SVM 0.93% accuracy respectively. Thus, ANN was proved to be a more suitable deep learning technique that tackles the issues produced by the use of imbalance dataset through the pre-processing of the data, normalization, followed by under-sampling of the same. Similarly outlined on different research carried out by Bin Sulaimen et al [13] outlines that ANN produced good performance and therefore the credit card fraud detection using ANN has result to be promising due to its ability to manage large datasets and distributed memory structure.

However, studied performed by Sohony et al. [14] produced that Random Forest and Neural Networks algorithm produced a higher accuracy, when analysing a large dataset of genuine credit card transactions, where an ensemble learning methodology was proposed for credit card fraud detection as the ratio from fraud transactions to standard transactions is bit suitable.

Nevertheless, on another project conducted by Kumar et al. [15] shown that Random Forest supervised machine learning model which yield an accuracy of 90% when detecting credit card fraud through the decision tree for classification that is used by Random Forest classifier.

Conversely, on a different study conducted by Bin et al. [13] states that random forest models are considered to be quite effective when it comes to predicting the class of regression problems, but it has been observed that on credit card fraud detection in real time could establish various limitations. This occurs as a consequence that random forest models perform well on datasets where limited data is available, but in real time scenarios have slower performance. Hence, when predicting credit card fraud detection in real time where a large volume of data is required, the random forest is lacking to effectively trained the dataset and make accurate predictions.

# Hadoop and Apache Stark structure for problem resolution.

A critical factor depicted by Hatua et al. [14] is that the choosing of the appropriate platform to perform the tests is the main step in order to succeed in that experiment. Thus, the amount of that being used along with how to get the best results are the two main concerns to be considered. [15] Cloudlitics outlines that Hadoop platform provides a complete distributed file system designed for storing and managing data across clusters of machines, which has been used since 2006. In contrast, Spark is considered a newer technology. [15] Cloudlitics stated that being both open sources frameworks designed for big data processing, Spark uses resilient distributed datasets (RDDs).

In comparison, Hadoop processing data using MapReduce, where data can be stored across different machines since has a distributed file system (HDSF). [16] Diep states that in terms of performance, Hadoop boots overall performance by accessing the data stored locally on HDFS. Nevertheless, Hadoop can never pair Spark in terms of in-memory processing.

[17] Apache Spark is considered easier to use and scale since its streamlined cluster operation and its ability to leverage multiple coding languages like Scala, python, Java and R languages [19]. In contrast, Hadoop essentially relies on expertise in MapReduce and Java. Adding to the fact that constant monitoring and maintenance is required for Hadoop due to its nodded architecture. As cited by Guan et al. [18] Hadoop is enjoying a vast market in medical treatment, e-commerce, finance as its great computing power, impressive compatibility as more data is stored at present on the cloud platform based on the Hadoop architecture.

As Hatua, [14] Hadoop platform using MapReduce was applied on study about Early detection of diabetic retinopathy from big data, which provided better results when processing that large dataset in comparison with some of the close competitive state of art techniques analysed in that study.

[19] Parmar outlines that Apache Spark was developed to overcome the limitations of Hadoop and introduced the term Resilient Distributed Dataset [RDD] and process the entire data into the memory RAM, which makes Spark 100 faster than Hadoop. That were the reason why Apache Spark became very popular as is very fast and could handle a lot of data.

As Iqbal points out [20] there are two ways of processing in Hadoop, one is MapRedude and second is call pySpark. As stated on cloudduggu site [36] Spark is based on two important concepts which are Resilient Distributed Dataset(RDD) and Directed Acyclic Graph (DAG).

[20] When using PySpark there is the creation of RDD two operations: actions and transformation, which is getting some values based on some values. Spark has RDD properties: immutability which means that data cannot be deleted, it cannot be updated, lineage is how to recover the missing information if happened. However, Apache Spark is most robust than Hadoop is because is doing all processing in RAM and there is not any hard drive disk involve [20]

As outlined by Azeroual and Fabre [21] the Hadoop three layers components are the Hadoop Distributed File System (HDFS), MapReduce. The third component is the Yet Another Resource Negotiator (YARN) which main function is to abstracts from MapReduce. HDFS is a Java-based distributed file system that allows reliable data storage. Nevertheless, as outlined by Iqbal [20] the Spark programming concept is more versatile and useful nowadays in the industry. As same author went on, that is the reason why Spark is very powerful as exposes a uniform programming model supporting the application development in multiple programming languages as people with expertise in these different languages can do the data processing either in Java, Python or R languages.

Therefore, when working with pySpark, the Hadoop yarn library is the one to control all the processing. In other words, it is managing all our network and all the communication [20]. As Iqbal went on [20] when using local drive, it cannot be exploited all the resources for the Hadoop, but if HDFS is being used, it is possible to use all the resources for Hadoop for read. Hence, Hadoop yarn is being used to process the clusters in a better way in the Apache Spark. Overall, Hadoop is better processing the Apache Spark as its architecture is more flexible [20 ].

Nevertheless, on a different approach stated by Madhugir [ 32] states that Apache Spark in terms of security is still on early stages in comparison with MapReduce, which has more projects and security features. On one hand, MapReduce requires little RAM capacity, but it needs a significant number of devices with a higher volume of hard disk space. On the other hand, Spark requires higher capacity in RAM, but it needs standard disk capacity in fewer devices. Thus, since hard disk capacity is less expensive, than RAM, MapReduce is a less expensive than Apache Spark.

# Hadoop and Apache Spark for data storage – Credit Card Fraud detection.

The approach chosen for the study conducted was based on Spark performance in contrast with MapReduce one. As listed previously above Spark exceeds in processing speed to MapReduce, as well as it performs excellent the iterative jobs. In contrast, MapReduce is dependent of Hadoop, whereas Apache Spark is not. Additionally, the machine Learning applications are excellent in Spark, whereas the ones for MapReduce are considered to be as an average [20]. Research carried out by Santosh and Ramesh [33] shows that a Spark-based integrated methodology produces good results when using bid data technologies to store a large volume of Credit Card Fraud detection data.

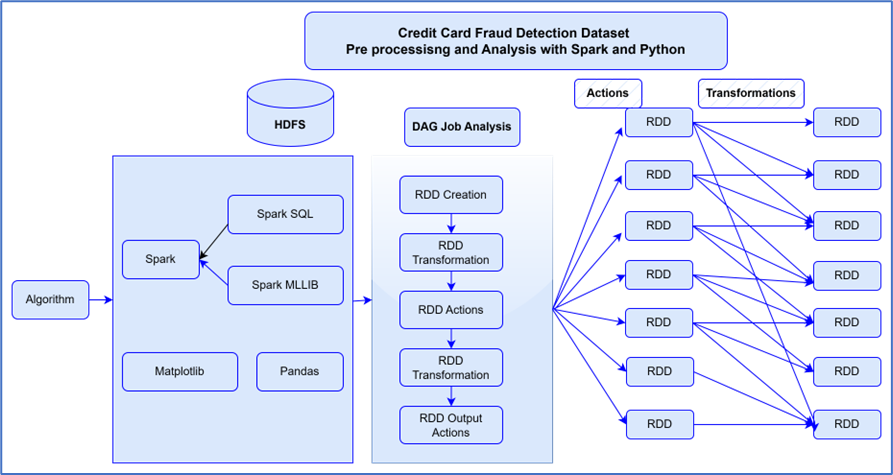


Fig. 1. Pre processing and Analysis with Spark and Python phases – Credit Card Fraud Detection Dataset

As shown in the above diagram, the dataset is going to be stored in Apache Spark and Hadoop. However, Hadoop is only to be used for dataset storage only. PySpark is going to be used to store and read the data using the Libraries of Machine learning for reading datasets. Matplotlib and Pandas libraries are going to be used for that purpose.

As image above, the credit card dataset was chosen for the purpose of this research paper. Hadoop yarn is going to be used to store the data in Hadoop and then optimize the processing of the data in Apache Spark. It is intended to use Spark SQl to write some queries and also Spark MLLIB to do some classification on my dataset using Numpy library from spark. Then DAG directed a cycle graph, then the RDD will be created for my dataset chosen, followed by the transformation and actions. As shown above, Apache Spark provides the RDD, which supports the distributed processing of a cluster. In the event of failure of those nodes during the processing of the dataset, the spark framework will be able to handle this failure by using the RDD properties of immutability [20]

* The exercise was performed to store the “creditcard\_2023” dataset into Hadoop as follows:
* Dataset “creditcard\_2023” was downloaded from MyDrive into “Download folders”.
* Start-sdh.sh & start-yarn.sh in Hadoop.
* New user creation: a new user “user04” was created by typing Hadoop -mkdir /user4
* Checked that used 4 was stored in Hadoop by typing: Hadoop fs -ls /
* Store dataset in Hadoop: Hadoop fs –put ./creditcard\_2023.csv /user4

Dataset has been only stored in Hadoop, but it is going to worked in Spark by typing “pypark” to open a jupyter notebook environment (figure 2, 3, 4, 5 and 6) Therefore, Apache Spark is going to be used to store and read the dataset using Spark MLLIB and Spark Sql. Those terminals mentioned above are going to be shown on the video attached.

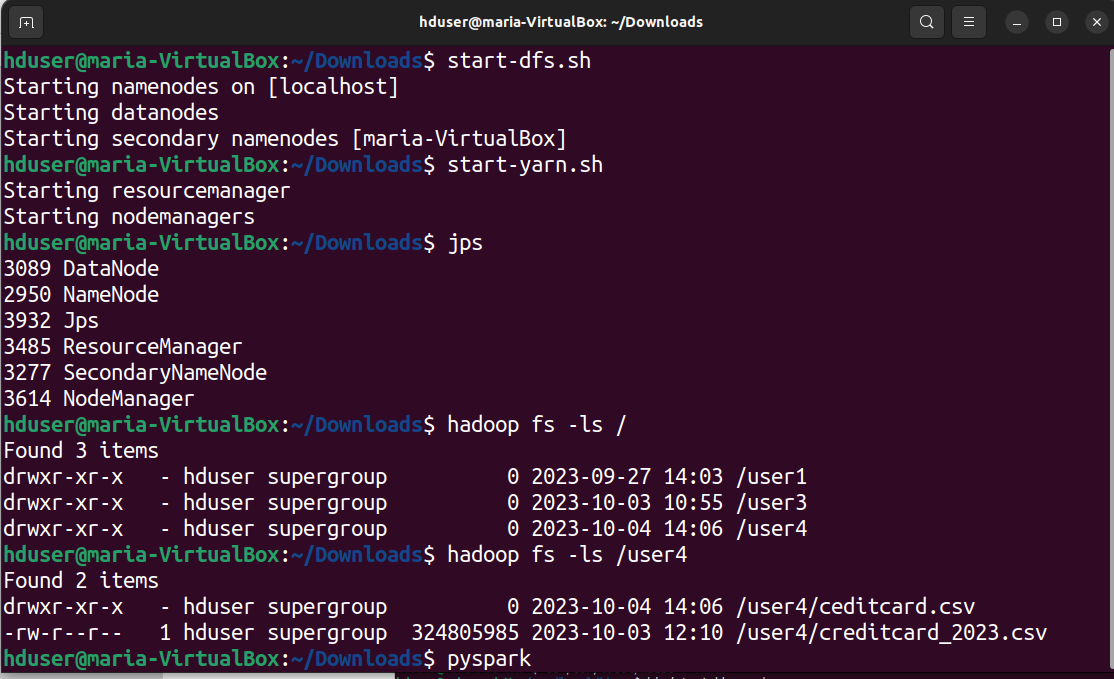


Fig. 2. Storage of dataset in Hadoop

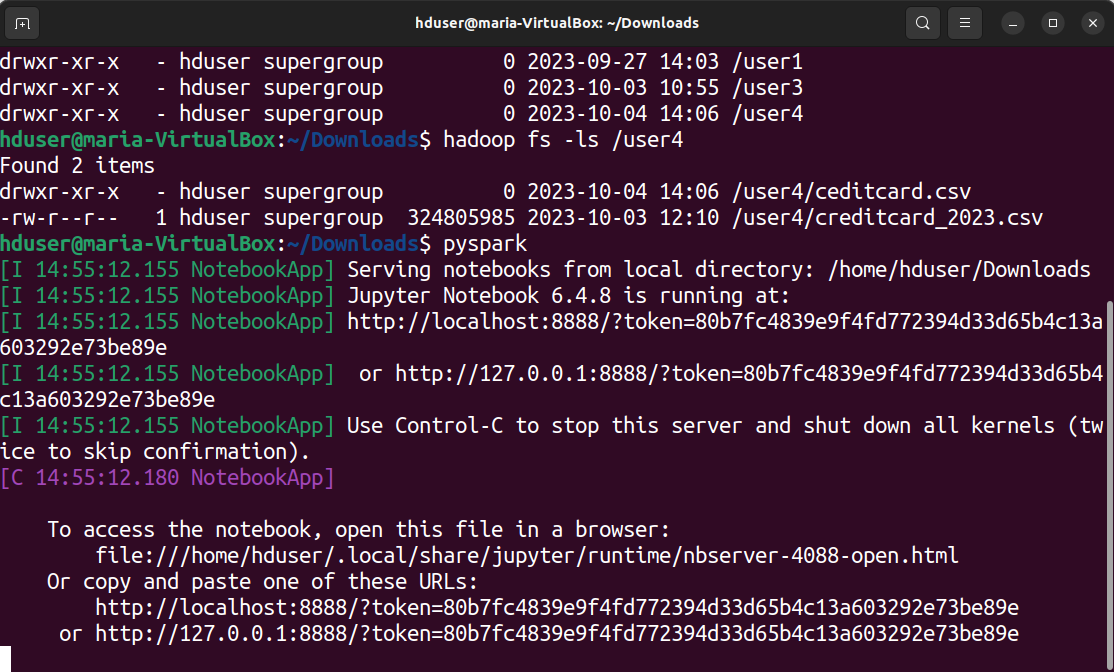


Fig. 3. Open PySpark

Jupyter notebook is open where it can be seen the dataset “creditcard\_2023.csv and the Jupiter notebook “Credit Card Fraud Detection”.

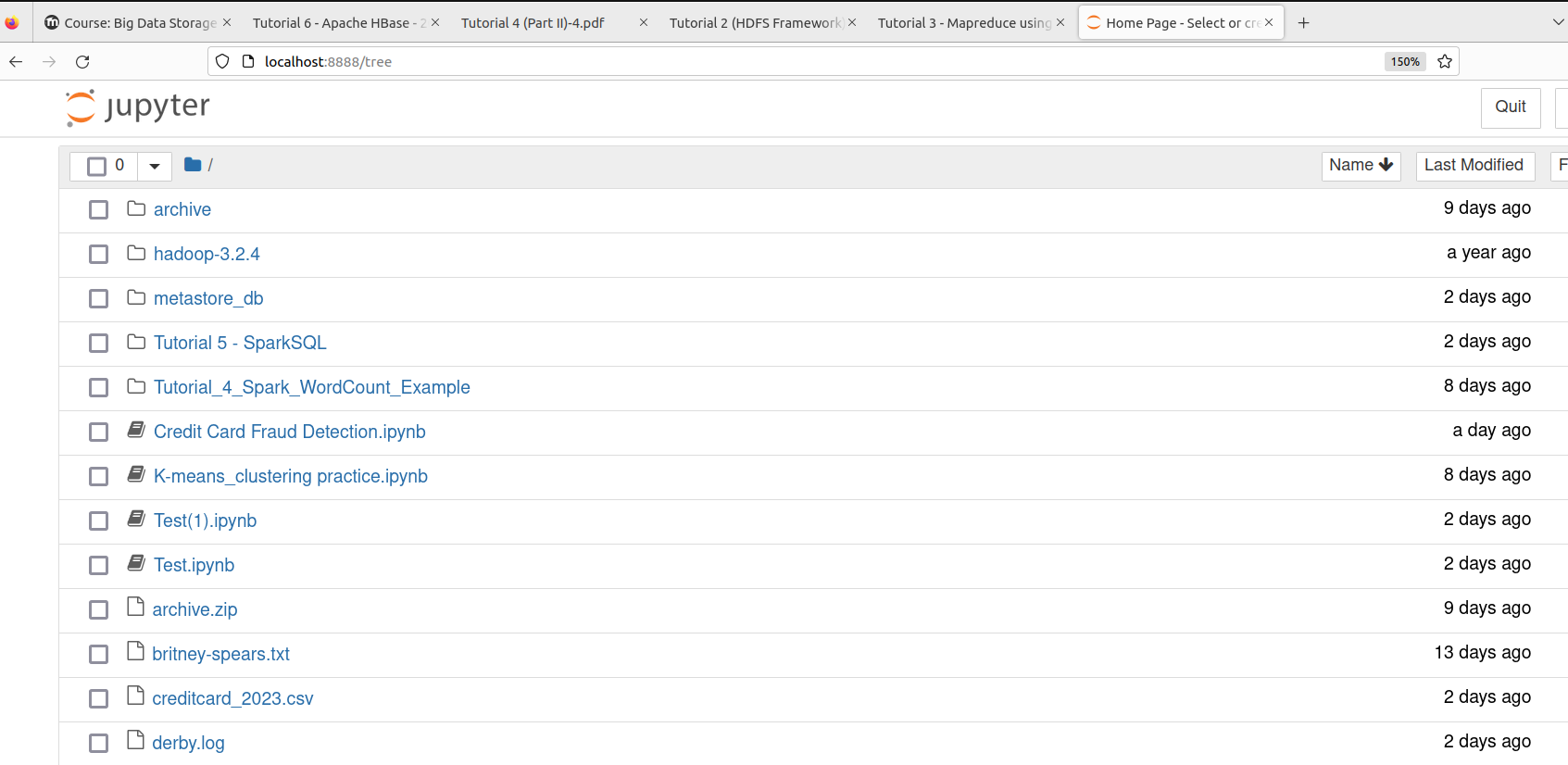


Fig. 4. Jupyter notebook for Credi Card Detection is shown

As seen below, the dataset is being read in Spark:

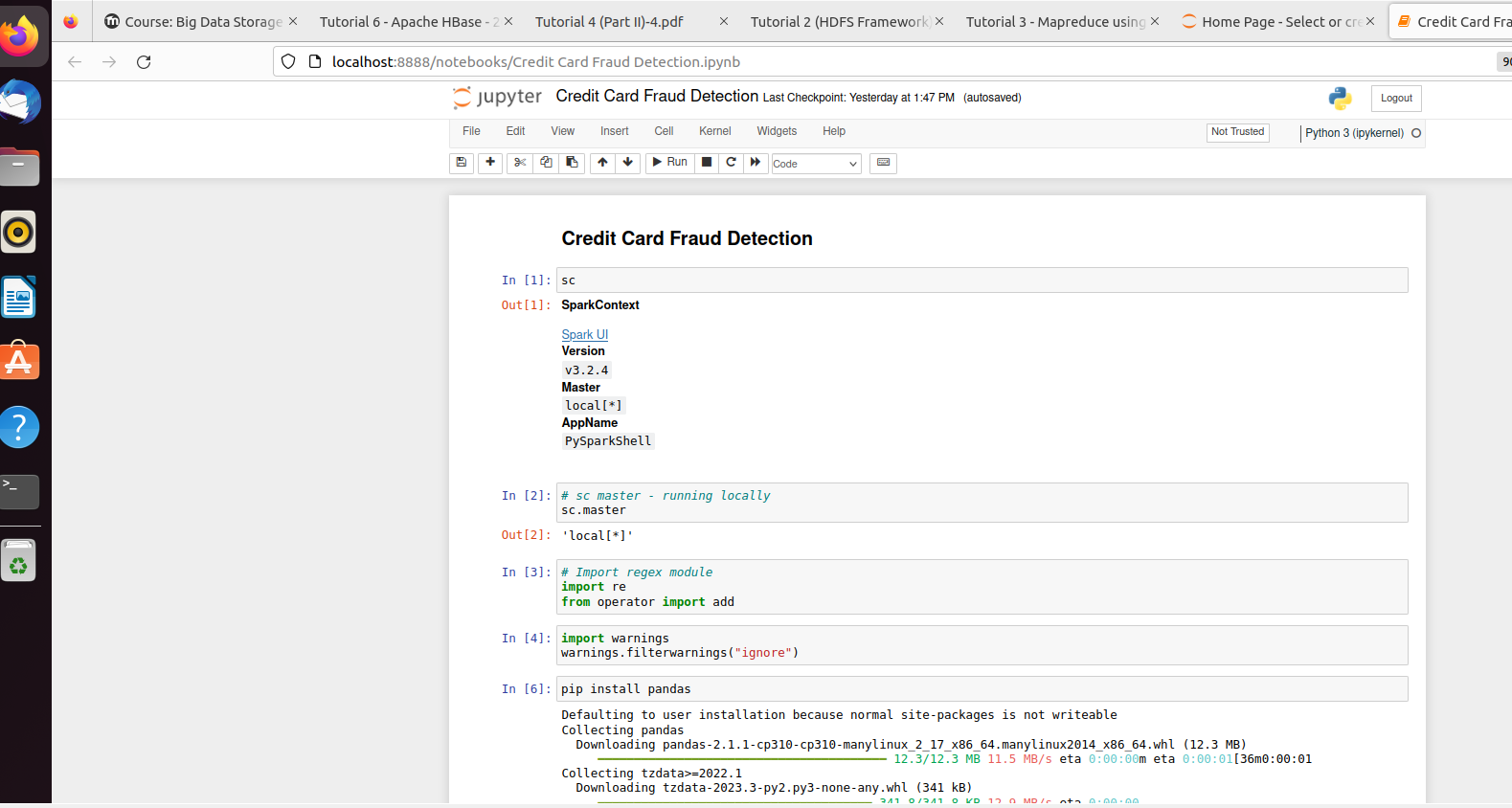


Fig. 5. Credit Card Fraud Detection Jupyter notebook

Some columns and correlation between variables has been read.

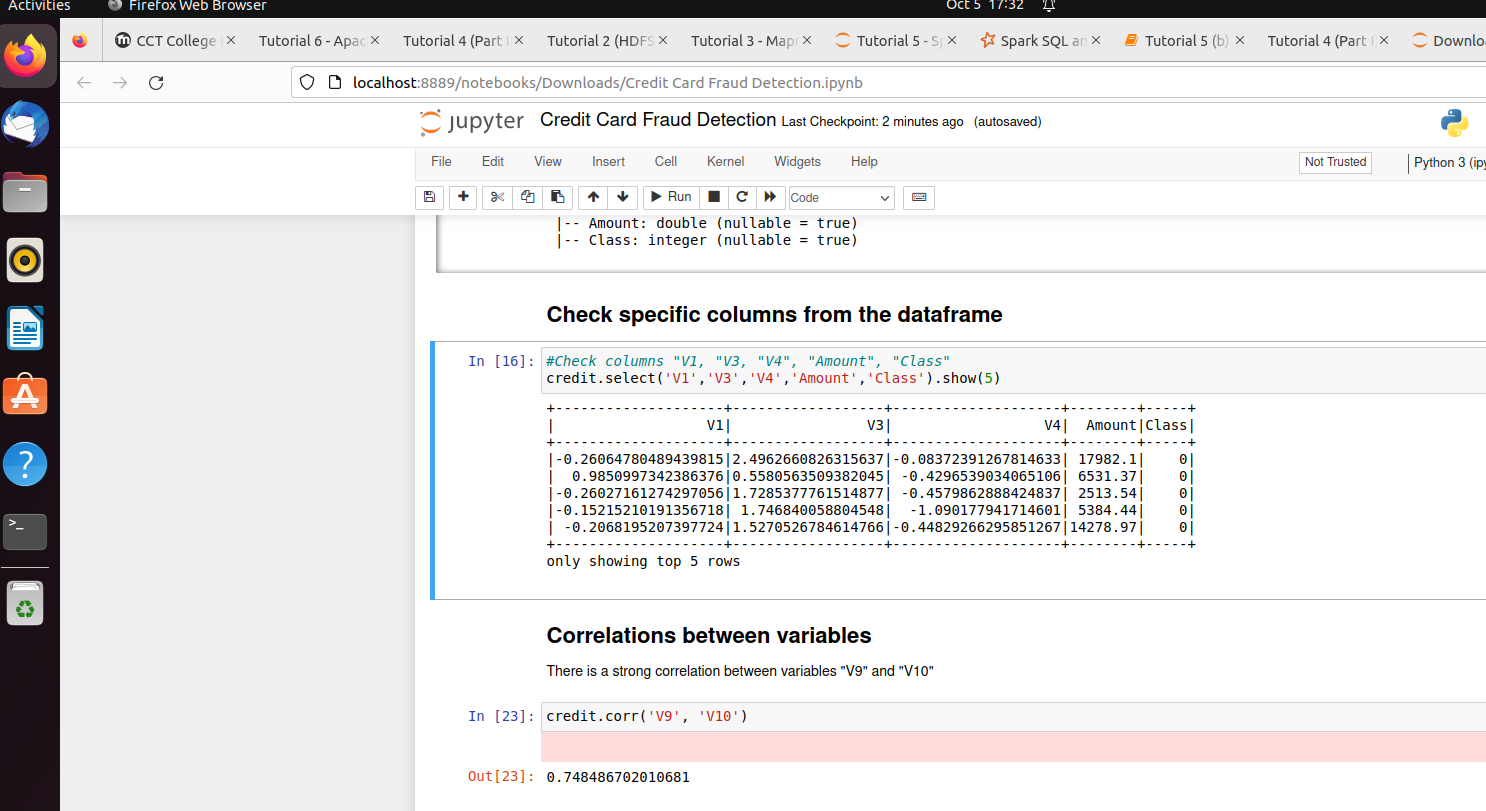


Fig. 6. Checking some variables and correlation of a variable in Jupyter notebook

# Deep Learning technique for solving the stated problem.

Jan et a. states that [3] through Neural Networks large volume can be processed in an efficient and faster way. However, as pointed out by McQuaid [22] there are 3 different neural network used in pre-trained models which are 1) Artificial Neural Networks (ANN), which can be used for classification and regression problems. However, ANN is unable to capture sequential information such as time series, audio, images among others in the input data. 2) Convolution Neural Networks, (CNN) which are used for computer vision problems, that through their kernel or filters are able to capture the most relevant features from the input data. CNN performs well in sequential information inputs. 3) Recurrent Neural Networks (RNN) which are used for time series problems. It is composition help that each hidden layers helps to train the next observation hidden layer in the dataset. In other words, the hidden layer not only generates an output, but the same output is stored in the same layer as input.

Ghofrani et al [23] outlines that Convolution Neural Networks and Big data have attracted the attention of researchers of Artificial Neural Networks techniques as software systems are developed on the basis of training a model of neural networks on a large dataset instead of writing complicated algorithms. However, ANN has been not able to resolve every challenge of Software engineering domain despite the fact of producing satisfying functionality and outstanding performance in classification and prediction of future events for industrial systems context.

Nevertheless, all neural networks experience the vanishing and exploding gradient, RNN in particular as they have a large number of time steps [22]. This happens when the gradient becomes too small or too big during the backpropagation process. When the gradient is too small causes the vanishing issue and gradient are too big causes the exploding gradient and creates an unstable model. As stated by McQuaid [22] exploding gradient issue can be solved by using truncated backpropagation, which limits the time steps numbers and before the exploding gradient occurs, the algorithm will stop. As well as that, penalties and gradient clipping could be used, this last one limits in term of how big the gradient become during the backpropagation algorithm by introducing an artificial ceiling. Weight initialization and echo state networks can be used for solving the vanishing gradient problem, but Short-Term Memory Networks (LSTMs) is the most significant solution to this problem [22]

Research conducted by Esenogho et al [24] stated that Long Short-Term Memory Neural Network [LSTM] has been proved to achieved excellent results in learning long-terms dependencies, which is a special type of RNN. This LSTM also avoids the vanishing and exploding gradient issue as were developed to overcome those issues. In contrast with ANN, LSTM presents feedback connections between the hidden units, which are connected with discrete time steps. Those feedback connections allow the long term sequencies to learn and to be able to predict a transaction label through the sequence of previous transactions.

A*. Limitations of Neural Networks.*

As stated by Thuwarakesh [25] ANN as they are composed by many intercommoned nodes, which are adjusted through the backpropagation process, requires a lot of computational power due to many parameters. As portraited by Urwin [26] The size of the data plays an important role in the amount of computational power required by the neural network as well as the complexity and depth of the network. As portraited by the same author [26] a Neural Network composed for one layer with 50 neurons is going to be faster that a random forest that contains 1,000 trees. Similarly, a random forest with 10 trees will be faster than the neural network with 50 layers.

Additionally, more extensive datasets for training are required for ANN. Also, Neural Networks are not easy to explain in comparison with traditional machine learning models, since they are more complex [25]. In addition to that, neural networks require a lot of training data which is the cost of being very flexible and learn to recognize the input patterns within the data. Therefore, they have the tendency to overfit in small datasets as they do not generalize well to new data as they memorize the training data. That is the reason why traditional machine learning models would perform better on smaller datasets rather than deep learning algorithms.

As explained in case of Boulle et al. [35] about using ANN to classify the cause of death from verbal autopsies, and the study was used to compare the performance of ANN and logistic regression model which demonstrated that ANN model performed better as predicted the CSMF as in 8 of 16 causes of death within 10% of true value. The same study concluded that cross-validation was critical in preventing the over fitting of the ANN model in terms of the training data as ANN models required large training data sets, which helps to improve their performance.

1. Proposed Approach – Artificial Neural Networks model.

The approach proposed by this study is to use deep learning applying Artificial Neural Network to identify fraud credit card transactions. As it was referenced above, Neural Networks have been proved to perform better on big data rather than conventional machine learning models as they require more training data. Artificial Neural Network is very good at fitting problem as is deep learning model system inspired by the human nervous system [27]. It is basically composed by three layers, input, hidden and an output layer. Its inputs are only processed in a forward direction [22].

Hence, it is going to be used the dataset “creditcard\_2023” from Kaggle, which is considered to be Big data with 309 MB, which contains numeric, continuous data. It has 568,630 observations and 31 features. As part of this experiment, it is going to be prepared a Random Forest classifier to compare both performances after processing the big data.

## Model Building

This study is based on X which are the independent variables and a target value “Class” of being either as a result of legitimate or not transaction.

It is proposed a multilayer neural network, which is processed under supervised learning. Each layer will be full connected as the units in the previous layer are connected to all the neurons in the next layer. The layers are going to be defined that way as more layers would allow the network to learn more complex relationships, but that would mean increasing the computational cost [28].

The ANN prepared will be composed by three layers, one input layer, one hidden layer and output layer. As stated by Khaldi et al [27] more hidden layers could provide better learning results but could also add complexity to the model and could solely surge the learning process time. The number of neurons were stablished by trial and error.

The ANN nodes are activated once they reached the threshold values 𝜃𝑖 , then a h unknow function is found as below [27 ]:

yj = h(x1, x2, … , xn)

1. EDA Analysis.

## Version control link:

<https://github.com/mariadominguez2023/ADA_BD_CA1.git>

The dataset is going to be load in the accompanied jupyter notebook. The Exploratory data Analysis was performed in the following sections:

Libraries used for analysis.

* Matplolib for visualization of graphs
* Seaborn for visualization of graphs
* Pandas for creation of data frames
* Numpy for calculation of numerical functions.
* Keras libraries to import models
* Keras library to import Sequential

## Data Cleaning.

There is no presence of null values, there has been one duplicated row which has been removed.



Fig. 7. Bar Plot of Class Distribution

As seen above, there is a balance distribution between the data coming from genuine and fraud transactions.

## Correlation of Variables

When checking the correlation of variables, there was identified a strong correlation of 0.77 between V16 and V18. There is also a strong correlation V4 and V11 features. Additionally, there was a moderate correlation of 0.59 between V5 and V7. There was a week correlation of 0.25 between V6 and V5. Also, there was a negative correlation of -0.22 between V17 and V20. As shown below there was a strong correlation of 0.77 between V9 and V10 variables.

Scatterplot was used to provide a visualization of the relationship of these two features [43] 2023). The default color was blue, histogram and box plot diagrams.

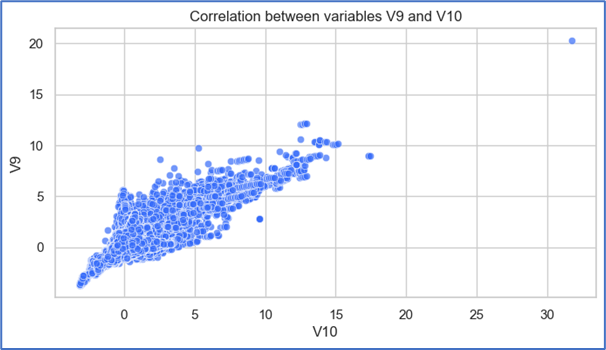


Fig. 8. Bar Strong correlation between variables

## Shape of the data.

An insight into the distribution of the data is needed to make decisions about data preprocessing, this will be obtained through the Histogram visualization of variables [43] As seen above the shape of the data follow a Gausian distribution, which is taken into consideration for the data preprocessing.

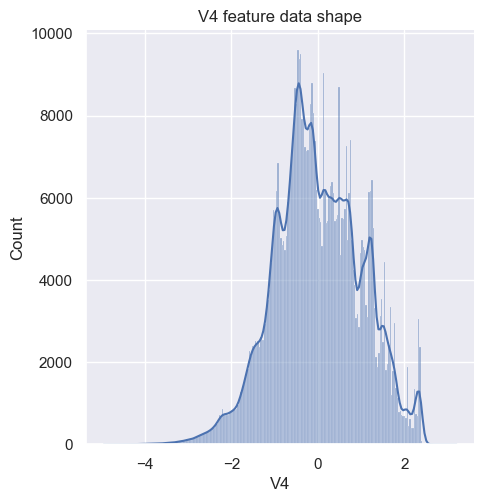
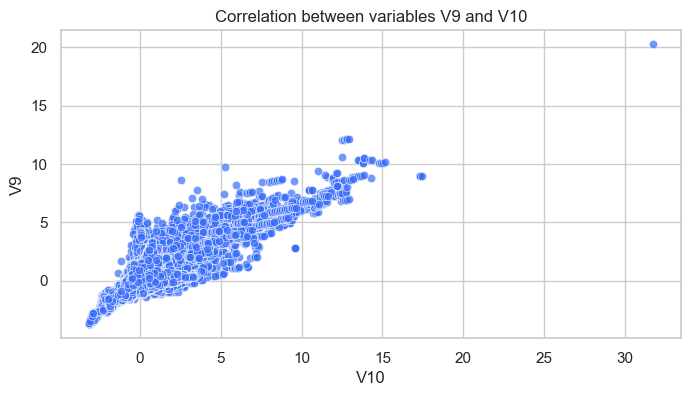
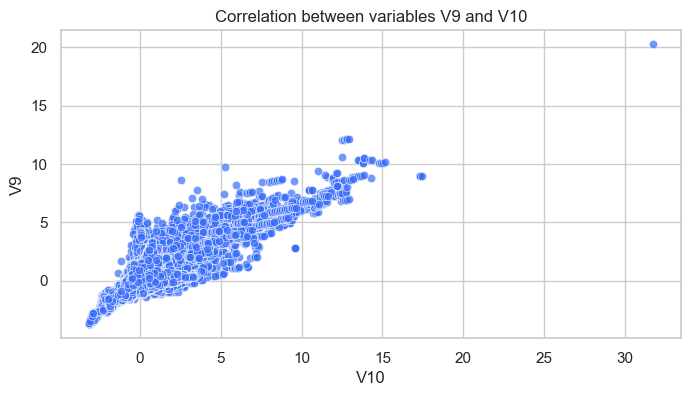
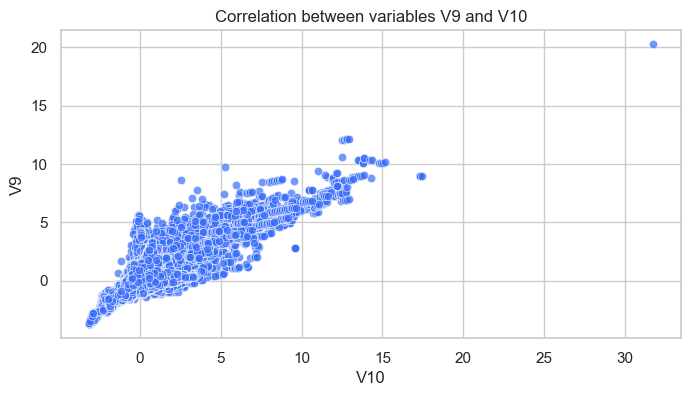


Fig. 9. Shape of feature V4

## Outliers Detection.

As seen below, extreme values were detected on the dataset used for this study. Box plots were used to get the visualization of those extreme values [43]

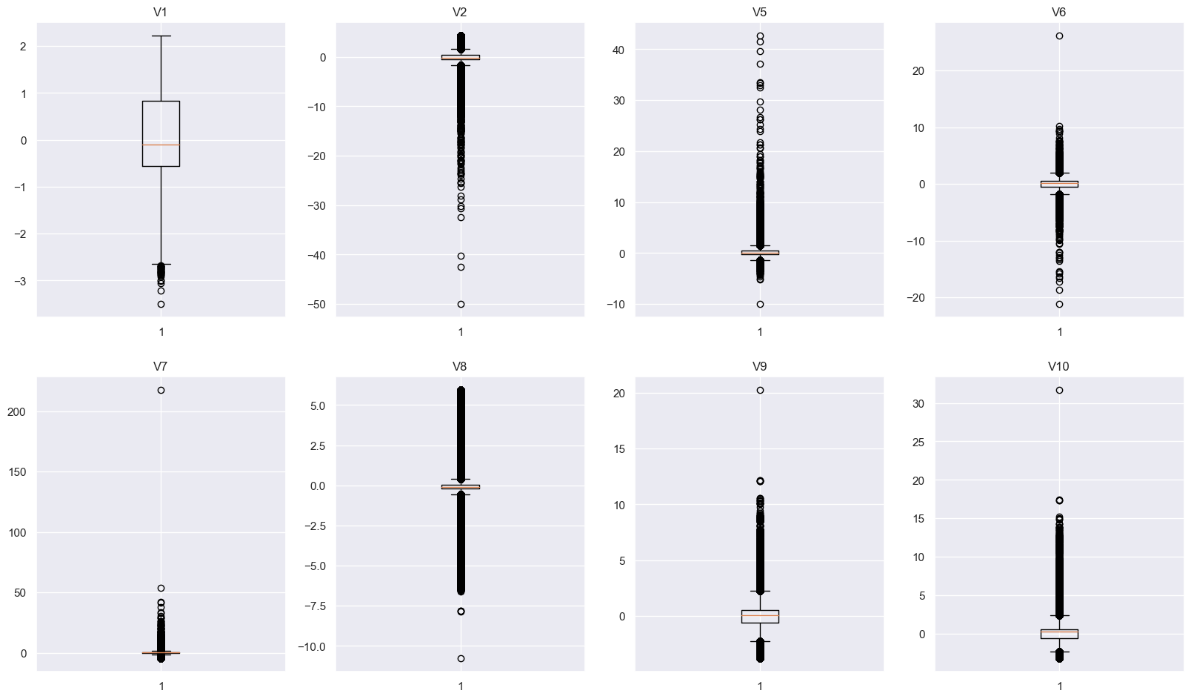


Fig. 10. Outliers on dataset

## Data Preprocessing.

When checking the shape of the data, it is observed that data follows a Gaussian distribution, it also has outliners shown above. The extreme values are not going to be removed from the dataset when performing this experiment. Therefore, the X train and X test is going to be standardized to get a scaled features that has a mean of 0 and a variance of 1 following the formula as below [34]:

x˜=x–mean(x)

sqrt(var(x))

Data preprocessing is very important for ANN to read better, which includes data cleaning and standardization of the data. Afterwards, the dataset has been divided into test data of 0.20%, training data of 80%, random\_state =1 and strastify=y. Models are going to be trained and tested to predict whether a transaction as not fraud or fraud.

## Verifying that shape after Standardization of X train and X test sets.

After data pre- processing technique, it is important to verify that the shape of the data has not changed. Hence, histograms were plotted to make sure that the shape of the data has not changed since these types of visualizations provided a better understanding of the shape distribution of the data [43]. It was confirmed that the shape of the data remains the same.

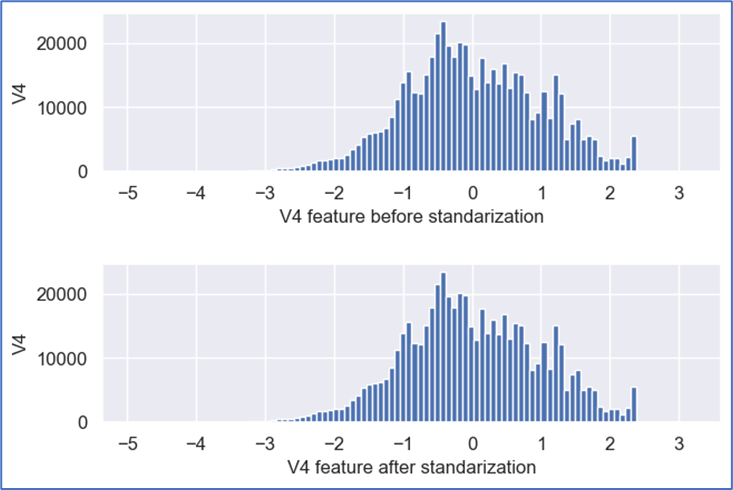


Fig. 11. Shape of data before and after feature scaling V4

1. Preparation of ANN model

Neural Network was built in Keras by using Sequential ( ) function. I set one input layer with 10 units, 17 feature values as it tells the first layer to expect each observation to have 17 values [29] and activation function “relu”. One hidden layer with 2 units, activation function “relu”. The output layer will have 1 unit with activation function of “Sigmoid” which is used in Binary classification where the target values is intended between 0 and 1. [28].

Subsequently, the model will be compiled by using an optimizer, which main purpose is to change all the values inside the neural network and the loss function. The optimizer chosen will be “Adam”. Research conducted by Kingma et al. [42] outlines that Adam optimizer has been found computationally efficient, which requires little memory requirement. This optimizer is going to backtrack to the entire neural network changing all the set values for the weight and biases [28]. Also, it is going to be chosen the loss function, which demonstrates that I have trained the model in the best possible way. If the loss is reduced; the ANN is better at making predictions. The loss function measures how well the predicted values matches the true values [28]

Thus, the loss function chosen is “Binary Cross entropy” for a binary classification problem [28] and the performance metrics chosen is “accuracy. The model will be fitted with 25 epochs, which defines the epoch parameters to use when training the data. Verbose will be ‘0’, which defines how much information during the training process is outputted. In this exercise, will be ‘0’ output. The batch size stablished is 100 which set the number of observations through the neural network before the parameters are updated [29].

## Results

The performance of both models can be summarized as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **F1\_score** | **Precision Score** | **Recall**  **score** |
| ANN | 0.97 | 0.9724 | 0.979 | 0.9651 |
| RF | 0.99 | 0.99 | 0.99 | 0.99 |

Fig. 12. Performance results for both models

## ANN Training and test set loss results.

As shown below (figure 5) the training loss was noted to be decreasing until get the lowest number at epoch 25. However, the test loss did not produce the expected results.



Fig. 13. Training and test set loss results

## ANN Training and Test set Accuracy results

Similar results were obtained on the accuracy training test which increased until it reached 0.97, but test set yield a different pattern.

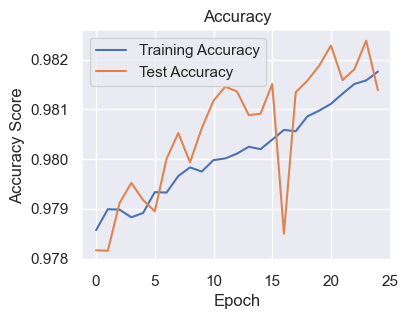


Fig. 14. Training and test set Accuracy results

As mentioned above, the number of neurons were determined by trial and error. The first time this model was tested, it had 10 units in the input layer, but 5 neurons in the hidden layer. However, the model was fitted with 6 epochs, batch size=100, but model was overfitted, with a loss of 0.0173 in the last epoch and yield an accuracy of 0.99% as below:

1. Further study

As stated above, a number of research have been conducted in regards with credit card fraud detection. However, there is the need for more efficient systems that can identify transaction fraud before this takes place.

1. Conclusion:

Through the research conducted above of deep learning and big data, it was possible to learn the functionality of deep learning models such as ANN, CNN and RNN as well as their limitations when using large datasets. It was also possible to learn about the disadvantages when using big data such as security concerns, data quality and expensive and complex to work with. As well as that, after performing the experiment on the dataset chosen referenced above, which was tabular data, using one input layer, one hidden layer and one output layer. The units in the layers were chosen by trial and error, it was possible to compare the performance of ANN and Random Forest algorithms, both for classification, using a large dataset. As found in previous studies referenced above, Random Forest lacked to effectively train the large datasets and make accurate predictions. Hence, RF was trained with the same parameters used for ANN model and it was found that RF as a standard machine learning algorithm did not perform well on big datasets. On the contrary, Artificial Neural Network yields better results. Nevertheless, after plotting the loss and accuracy in the training and test sets, the loss shown a better performance in the training test as it was shown a reduction when the epochs were decreasing. Conversely, the loss and accuracy in the test set did not produce better results. Hence, it can be concluded that ANN was a possible way to conduct this experiment, but it could have also been tested with other deep learning models such as CNN or RNN using Long Short-Term Memory (LSTM) to produce better results.

1. References

[1] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. Journal of big data. 2015 Dec;2(1):1-21.

[2] Tang, L., Li, J., Du, H., Li, L., Wu, J. and Wang, S., 2022. Big data in forecasting research: a literature review. Big Data Research, 27, p.100289.

[3] Jan, B., Farman, H., Khan, M., Imran, M., Islam, I.U., Ahmad, A., Ali, S. and Jeon, G., 2019. Deep learning in big data analytics: a comparative study. Computers & Electrical Engineering, 75, pp.275-287.

[4] Gheisari, M., Wang, G. and Bhuiyan, M.Z.A., 2017, July. A survey on deep learning in big data. In 2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC) (Vol. 2, pp. 173-180). IEEE.

[5] Vangumalli, D., (2017). Applications of Deep Learning (CNN) in Big Data Analytics. [online] Available at: <https://medium.com/@dinnuv/applications-of-deep-learning-in-big-data-analytics-c0a8d8b46c51> [Accessed 12 September 2023]

[6] Kaggle datasets (2023). Credit Card Fraud Detection Dataset 2023. [online] Available at: <https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023> [Accessed 01 October 2023]

[7] Asha, R.B. and KR, S.K., 2021. Credit card fraud detection using artificial neural network. Global Transitions Proceedings, 2(1), pp.35-41.

[8] Khan, M., Jan, B., Farman, H., Talha, M., Ali, S., Shah, S., Khan, F.G. and Iqbal, J., 2019. Integration of big data and deep learning. Deep Learning: Convergence to Big Data Analytics, pp.43-52.

[9] Chen, X.W. and Lin, X., 2014. Big data deep learning: challenges and perspectives. IEEE access, 2, pp.514-525.

[10] Kumar, P. and Iqbal, F., 2019, April. Credit card fraud identification using machine learning approaches. In *2019* 1st International conference on innovations in information and communication technology (ICIICT) (pp. 1-4). IEEE.

[11] Benchaji, I., Douzi, S. and El Ouahidi, B., 2021. Credit card fraud detection model based on LSTM recurrent neural networks. Journal of Advances in Information Technology, 12(2).

[12] Figuerola Ullastres, E., 2022. Credit Card Fraud Detection using Ensemble Learning Algorithms (Doctoral dissertation, Dublin, National College of Ireland).

[13] Bin Sulaiman, R., Schetinin, V. and Sant, P., 2022. Review of machine learning approach on credit card fraud detection. Human-Centric Intelligent Systems, 2(1-2), pp.55-68.

[14] Sohony, I., Pratap, R. and Nambiar, U., 2018, January. Ensemble learning for credit card fraud detection. In Proceedings of the ACM India joint international conference on data science and management of data (pp. 289-294).

[15] Kumar, M.S., Soundarya, V., Kavitha, S., Keerthika, E.S. and Aswini, E., 2019, February. Credit card fraud detection using random forest algorithm. In 2019 3rd International Conference on Computing and Communications Technologies (ICCCT) (pp. 149-153). IEEE.

[14] Hatua, A., Subudhi, B.N., Veerakumar, T. and Ghosh, A., 2021. Early detection of diabetic retinopathy from big data in hadoop framework. Displays, 70, p.102061.

[15] Cloudlitics (2023). Hadoop vs Spark: A Compartative Study. [online] Available at: <https://cloudlytics.com/hadoop-vs-spark-a-comparative-study/> [Accessed 22 September 2023]

[16] Diep, N. (2020). Big Data Analytics: Apache Spark vs. Apache Hadoop. [online] Available at: <https://towardsdatascience.com/big-data-analytics-apache-spark-vs-apache-hadoop-7cb77a7a9424> [Accessed 22 September 2023]

[17] Anderson, S. (2022). Hadoop (MapReduce) vs Apache Spark: A Deep Dive Comparison. [online] Available at: [https://streamsets.com/blog/hadoop-mapreduce-vs-apache-spark/ [Accessed](https://streamsets.com/blog/hadoop-mapreduce-vs-apache-spark/%20%5bAccessed) 22 September 2023]

[18] Guan, S., Zhang, C., Wang, Y. and Liu, W., 2023. Hadoop-based secure storage solution for big data in cloud computing environment. Digital Communications and Networks.

[19] Parmar, D. (2023). Learn Apache Spark in 10 Minutes, Step by Step Guide. [online] Available at: <https://www.youtube.com/watch?v=v_uodKAywXA&t=383s> [Accessed 27 September 2023]

[20] Iqbal, M. (2023). Lecture 4, page 18 Case Study (Spark Programming Paradigm). Big Data Storage and Processing for Data Analytics module. [Accessed 29 September 2023].

[21] Azeroual, O. and Fabre, R., 2021. Processing big data with apache hadoop in the current challenging era of COVID-19. Big Data and Cognitive Computing, 5(1), p.12.

[22] McQuaid, D. (2023). CNN vs. RNN vs. ANN – Analysing 3 Types of Neural Networks in Deep Learning. Advanced Data Analytics Module. [Accessed 24 September, 2023].

[23 ] Ghofrani, J., Kozegar, E., Bozorgmehr, A. and Soorati, M.D., 2019, September. Reusability in artificial neural networks: An empirical study. In Proceedings of the 23rd International Systems and Software Product Line Conference-Volume *B* (pp. 122-129).

[24] Esenogho, E., Mienye, I.D., Swart, T.G., Aruleba, K. and Obaido, G., 2022. A neural network ensemble with feature engineering for improved credit card fraud detection. IEEE Access, 10, pp.16400-16407.

[25] Thuwarakesh (2023). Disadvantages of Artificial Neural Networks And Workarounds. [online] Available at: <https://www.the-analytics.club/disadvantages-of-artificial-neural-networks/> [Accessed 24 September 2023]

[26 ] Urwin, M., (2023). 4 Disadvantages of Neural Networks. [online] Available at: <https://builtin.com/data-science/disadvantages-neural-networks> [Accessed 24 September 2023]

[27] Khaldi, R., El Afia, A., Chiheb, R. and Faizi, R., 2017, March. Artificial neural network based approach for blood demand forecasting: Fez transfusion blood center case study. In Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (pp. 1-6).

[28] McQuaid, D. (2023). Loss Functions – Advanced Data Analytics Module. [Accessed 20 September, 2023].

[29] Iqbal, M. (2022). Tutorial 6 ANN. Case Study (Artificial Neural Network lecture). Machine Learning for Data Analytics module. [Accessed 20 September 2023].

[30] Puligadda, m. And chapala, h.m., deep learning challenges in big data analytics.

[31] Diagrams.net (2023). [online] Available at: <https://app.diagrams.net/#Wa27be703f96b30ad%2FA27BE703F96B30AD!3523> [Accessed 02 October 2023].

[32] Madhugir, D. (2022). Apache Spark Vs. Hadoop MapReduce – Top 7 Differences. [online] Available at: <https://www.analyticsvidhya.com/blog/2022/06/apache-spark-vs-hadoop-mapreduce-top-7-differences/> [Accessed 02 October 2023]’

[3] Santosh, T., and Ramesh, D., (2019). Machine Learning Approach on Apache Spark for Credit Card Fraud Detection. [online] Available at: <https://www.iieta.org/journals/isi/paper/10.18280/isi.250113#:~:text=The%20proposed%20hybrid%20approach%20is%20the%20integration%20of,and%20fraud%20instance%20in%20the%20dataset%20is%20classified>. [Accessed 02 October 2022]

[34] McQuaid, D. (2023).– Feature Scaling or Normalization – Data Preparation and Visualization for Data Analytics Module. [Accessed 20 September 2023].

[35] Boulle, A., Chandramohan, D. and Weller, P., 2001. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. International journal of epidemiology, 30(3), pp.515-520.

[36] Cloudduggu (2023) Apache Spark Architecture. [online] Available at: <https://www.cloudduggu.com/spark/architecture/> [Accessed 01 October 2023].

[37] Aru, J., Labash, A., Corcoll, O. and Vicente, R., 2023. Mind the gap: challenges of deep learning approaches to Theory of Mind. Artificial Intelligence Review, pp.1-16.

[38] Chollet, F. (2017) The limitations of Deep Learning. [online] Available at: <https://blog.keras.io/the-limitations-of-deep-learning.html> [Accessed 01 October 2023].

[39]Thompson, N.C., Greenewald, K., Lee, K. and Manso, G.F., 2020. The computational limits of deep learning. arXiv preprint arXiv:2007.05558.

[40] JavaPoint (2023). Advantages and Disadvantages of Big Data. [online] Available at: [https://www.javatpoint.com/advantages-and-disadvantages-of-big-data [Accessed 05 October 2023]

[41] Krishnan, H. (2023). Disadvantages of Big Data [online] Available at: <https://www.knowledgehut.com/blog/big-data/disadvantages-of-big-data> [Accessed 02 October]

[42] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[43] McQuaid, D. (2023). Data Visualization and Communication. What visualization should I used? Data Preparation and Visualization module. (Slides 5,6,10, 12) [accessed 01 October 2023].