A Project report on

PREDICTIVE ANALYTICS WITH MACHINE LEARNING FOR FRAUD DETECTION OF ONLINE TRANSACTIONS

Submitted in partial fulfillment of the requirements

For the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

By

L. NARAYANA REDDY (204G1A0565)

Under the Guidance of

Mrs. N. Ushasree, M.Tech

Assistant Professor

Department of Computer Science & Engineering

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY
(AUTONOMOUS)
Rotarypuram Village, B K Samudram Mandal, Ananthapuramu- 515701

2023-2024

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY

(AUTONOMOUS)

(Affiliated to JNTUA, Accredited by NAAC with 'A' Grade, Approved by AICTE, New Delhi &
Accredited by NBA (EEE, ECE & CSE)
Rotarypuram Village, BK Samudram Mandal, Ananthapuramu-515701

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Certificate

This is to certify that the Project report entitled PREDICTIVE ANALYTICS WITH MACHINE LEARNING FOR FRAUD DETECTION OF ONLINE MARKETING TRANSACTIONS is the bonafide work carried out by **L. Narayana Reddy** bearing Roll Number **204G1A0565** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** during the academic year 2023-2024.

Project Guide
Mrs. N. Ushasree M.Tech
Assistant Professor

Head of the Department
Mr. P. Veera Prakash M.Tech.,(Ph.D)
Assistant Professor

Date:

External Examiner

Place: Rotarypuram

DECLARATION

I'm, **Mr. L. Narayana Reddy** with reg no: 204G1A0565 students of **SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY**, Rotarypuram, hereby declare that the dissertation entitled **"PREDICTIVE ANALYTICS WITH MACHINE LEARNING FOR FRAUD DETECTION OF ONLINE MARKETING TRANSACTIONS"** embodies the report of our project work carried out by us during IV year Bachelor of Technology under the guidance of **Mrs. N. Ushasree,** Department of CSE, and this work has been submitted for the partial fulfillment of the requirements for the award of the Bachelor of Technology degree.

The results embodied in this project have not been submitted to any other University of Institute for the award of any Degree or Diploma.

NARAYANA REDDY L Reg no: 204G1A0565

VISION & MISSION OF THE INSTITUTION

Vision:

To become a premier Educational Institution in India offering the best teaching and learning environment for our students that will enable them to become complete individuals with professional competency, human touch, ethical values, service motto, and a strong sense of responsibility towards environment and society at large.

Mission:

- Continually enhance the quality of physical infrastructure and human resources to evolve into a center of excellence in engineering education.

- Provide comprehensive learning experiences that are conducive for the students to acquire professional competences, ethical values, life-long learning abilities and understanding of the technology, environment and society.

- Strengthen industry institute interactions to enable the students work on realistic problems and acquire the ability to face the ever-changing requirements of the industry.

- Continually enhance the quality of the relationship between students and faculty which is a key to the development of an exciting and rewarding learning environment in the college.

VISION & MISSION OF THE DEPARTMENT OF CSE

Vision:

To evolve as a leading department by offering best comprehensive teaching and learning practices for students to be self-competent technocrats with professional ethics and social responsibilities.

Mission:

DM 1: Continuous enhancement of the teaching-learning practices to gain profound knowledge in theoretical& practical aspects of computer science applications.

DM 2: Administer training on emerging technologies and motivate the students to inculcate self-learning abilities, ethical values and social consciousness to become competent professionals.

DM 3: Perpetual elevation of Industry-Institute interactions to facilitate the students to work on real-time problems to serve the needs of the society.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that I have now the opportunity to express our gratitude for all of them.

It is with immense pleasure that I would like to express our indebted gratitude to our Guide **Mrs. N. Ushasree, Assistant Professor, Computer Science & Engineering**, who has guided us a lot and encouraged us in every step ofthe project work. I thank him for the stimulating guidance, constant encouragement and constructive criticism which have made possible to bring out this project work.

I express our deep felt gratitude to **Mr. C. Lakshminatha Reddy, Assistant Professor** and **Mr. M. Narasimhulu, Assistant Professor,** Project Coordinators for their valuable guidance and unstinting encouragement enabled us to accomplish our project successfully in time

I'm very much thankful to **Mr. P. Veera Prakash, Assistant Professor & Head of the Department, Computer Science & Engineering,** for his kind support and for providing necessary facilities to carry out the work.

I wish to convey our special thanks to **Dr. G. Bala Krishna, Principal** of **Srinivasa Ramanujan Institute of Technology** for giving the required information in doing our project work. Not to forget, I thank all other faculty and non-teaching staff, and our friends who had directly or indirectly helped and supported us in completing our project in time.

I also express our sincere thanks to the Management for providing excellent facilities**.**

Finally, I wish to convey our gratitude to our families who fostered all the requirements and facilities that I need.

Narayana Reddy L

(204G1A0565)

ABSTRACT

With digital strategies coping up with online marketing system, enormous data passed to these sectors, online transactions are becoming more prone to frauds and threats resulting in data leakage and personal detail exposed to fraudsters leading to huge loss to customer. This makes online marketing systems adapt to high-level security and data handling technology solutions like machine learning, deep learning and predictive analytics which are efficient enough to deal with highly sensitive data, predict frauds and unwanted behavioral patterns in this data. This paper reviews the different advance technologies commonly used to deal with this type of data forms a comparison among them and suggests the most efficient and informative method to use in this sector. Through the end of the review, feature engineering and its selection of parameters for achieving better performance are discussed.

Keywords: Supervised Learning, Fraud Detection, and Machine Learning.

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
PCA	Principle Component Analysis
KNN	K – Nearest Neighbors
GBM	Gradient Boosting Machine
NLP	Natural Language Processing
SRS	System Requirements Specification
LR	Logistic Regression
UML	Unified Modelling Language
XG Boost	Extreme Gradient Boosting
OMT	Online Marketing Transactions
ER	Entity Relationship
DFD	Data Flow Diagram

CHAPTER 1
INTRODUCTION

In the dynamic landscape of online transactions, the unprecedented growth in digital interactions has ushered in a new era of convenience and accessibility. However, this rapid evolution has also presented a significant challenge — the increasing susceptibility of online marketing systems to fraudulent activities. The ramifications of fraud extend beyond financial losses, encompassing data breaches and compromised personal information, resulting in a loss of trust among consumers.

As traditional security protocols prove inadequate in countering sophisticated cyber threats, the imperative to adopt advanced technologies becomes paramount. This paper addresses this critical need by investigating the integration of machine learning (ML), deep learning, and predictive analytics into the fabric of online marketing systems. These technologies stand poised as efficient guardians against the rising tide of fraud, offering the capability to handle vast amounts of sensitive data and predict intricate patterns indicative of fraudulent behaviour.

This study embarks on a comprehensive exploration, comparing and evaluating the effectiveness of diverse advanced technologies. The focus extends to feature engineering and parameter selection to optimise the performance of these technologies. Recognising the limitations of existing methodologies, particularly Logistic Regression and K-Means clustering, the research advocates for a paradigm shift towards Decision Trees, KNN Classifier, and Random Forest algorithms and parameter selection for optimal performance in securing online marketing systems.

1.1 Problem Statement

The Problem Statement revolves around With the surge in online transactions, there is an increase in data leakage, fraud, and threats causing significant losses. The existing security protocols are becoming obsolete and unable to handle sophisticated cyber threats efficiently. The challenge is to implement ML, DL, and Predictive Analytics effectively to analyze massive datasets, learn patterns, and make accurate predictions to enhance the security of online marketing systems.

1.2 Objectives

The objective is to integrate Machine Learning (ML) and Predictive Analytics into online marketing systems to enhance security. These technologies aim to analyze extensive data, identify potential fraud, and ensure secure transactions.

Through continuous learning, they adapt to new threats, providing robust protection and instilling confidence among users, thus safeguarding digital assets and personal data efficiently.

1.3 Scope of the Project

This review explores advanced technologies safeguarding online transactions against fraud and data breaches, emphasizing machine learning, deep learning, and predictive analytics.

By comparing their efficacy in handling sensitive data, predicting fraud, and detecting abnormal behavior, it provides guidance on the optimal technology for secure online marketing. The discussion concludes with an examination of feature engineering and parameter selection to enhance performance.

1.4 Machine Learning

Machine Learning, a subset of artificial intelligence, enables computers to learnfrom data and improve performance without explicit programming. It involves algorithms trained on labeled datasets, predicting outcomes and making decisions. This documentation provides a concise introduction to key concepts, workflow, and applications, facilitating a foundational understanding of machine learning principles.

1.5 Deep Learning

Deep Learning, a subset of machine learning, leverages neural networks with multiple layers to model and solve complex tasks. This documentation offers a brief yet comprehensive overview, introducing key concepts and applications for understanding and implementing deep learning in diverse domains. In deep learning we have architectures like CNN, FNN, RNN, LSTM, GAN, and Autoencoders.

CHAPTER 2

LITERATURE SURVEY

[1] L. Bhavya , V. Sasidhar Reddy , U. Anjali Mohan , S. Karishma, 2020, Credit Card Fraud Detection using Classification, Unsupervised, Neural Networks Models, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020). , Nowadays online transactions have grown in large quantities. Among them, online credit card transactions hold a huge share. Therefore, there is much need for credit card fraud detection applications in bans and financial business. Credit card fraud purposes may be to obtain goods without paying or to obtain unauthorized funds from an account. With the demand for money credit card fraud events became common. This results in a huge loss in finances to the cardholder.

[2] Renjith, Shini. (2018). Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach. International Journal of Engineering Trends and Technology,

The e-commerce share in the global retail spend is showing a steady increase over the years indicating an evident shift of consumer attention from bricks and mortar to clicks in retail sector. In recent years, online marketplaces have become one of the key contributors to this growth. Fraudulent e-commerce buyers and their transactions are being studied in detail and multiple strategies to control and prevent them are discussed. Another area of fraud happening in marketplaces are on the seller side and is called merchant fraud. Goods/services offered and sold at cheap rates, but never shipped is a simple example of this type of fraud. This paper attempts to suggest a framework to detect such fraudulent sellers with the help of machine learning techniques.

[3] Saputra, Adi & Suharjito, Suharjito. (2019). Fraud Detection using Machine Learning in e-Commerce. 10.14569/IJACSA.2019.0100943, The volume of internet users is increasingly causing transactions on e-commerce to increase as well. We observe the quantity of fraud on online transactions is increasing too. Fraud Predictive Analytics with Machine Learning for Fraud Detection of online Marketing Transactions Computer Science & Engineering, SRIT Page 4 prevention in e-commerce shall be developed using machine learning, this work to analyze the suitable machine learning algorithm, the algorithm to be used is the Decision Tree, Naive Bayes, Random Forest, and Neural Network. Result of evaluation using confusion matrix achieve the highest accuracy of the neural network by 96 percent, random forest is 95 percent, Naïve Bayes is 95 percent, and Decision tree is 91 percent. Synthetic Minority Over-sampling Technique (SMOTE) is able to increase the average of F1-Score from 67.9 percent to 94.5 percent and the average of G-Mean from 73.5 percent to 84.6 percent.

[4] A. K. Rai and R. K. Dwivedi, "Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 421-426, doi: 10.1109/ICESC48915.2020.9155615, Development of communication technologies and e-commerce has made the credit card as the most common technique of payment for both online and regular purchases. So, security in this system is highly expected to prevent fraud transactions. Fraud transactions in credit card data transaction are increasing each year. In this direction, researchers are also trying the novel techniques to detect and prevent such frauds. However, there is always a need of some techniques that should precisely and efficiently detect these frauds. This paper proposes a scheme for detecting frauds in credit card data which uses a Neural Network (NN) based unsupervised learning technique. Proposed method outperforms the existing approaches of Auto Encoder (AE), Local Outlier Factor (LOF), Isolation Forest (IF) and K-Means clustering. Proposed NN based fraud detection method performs with 99.87% accuracy whereas existing methods AE, IF, LOF and K Means gives 97%, 98%, 98% and 99.75% accuracy respectively.

CHAPTER 3

PLANNING

3.1 Machine Learning

Machine Learning is undeniably one of the most influential and powerful technologies in today‘s world. Machine learning is a tool for turning information into knowledge. In the past 50 years, there has been an explosion of data. This mass of data is useless; we analyse it and find the patterns hidden within. Machine learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover.

The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making. To learn the rules governing a phenomenon, machines have to go through a learning process, trying different rules and learning from how well they perform. Hence, why it‘s known as Machine Learning.

Basic Terminology:

- Dataset: A set of data examples, which contain features important to solving the problem.
- Features: Important pieces of data that help us understand a problem. These are fed into a Machine Learning algorithm to help it learn.
- Model: The representation (internal model) of a phenomenon that a Machine Learning algorithm has learnt. It learns this from the data it is shown during training. The model is the output you get after training an algorithm. For example, a decision tree algorithm would be trained and produce a decision tree model.

Types of Machine Learning:

There are multiple forms of Machine Learning; supervised, unsupervised, semi supervised and reinforcement learning. Each form of Machine Learning has differing approaches, but they all follow the same underlying process and theory.

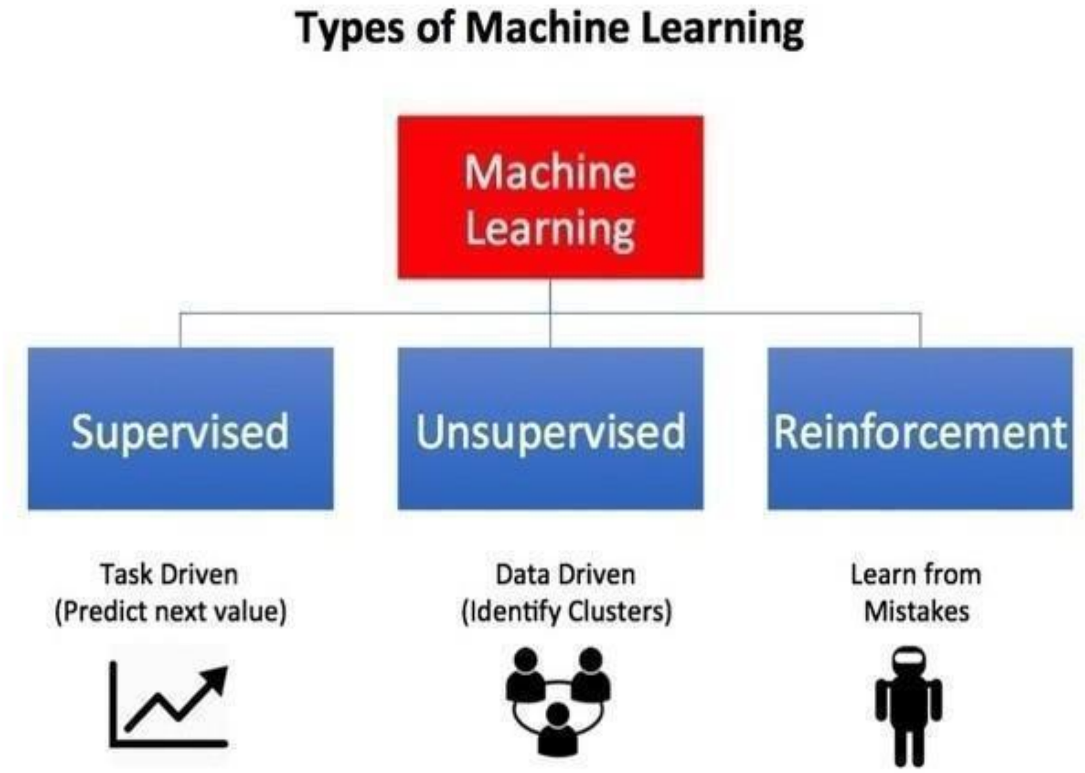

Types of Machine Learning
Machine Learning
Supervised
Unsupervised
Reinforcement
Task Driven (Predict next value)
Data Driven (Identify Clusters)
Learn from Mistakes

Fig. 3.1: Types of Machine Learning

Supervised Learning: It is the most popular paradigm for machine learning. Given data in the form of examples with labels, we can feed a learning algorithm these example-label pairs one by one, allowing the algorithm to predict the label for each example, and giving it feedback as to whether it predicted the right answer or not. Over time, the algorithm will learn to approximate the exact nature of the relationship between examples and their labels. When fully-trained, the supervised learning algorithm will be able to observe a new, never before-seen example and predict a good label for it.

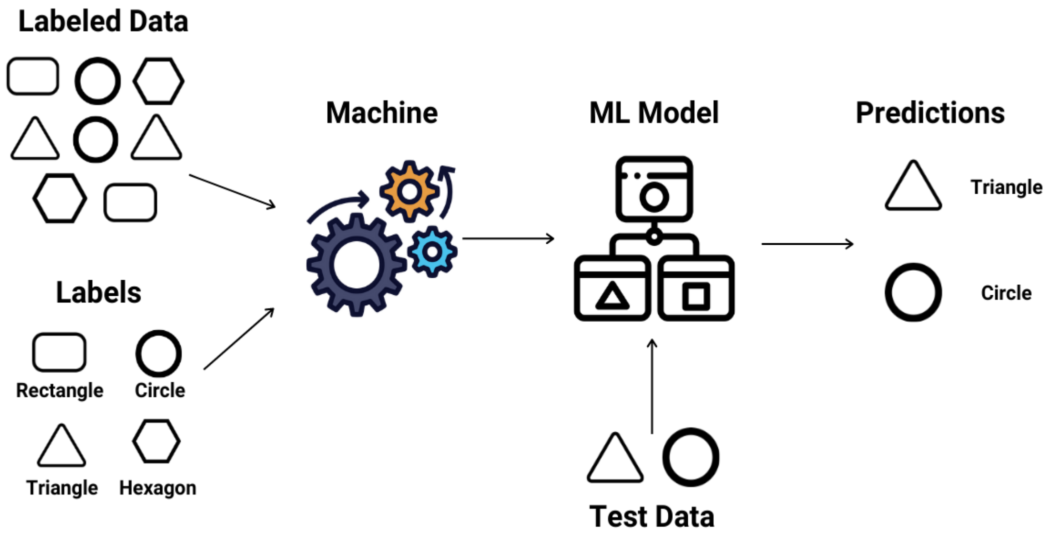


Fig. 3.2: Process of Supervised Learning

Unsupervised learning: It is very much the opposite of supervised learning. It features no labels. Instead, the algorithm would be fed a lot of data and given the tools to understand the properties of the data. From there, it can learn to group, cluster, and organize the data in a way such that a human can come in and make sense of the newly organized data. Because unsupervised learning is based upon the data and its properties, we can say that unsupervised learning is data- driven. The outcomes from an unsupervised learning task are controlled by the data and the way it‘s formatted.

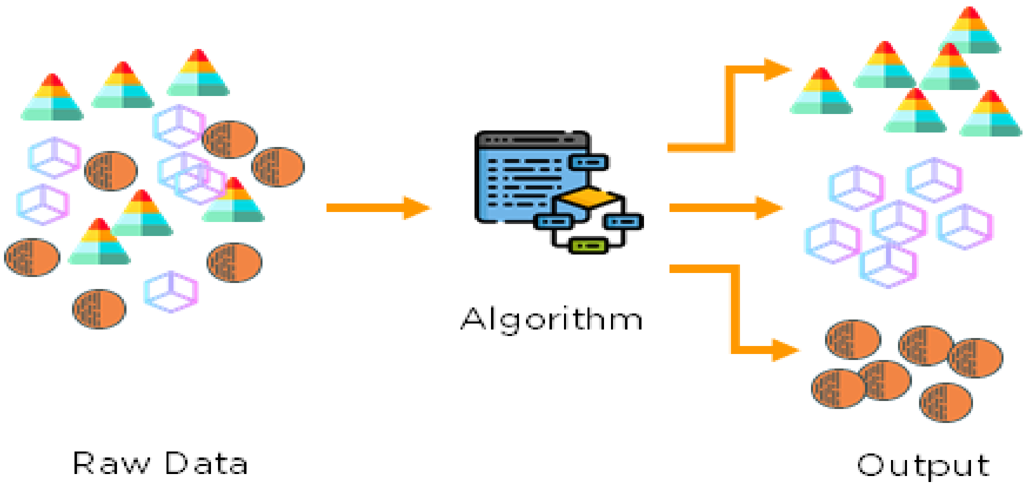


Fig. 3.3: Process of Unsupervised Learning

Reinforcement learning: It is fairly different when compared to supervised and unsupervised learning. Reinforcement learning is very behaviour driven. It has influences from the fields of neuroscience and psychology. For any reinforcement learning problem, we need an agent and an environment as well as a way to connect the two through a feedback loop. To connect the agent to the environment, we give it a set of actions that it can take that affect the environment. To connect the environment to the agent, we have it continually issue two signals to the agent: an updated state and a reward (our reinforcement signal for behaviour).

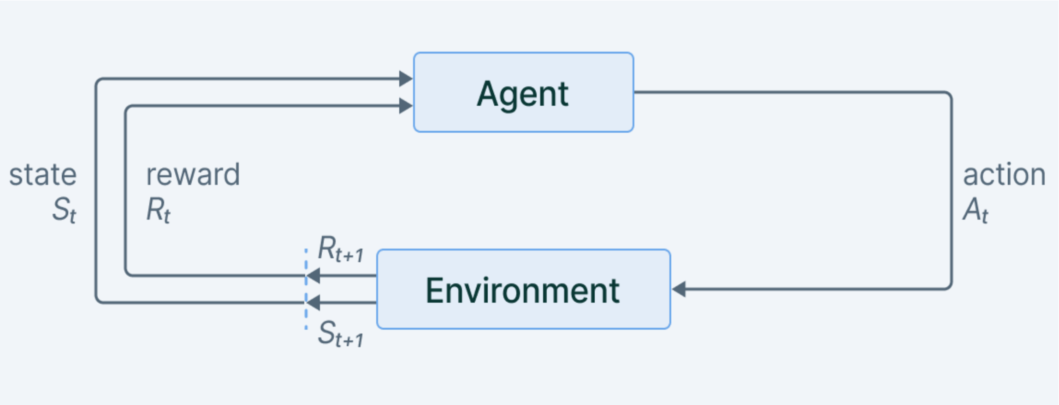


Fig. 3.4: Reinforcement Learning

3.2 Algorithm Used

K Nearest Neighbors:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.

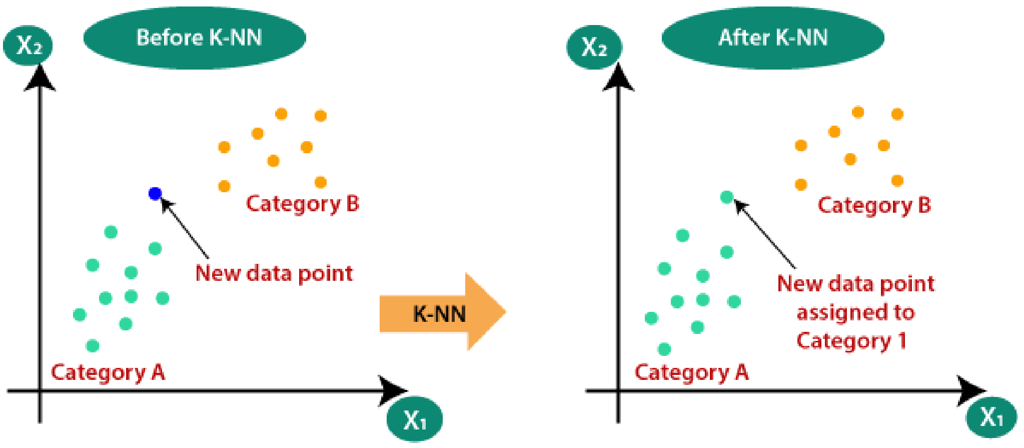


Fig. 3.5: KNN Classifier

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

Decision Trees:

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal.

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

Although, a real dataset will have a lot more features and this will just be a branch in a much bigger tree, but you can't ignore the simplicity of this algorithm. The feature importance is clear and relations can be viewed easily. This methodology is more commonly known as learning decision tree from data and above tree is called Classification tree as the target is to classify passenger as survived or died. Regression trees are represented in the same manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees.

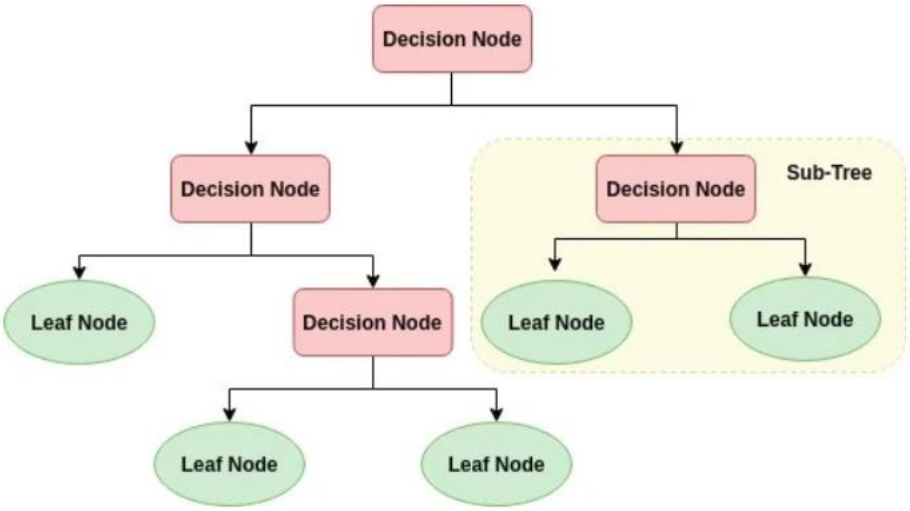


Fig. 3.6: Decision Tree Algorithm

So, what is actually going on in the background? Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop. As a tree generally grows arbitrarily, you will need to trim it down for it to look beautiful. Let's start with a common technique used for splitting.

Random Forest:

First, Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

The author gives four advantages to illustrate why we use Random Forest algorithm. The one mentioned repeatedly by the author is that it can be used for both classification and regression tasks. Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model. The third advantage is the classifier of Random Forest can handle missing values, and the last advantage is that the Random Forest classifier can be modeled for categorical values.

There are two stages in Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first stage.

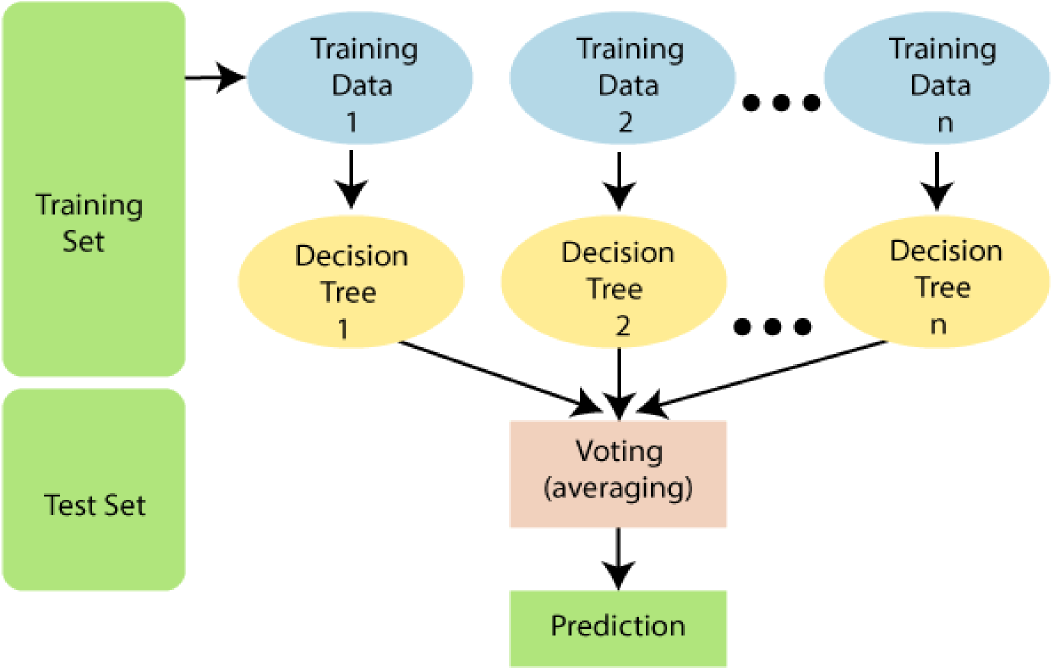


Fig. 3.7: Random Forest Algorithm

STEPS:

1. Randomly select “K” features from total “m” features where k << m
2. Among the “K” features, calculate the node “d” using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat the a to c steps until “l” number of nodes has been reached
5. Build forest by repeating steps a to d for “n” number times to create “n” number of trees

XGBoost:

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

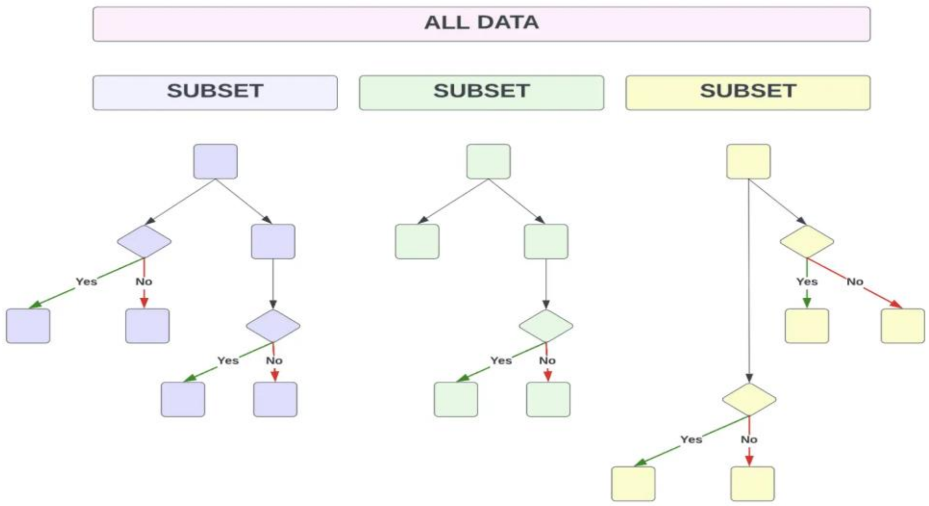


Fig. 3.8: XG Boost Algorithm

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

CHAPTER 4

SYSTEM REQUIREMENTS SPECIFICATIONS

4.1 Functional Requirements

These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

Examples of functional requirements:

1) Authentication of user whenever he/she logs into the system
2) System shutdown in case of a cyber-attack
3) A verification email is sent to user whenever he/she register for the first time on some software system.

Benefits of functional requirements:

- Helps you to check whether the application is providing all the functionalities that were mentioned in the functional requirement of that application
- A functional requirement document helps you to define the functionality of a system or one of its subsystems.
- Functional requirements along with requirement analysis help identify missing requirements. They help clearly define the expected system service and behavior.
- Errors caught in the Functional requirement gathering stage are the cheapest to fix.
- Support user goals, tasks, or activities

4.2 Non-Functional Requirements:

These are basically the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to other. They are also called non-behavioral requirements. They basically deal with issues like:

- Portability
- Security
- Maintainability
- Reliability
- Scalability
- Performance
- Reusability
- Flexibility

Benefits of Non-Functional Requirements:

- The nonfunctional requirements ensure the software system follows legal and compliance rules.
- They ensure the reliability, availability, and performance of the software system.
- They ensure good user experience and ease of operating the software.
- They help in formulating security policy of the software system.

4.3 Python Libraries:

Normally, a library is a collection of books or is a room or place where many books are stored to be used later. Similarly, in the programming world, a library is a collection of precompiled codes that can be used later on in a program for some specific well-defined operations. Other than pre-compiled codes, a library may contain documentation, configuration data, message templates, classes, and values, etc.

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer. As we don’t need to write the same code again and again for different programs. Python libraries play a very vital role in fields of Machine Learning, Data Science, Data Visualization, etc.

Working of Python Library

As is stated above, a Python library is simply a collection of codes or modules of codes that we can use in a program for specific operations. We use libraries so that we don’t need to write the code again in our program that is already available. But how it works. Actually, in the MS Windows environment, the library files have a DLL extension (Dynamic Load Libraries). When we link a library with our program and run that program, the linker automatically searches for that library. It extracts the functionalities of that library and interprets the program accordingly. That's how we use the methods of a library in our program. We will see further, how we bring in the libraries in our Python programs.

Python standard library

The Python Standard Library contains the exact syntax, semantics, and tokens of Python. It contains built-in modules that provide access to basic system functionality like I/O and some other core modules. Most of the Python Libraries are written in the C programming language. The Python standard library consists of more than 200 core modules. All these work together to make Python a high-level programming language. Python Standard Library plays a very important role. Without it, the programmers can't have access to the functionalities of Python. But other than this, there are several other libraries in Python that make a programmer's life easier. Let's have a look at some of the commonly used libraries:

1.Pandas: Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

2. Numpy: The name "Numpy" stands for "Numerical Python". It is the commonly used library. It is a popular machine learning library that supports large matrices and multi-dimensional data. It consists of in-built mathematical functions for easy computations. Even libraries like TensorFlow use Numpy internally to perform several operations on tensors. Array Interface is one of the key features of this library.

3. Flask: Flask is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries.[2] It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

4. OpenCV: OpenCV is an open-source software library for computer vision and machine learning. The OpenCV full form is Open Source Computer Vision Library. It was created to provide a shared infrastructure for applications for computer vision and to speed up the use of machine perception in consumer products. OpenCV, as a BSD-licensed software, makes it simple for companies to use and change the code. There are some predefined packages and libraries that make our life simple and OpenCV is one of them.

Use of Libraries in Python Program

As we write large-size programs in Python, we want to maintain the code's modularity. For the easy maintenance of the code, we split the code into different parts and we can use that code later ever we need it. In Python, modules play that part. Instead of using the same code in different programs and making the code complex, we define mostly used functions in modules and we can just simply import them in a program wherever there is a requirement. We don't need to write that code but still, we can use its functionality by importing its module. Multiple interrelated modules are stored in a library. And whenever we need to use a module, we import it from its library. In Python, it's a very simple job to do due to its easy syntax. We just need to use **import**.

4.4 Hardware Requirements

The hardware requirements include the requirements specification of the physical computer resources for a system to work efficiently. The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. The Hardware Requirements are listed below:

Processor	- I7/Intel Processor
Hard Disk	- 160GB
Key Board	- Standard Windows Keyboard
Mouse	- Two or Three Button Mouse
Monitor	- SVGA
RAM	- 8GB

1. **Processor:** A processor is an integrated electronic circuit that performs the calculations that run a computer. A processor performs arithmetical, logical, input/output (I/O) and other basic instructions that are passed from an operating

system (OS). Most other processes are dependent on the operations of a processor. A minimum 1 GHz processor should be used, although we would recommend S2GHz or more. A processor includes an arithmetical logic and control unit (CU), which measures capability in terms of the following:

- Ability to process instructions at a given time
- Maximum number of bits/instructions
- Relative clock speed

Fig. 4.1: Processor

The proposed system requires a 2.4 GHz processor or higher.

2. Ethernet connection (LAN) OR a wireless adapter (Wi-Fi): Wi-Fi is a family of radio technologies that is commonly used for the wireless local area networking (WLAN) of devices which is based around the IEEE 802.11 family of standards. Devices that can use Wi-Fi technologies include desktops and laptops, smartphones and tablets, TV‘s and printers, digital audio players, digital cameras, cars and drones. Compatible devices can connect to each other over Crop Yield Prediction and Fertilizer Analysis Using Machine Learning Wi- Fi through a wireless access point as well as to connected Ethernet devices and may use it to access the Internet. Such an access point (or hotspot) has a range of about 20 meters (66 feet) indoors and a greater range outdoors. Hotspot coverage can be as small as a single room with walls that block radio waves, or as large as many square 18ilometers achieved by using multiple overlapping access points.

Fig. 4.2: Ethernet Connection

3. Hard Drive: A hard drive is an electro-mechanical data storage device that uses magnetic storage to store and retrieve digital information using one or more rigid rapidly rotating disks, commonly known as platters, coated with magnetic material. The platters are paired with magnetic heads, usually arranged on a moving actuator arm, which reads and writes data to the platter surfaces. Data is accessed in a random-access manner, meaning that individual blocks of data can be stored or retrieved in any order and not only sequentially. HDDs are a type ofen on volatile storage, retaining stored data even when powered off. 32 GB or higher is recommended for the proposed system.

Fig. 4.3: Hard Disk

4. Memory (RAM): Random-access memory (RAM) is a form of computer data storage that stores data and machine code currently being used. A random- access memory device allows data items to be read or written in almost the same amount of time irrespective of the physical location of data inside the memory. In today's technology, random-access memory takes the form of integrated chips. RAM is normally associated with volatile types of memory (such as DRAM modules), where stored information is lost if power is removed, although non- volatile RAM has also been developed. A minimum of 8 GB RAM is recommended for the proposed system.

Fig. 4.4: RAM

4.5 Software Requirements

The software requirements are description of features and functionalities of the target system. Requirements convey the expectations of users from the software product. The requirements can be obvious or hidden, known or unknown, expected or unexpected from client‘s point of view.

Operating System	:	Windows 11
Server side Script	:	HTML, CSS & JS
Programming Language	:	Python
Libraries	:	Flask, Pandas, Numpy
IDE/Workbench	:	PyCharm
Technology	:	Python 3.6+

1.PyCharm: Py Charm is the most popular IDE for Python, and includes great features such as excellent code completion and inspection with advanced debugger and support for web programming and various frameworks. The intelligent code editor provided by PyCharm enables programmers to write high quality Python code. The editor enables programmers to read code easily through colour schemes, insert indents on new lines automatically, pick the appropriate coding style, and avail context-aware code completion suggestions.

At the same time, the programmers can also use the editor to expand a code block to an expression or logical block, avail code snippets, format the code base, identify errors and misspellings, detect duplicate code, and auto-generate code. PyCharm offers some of the best features to its users and developers in the following aspects

- Code completion and inspection
- Advanced debugging
- Support for web programming and frameworks such as Django and Flask

Fig. 4.5: Pycharm image

2. Python: It is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options. Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability. Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers. Additionally, Python supports the use of modules and a package, which means that programs can be designed in a modular style and code can be reused across a variety of projects.

Fig. 4.6: Python Icon

5.Flask Framework: Web Application Framework or simply Web Framework represents a collection of libraries and modules that enables a web application developer to write applications without having to bother about low-level details such as protocols, thread management etc. Flask is a web application framework written in Python. It is developed by Armin Ronacher, who leads an international group of Python enthusiasts named Pocco. Flask is based on the Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects. Web Server Gateway Interface (WSGI) has been adopted as a standard for Python web application development. WSGI is a specification for a universal interface between the web server and the web applications.

werkzeug is a WSGI toolkit, which implements requests, response objects, and other utility functions. This enables building a web framework on top of it. The Flask framework uses Werkzeug as one of its bases.

Jinja2 is a popular templating engine for Python. A web templating system combines a template with a certain data source to render dynamic web pages.

Flask is often referred to as a micro framework. It aims to keep the core of an application simple yet extensible. Flask does not have built-in abstraction layer for database handling, nor does it have form a validation support. Instead, Flask supports the extensions to add such functionality to the application.

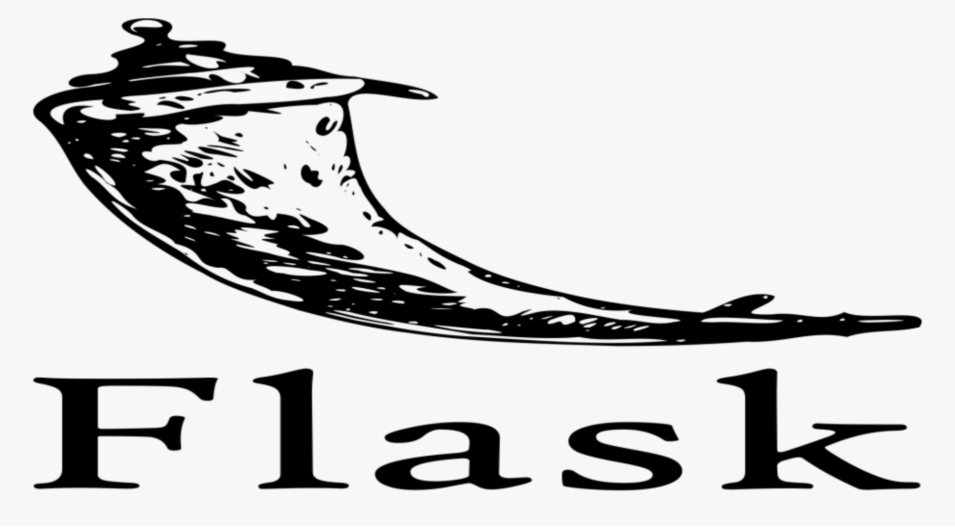


Fig. 4.7: Flask Python Logo

CHAPTER 5
DESIGN

Systems development is a systematic process which includes phases such as planning, analysis, design, deployment, and maintenance. System Analysis is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components. System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose. Analysis specifies what the system should do.

System Design is a process of planning a new business system or replacing an existing system by defining its components or modules to satisfy the specific requirements. Before planning, you need to understand the old system thoroughly and determine how computers can best be used in order to operate efficiently. System Design focuses on how to accomplish the objective of the system.

5.1 UML DIAGRAMS:

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

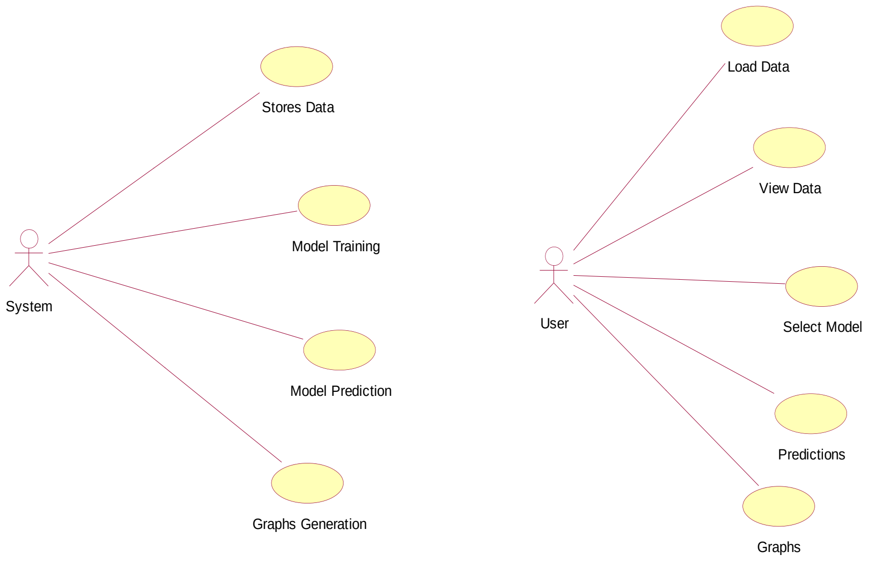


Fig. 5.1: Use case Diagram

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

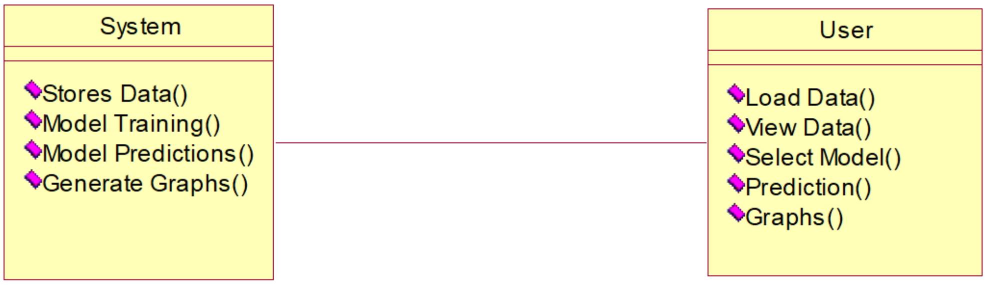


Fig. 5.2: Class Diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

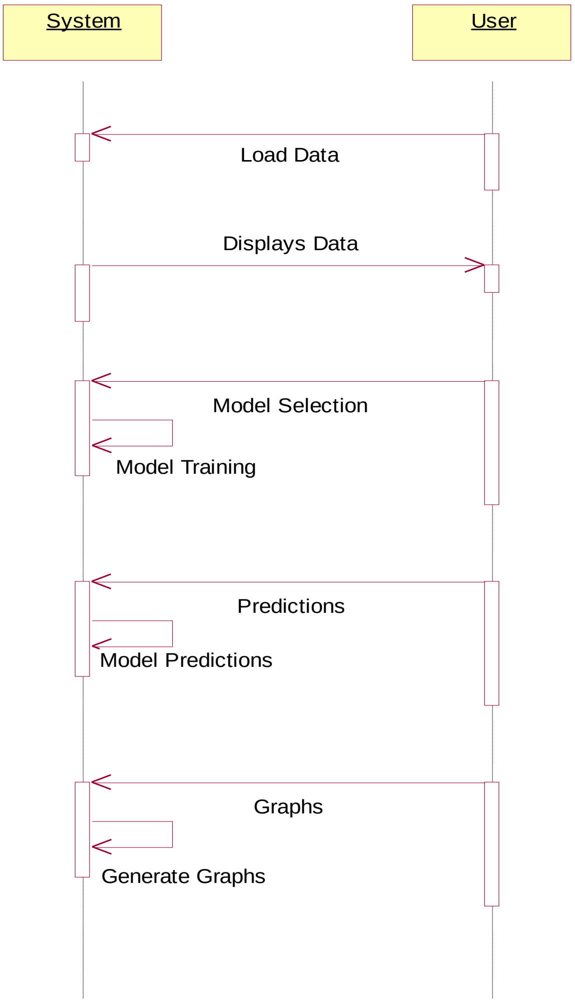


Fig. 5.3: Sequence Diagram

COLLABORATION DIAGRAM:

In collaboration diagram the method call sequence is indicated by some numbering technique as shown below. The number indicates how the methods are called one after another. We have taken the same order management system to describe the collaboration diagram. The method calls are similar to that of a sequence diagram. But the difference is that the sequence diagram does not describe the object organization where as the collaboration diagram shows the object organization.

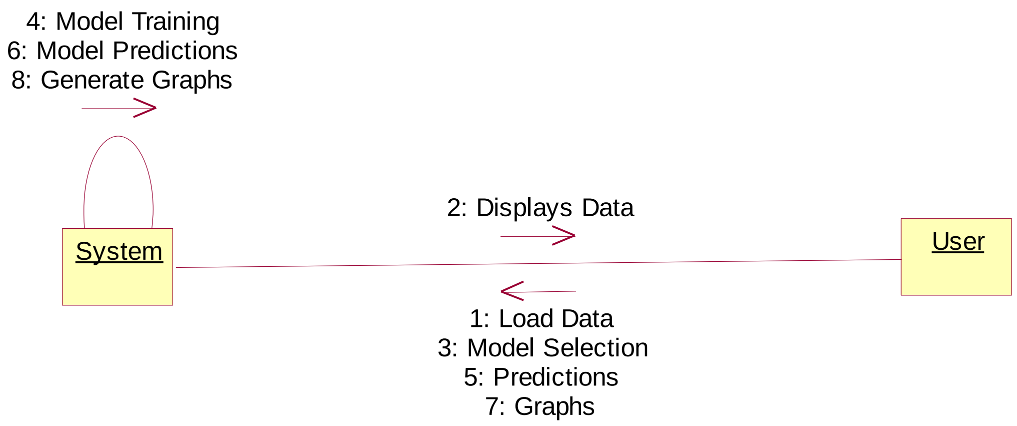


Fig. 5.4: Collaboration Diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

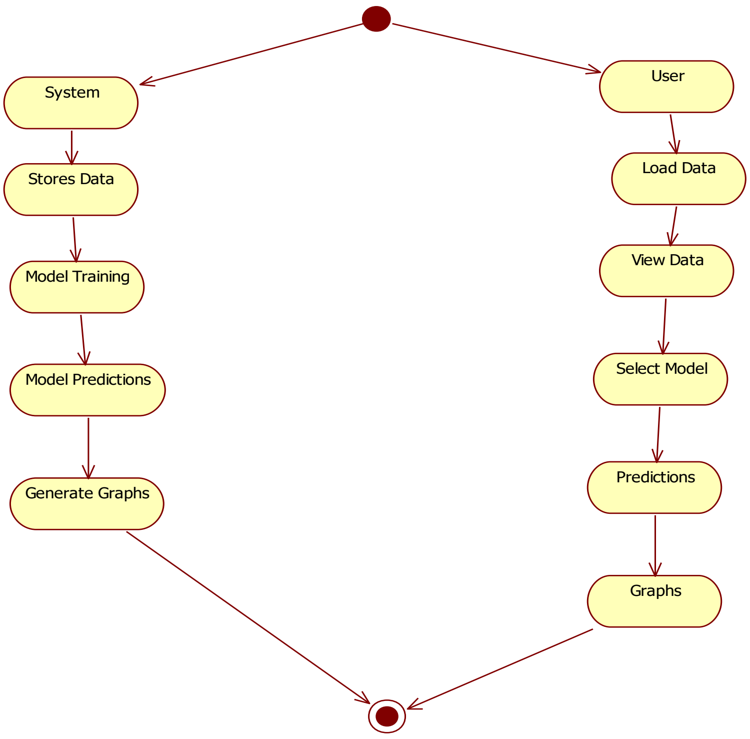


Fig. 5.5: Activity Diagram

1. Usage of UML in Project

As the strategic value of software increases for many companies, the industry looks for techniques to automate the production of software and to improve quality and reduce cost and time to the market. These techniques include component technology, visual programming, patterns and frameworks. Additionally, the development for the World Wide Web, while making somethings simpler, has exacerbated these architectural problems. The UML was designed to respond to these needs. Simply, systems design refers to the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements which can bed one easily through UML diagrams.

5.2 System Architecture

Architecture diagrams can help system designers and developers visualize the high-level, overall structure of their system or application for the purpose of ensuring the system meets their users' needs. They can also be used to describe patterns that are used throughout the design. It's somewhat like a blueprint that can be used as a guide for the convenience of discussing, improving, and following among a team.

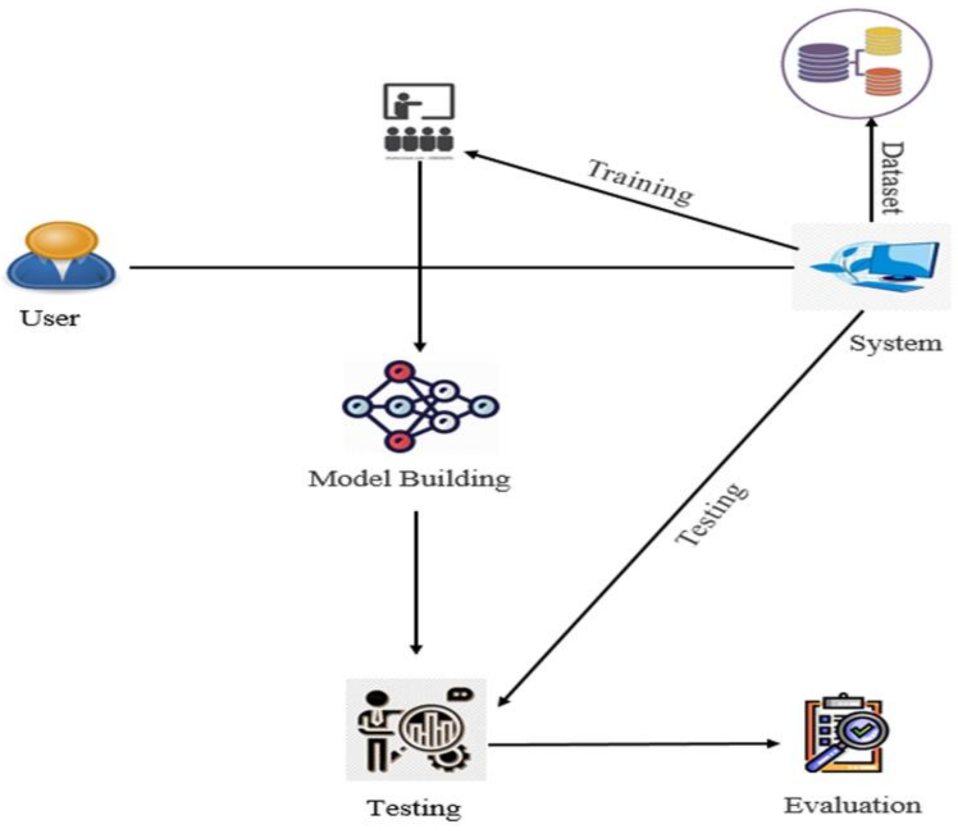


Fig. 5.6: System Architecture

5.3 Flowchart

A flowchart is simply a graphical representation of steps. It shows steps in sequential order and is widely used in presenting the flow of algorithms, workflow or processes. Typically, a flowchart shows the steps as boxes of various kinds, and their order by connecting them with arrows. It originated from computer science as a tool for representing algorithms and programming logic but had extended to use in all other kinds of processes. Nowadays, flowcharts play an extremely important role in displaying information and assisting reasoning. They help us visualize complex processes, or make explicit the structure of problems and tasks. A flowchart can also be used to define a process or project to be implemented.

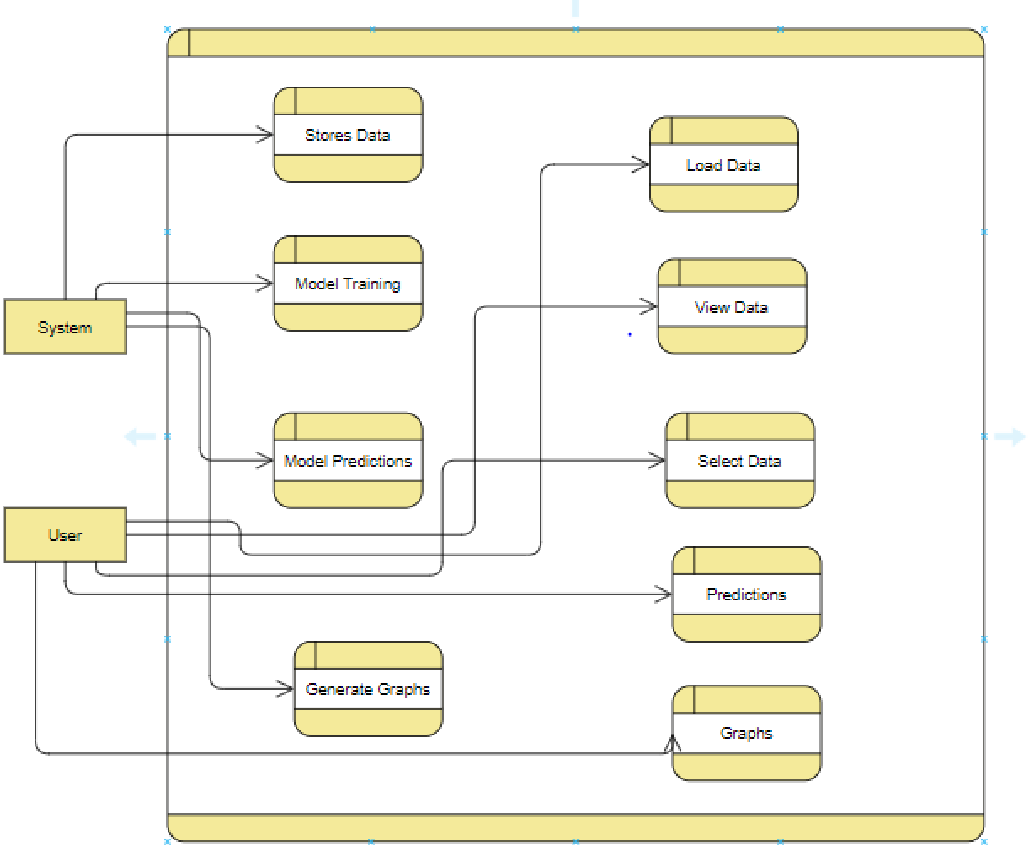


Fig. 5.7: Flowchart of the system

CHAPTER 6

IMPLEMENTATION

Predictive analytics, coupled with machine learning, plays a pivotal role in the realm of fraud detection within online marketing transactions. These systems delve into historical transaction data to identify anomalies such as irregular spending behaviors or suspicious account activities. The integration of predictive analytics with machine learning bolsters security measures within online marketing transactions, bolstering trust and integrity in digital commerce ecosystems.

1.System

Store Dataset: The System stores the dataset given by the user.

Model Training: The system takes the data from the user and fed that data to the selected model.

Model Predictions: The system takes the data given by the user and predict the output based on the given data.

Graphs Generation: The system takes the dataset given by the user, selects the model and generates the graph corresponding to the selected model

2. User:

Load Dataset: The user can load the dataset he/she want to work on.

View Dataset: The User can view the dataset.

Select model: User can apply the model to the dataset for accuracy.

Predictions: User can enter random values for prediction.

Working Flow of the System

Sure, here's a proposed workflow for a predictive analytics system using machine learning for fraud detection in online marketing transactions, outlined in points

1. Data Collection: Gather relevant data from various sources such as transaction logs, customer profiles, IP addresses, device information, and historical fraud records. Secure Voting System Through Face Recognition Computer Science & Engineering, SRIT Page 34

2. Data Preprocessing: Cleanse and preprocess the collected data by handling missing values, standardizing formats, and encoding categorical variables.

3. Feature Engineering: Extract relevant features from the data that are indicative of fraudulent behavior, such as transaction amount, frequency, time of day, geolocation, and user behavior patterns.

4. Model Training: Utilize machine learning algorithms such as logistic regression, random forests, or gradient boosting machines to train a predictive model on the preprocessed data. Train the model using labeled data, where fraudulent transactions are labeled as such.

5. Model Evaluation: Evaluate the trained model's performance using metrics such as accuracy, precision, recall, and F1-score on a separate validation dataset. Adjust hyperparameters and model architecture as necessary to improve performance.

6. Deployment: Deploy the trained model into a production environment where it can receive incoming transaction data in real-time.

7. Real-time Prediction: Apply the trained model to incoming transactions in realtime to predict the likelihood of fraud. Set a threshold probability above which transactions are flagged as potentially fraudulent.

8. Alert Generation: Generate alerts for transactions that exceed the predefined threshold probability for fraud. These alerts can be sent to fraud analysts or automated systems for further investigation.

9. Human Review: Review flagged transactions to validate the model's predictions and take appropriate action, such as blocking suspicious accounts or initiating further verification steps.

10. Model Monitoring and Maintenance: Continuously monitor the model's performance over time and retrain it periodically with updated data to ensure its effectiveness in detecting evolving fraud patterns.

11. Model Prediction Trends: Create a line or bar graph depicting the trends in model predictions over time, such as the frequency of flagged transactions and their corresponding outcomes (e.g., confirmed fraud, false positives)

6.1 Datasets

Machine Learning depends heavily on data. The dataset used for the research is a synthetic dataset generated for the purpose of this study, appendix 1. It contains information about financial transactions, including transaction IDs, customer IDs, transaction amounts, transaction timestamps, regions, states, customer categories, and account balances. The dataset consists of 10000 records and includes characteristic such as geographical information, customer profiles, and transaction details.

6.2 Data Pre-Processing

Data Pre-Processing is a Data Mining method that entails converting raw data into a format that can be understood. Real-world data is frequently inadequate, inconsistent, and/or lacking in specific activities or trends, as well as including numerous inaccuracies. This might result in low-quality data collection and, as a result, low-quality models based on that data. Preprocessing data is a method of resolving such problems. Machines do not comprehend free text, image, or video data; instead, they comprehend 1s and 0s. So putting on a slideshow of all our photographs and expecting our machine learning model to learn from it is probably not going to be adequate. Data Pre-processing is the step in any Machine Learning process in which the data is changed, or encoded, to make it easier for the machine to parse it. In other words, the algorithm can now easily interpret the data's features. Data Pre-processing can be done in four different ways. Data cleaning/cleaning, data integration, data transformation, and data reduction are the four categories.

1. Data Cleaning:

Data in the real world is frequently incomplete, noisy, and inconsistent. Many bits of the data may be irrelevant or missing. Data cleaning is carried out to handle this aspect. Data cleaning methods aim to fill in missing values, smooth out noise while identifying outliers, and fix data discrepancies. Unclean data can confuse data and the model. Therefore, running the data through various Data Cleaning/Cleansing methods is an important Data Pre-processing step.

2. Data Integration:

It is involved in a data analysis task that combines data from multiple sources into a coherent data store. These sources may include multiple databases. Do you think how data can be matched up?? For a data analyst in one database, he finds Customer_ID and in another he finds cust_id, How can he sure about them and say these two belong to the same entity.

3. Data Reduction :

Because data mining is a methodology for dealing with large amounts of data. When dealing with large amounts of data, analysis becomes more difficult. We employ a data reduction technique to get rid of this. Its goal is to improve storage efficiency while lowering data storage and analysis expenses.

4. Dimensionality Reduction

A huge number of features may be found in most real-world datasets. Consider an image processing problem: there could be hundreds of features, also known as dimensions, to deal with. As the name suggests, dimensionality reduction seeks to minimize the number of features but not just by selecting a sample of features from the feature set, which is something else entirely Feature Subset Selection or feature selection.

CHAPTER 7

SYSTEM STUDY &TESTING

7.1 Feasibility Study

The feasibility of the project is analysed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company.

Three key considerations involved in the feasibility analysis are:

- Economical feasibility.
- Technical feasibility
- Social feasibility

7.2 Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization.The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available.

7.3 Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources.This will lead to high demands being placed on the client.

7.4 Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it.

7.5 System Testing

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.6 Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

7.7 Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.8 Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.9 Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

7.10 White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

7.11 Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.12 Test Objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

CHAPTER 8

RESULTS

In the final implementation of the application the first screen the user can view is the Home page that contain five functionalities Load Data, View Data, Select Model, Prediction and Graphs.

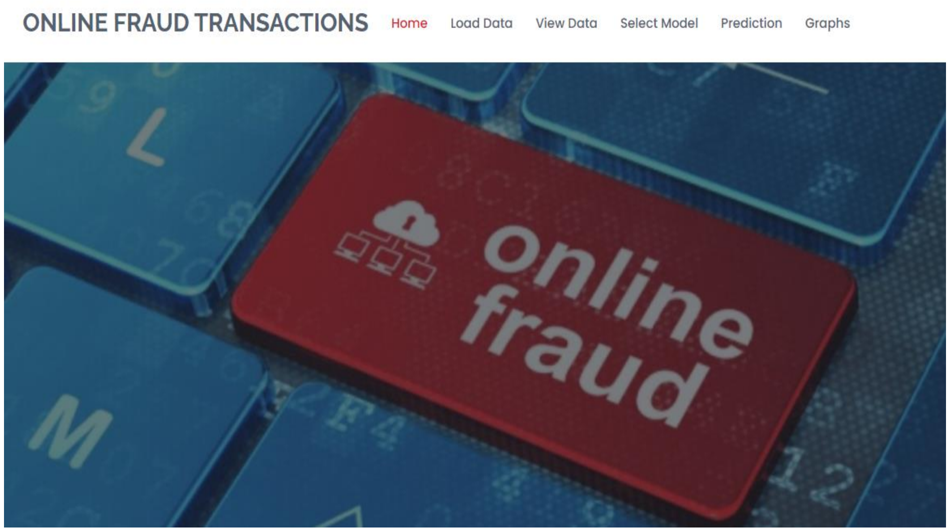


Fig. 8.1: Home Page

The above figure represents the Initial web page of the system that has five functionalities Load Data, View Data, Select Model, Prediction and Graphs. On selecting the required one we get the new webpage and required to give the inputs.

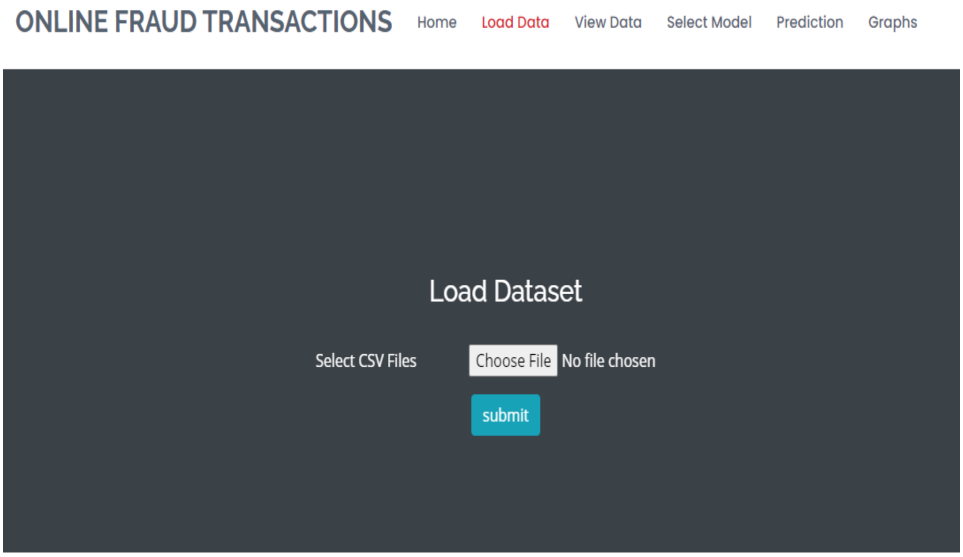
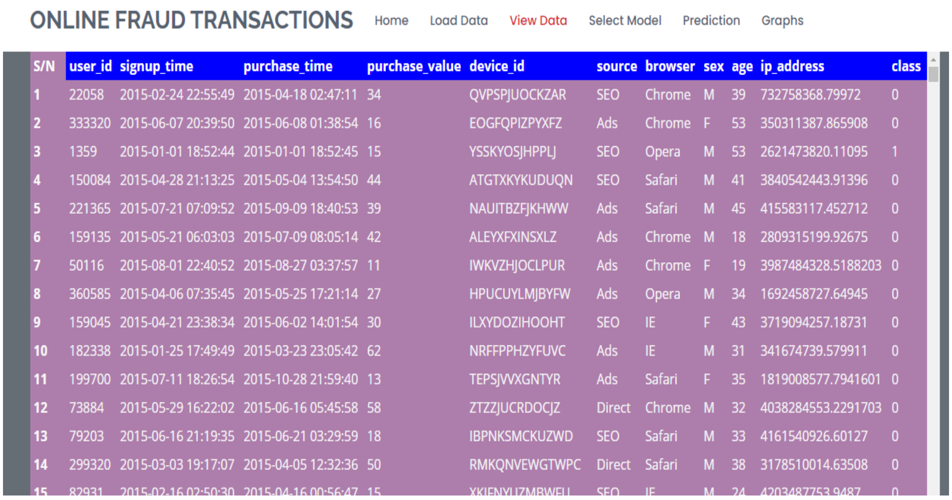


Fig. 8.2: Load Data Page

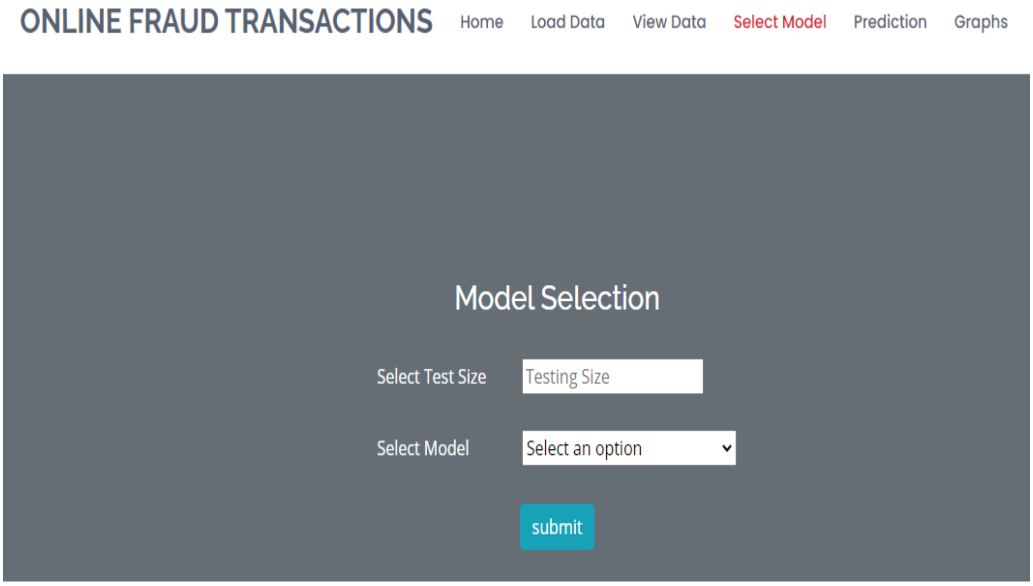
The above page represents the initial page of Loading Dataset. Here we are required to give input Dataset.



S/N	user_id	signup_time	purchase_time	purchase_value	device_id	source	browser	sex	age	ip_address	class
1	22058	2015-02-24 22:55:49	2015-04-18 02:47:11	34	QVPSPJUOCKZAR	SEO	Chrome	M	39	732758368.79972	0
2	333320	2015-06-07 20:39:50	2015-06-08 01:38:54	16	EOGFQPIZPYXFZ	Ads	Chrome	F	53	350311387.865908	0
3	1359	2015-01-01 18:52:44	2015-01-01 18:52:45	15	YSSKYOSJHPPLJ	SEO	Opera	M	53	2621473820.11095	1
4	150084	2015-04-28 21:13:25	2015-05-04 13:54:50	44	ATGTXKYKUDUQN	SEO	Safari	M	41	3840542443.91396	0
5	221365	2015-07-21 07:09:52	2015-09-09 18:40:53	39	NAUITBZFJKHWW	Ads	Safari	M	45	415583117.452712	0
6	159135	2015-05-21 06:03:03	2015-07-09 08:05:14	42	ALEYXFXINSXLZ	Ads	Chrome	M	18	2809315199.92675	0
7	50116	2015-08-01 22:40:52	2015-08-27 03:37:57	11	IWKVZHJOCLPUR	Ads	Chrome	F	19	3987484328.5188203	0
8	360585	2015-04-06 07:35:45	2015-05-25 17:21:14	27	HPUCUYLMJBYFW	Ads	Opera	M	34	1692458727.64945	0
9	159045	2015-04-21 23:38:34	2015-06-02 14:01:54	30	ILXYDOZIHOOHT	SEO	IE	F	43	3719094257.18731	0
10	182338	2015-01-25 17:49:49	2015-03-23 23:05:42	62	NRFFPPHZYFUVC	Ads	IE	M	31	341674739.579911	0
11	199700	2015-07-11 18:26:54	2015-10-28 21:59:40	13	TEPSJVVXGNTYR	Ads	Safari	F	35	1819008577.7941601	0
12	73884	2015-05-29 16:22:02	2015-06-16 05:45:58	58	ZTZZJUCRDOCJZ	Direct	Chrome	M	32	4038284553.2291703	0
13	79203	2015-06-16 21:19:35	2015-06-21 03:29:59	18	IBPNKSMCKUZWD	SEO	Safari	M	33	4161540926.60127	0
14	299320	2015-03-03 19:17:07	2015-04-05 12:32:36	50	RMKQNVEWGTWPC	Direct	Safari	M	38	3178510014.63508	0
15	82931	2015-02-16 02:50:30	2015-04-16 00:56:47	15	XKIFNYUZMBWFU	SEO	IE	M	24	4203487753.9487	0

Fig. 8.3: View Data page

The above page shows us Loaded Data Set. User can see Loaded Data in the above webpage.



Model Selection

Select Test Size: Testing Size

Select Model: Select an option

Fig. 8.4: Model Selection Page

The above page is the third page where user can enter Test Size and Select Model.

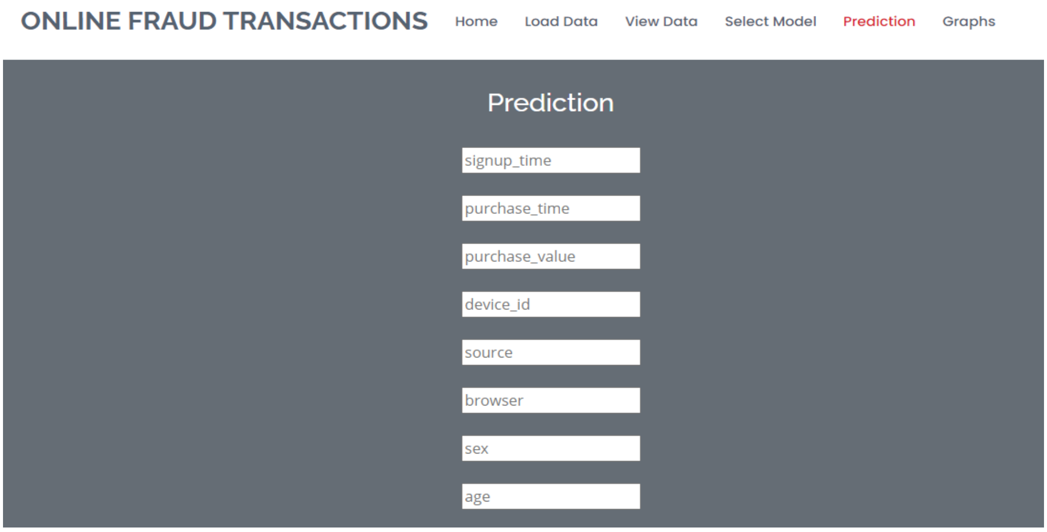


Fig. 8.5: Prediction Page

The above page User has to enter signup_time, purchase_time, purchase_value, device_id, source, browser, sex, age.

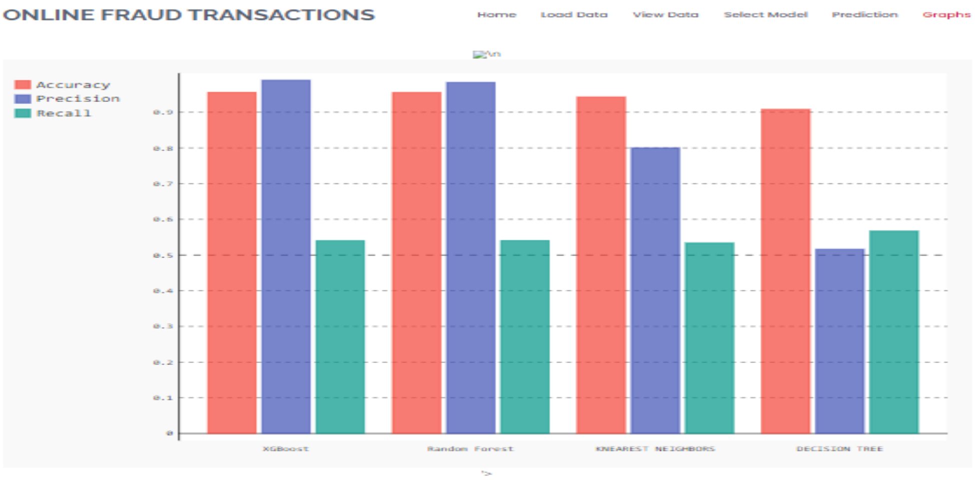


Fig. 8.6: Results Page

The above page user can see the Graph results like Accuracy, Precision, Recall values.

CONCLUSION

In conclusion, In the study, we focused on developing unsupervised machine learning (ML) models with the primary goal of effectively detecting fraudulent transactions. After rigorous evaluation of various algorithms including Extreme Gradient Boosting, K-Nearest Neighbors (KNN), and Random Forest, it became evident that the Random Forest model stood out with superior performance.

The Random Forest model not only showcased high accuracy in identifying fraudulent transactions but also exhibited commendable precision and recall scores. This combination of accuracy, precision, and recall makes the Random Forest algorithm a robust and reliable choice for detecting fraudulent activities within transaction data.

By leveraging the capabilities of the Random Forest model, we can significantly enhance the security and reliability of online transactions. The model's ability to accurately identify fraudulent transactions minimizes the risks associated with financial fraud, thereby safeguarding both businesses and consumers from potential losses.

In summary, the implementation of the Random Forest model as part of fraud detection systems represents a substantial step forward in mitigating the risks posed by fraudulent activities in financial transactions. Its robust performance ensures a proactive approach to identifying and preventing fraudulent behavior, ultimately fostering trust and confidence in online transaction processes.

REFERENCES

[1] Jehovah Jireh Arputhamoni and Gnana Saravanan "Online Smart Voting System Using Biometrics Based Facial and Fingerprint Detection on Image Processing and CNN", [1] Taha, Altyeb & Malebary, Sharaf. (2020). An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. IEEE Access. 8. 25579-25587.

[2] Assaghir, Zainab & Taher, Yehia & Haque, Rafiqul & Hacid, Mohand-Said & Zeineddine, Hassan. (2019). An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. IEEE Access.

[3] L. Meneghetti, M. Terzi, S. Del Favero, G. A Susto, C. Cobelli, "DataDriven Anomaly Recognition for Unsupervised Model-Free Fault Detection in Artificial Pancreas", Ieee Transactions On Control Systems Technology, (2018) pp. 1-15

[4] F. Carcillo, Y.-A. Le Borgne and O. Caelen et al., "Combining unsupervised and supervised learning in credit card fraud detection", Information Sciences, Elsevier (2019), pp. 1-15.

[5] Ashphak, Mr. & Singh, Tejpal & Sinhal, Dr. Amit. (2012). A Survey of Fraud Detection System using Hidden Markov Model for Credit Card Application Prof. Amit Sinhal. 1.

[6] Renjith, Shini. (2018). Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach. International Journal of Engineering Trends and Technology. 57. 48-53. 10.14445/22315381/IJETT-V57P210.

[7] Saputra, Adi & Suharjito, Suharjito. (2019). Fraud Detection using Machine Learning in e-CommercePREDICTIVE ANALYTICS WITH MACHINE LEARNING FOR FRAUD DETECTION OF ONLINE MARKETING TRANSACTIONS (3).docx. 10.14569/IJACSA.2019.0100943.

[8] A. K. Rai and R. K. Dwivedi, "Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 421-426, doi: 10.1109/ICESC48915.2020.9155615.

[9] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare et al., "Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis", IEEE, 2017.

[10] Rajendra Kumar Dwivedi, Sonali Pandey, Rakesh Kumar "A study on Machine Learning Approaches for Outlier Detection in Wireless Sensor Network" IEEE International Conference Confluence, (2018).

Predictive Analytics With Machine Learning For Fraud Detection Of Online Marketing Transactions

Nyasala. UshaSree,[1, a)] Lingannagari. Narayana Reddy, Kakarla. Pushpa, Gujjala. Uday Kiran, Neeruganti. Sai Harsha Vardhan[2,3,4,5 b)]

[1]Assistant Professor, Computer Science and Engineering, Srinivasa Ramanujan Institute of Technology, Anantapur, India.

[2,3,4,5] Computer Science and Engineering, Srinivasa Ramanujan Institute of Technology.

[a)]ushasreen.cse@srit.ac.in,
[b)]204g1a0565@srit.ac.in
[c)]204g1a0574@srit.ac.in
[d)]204g1a05b4@srit.ac.in
[e)]204g1a0585@srit.ac.in

Abstract: The surge in online transactions exposes digital platforms to escalating fraud risks, necessitating advanced security measures. This research explores the integration of ML and predictive analytics to fortify online marketing systems against fraudulent activities. The objective is to proactively identify and prevent data breaches, minimising losses to customers. The study compares advanced technologies, emphasising feature engineering and parameter selection for optimal performance. Current methodologies, relying on Logistic Regression and K-Means clustering, exhibit drawbacks such as low precision, poor recall scores, and high computational time. The proposed system leverages Decision Trees, KNN Classifier, and Random Forest algorithms for enhanced fraud detection, providing accurate predictions, adaptive learning, and efficient processing. This research contributes to securing online transactions in ecommerce and banking, fostering user trust in digital transactions.

Keywords: Predictive Analytics, Machine Learning, Fraud Detection, Online Marketing, Decision Trees, KNN Classifier, Random Forest.

INTRODUCTION

In the dynamic landscape of online transactions, the unprecedented growth in digital interactions has ushered in a new era of convenience and accessibility. However, this rapid evolution has also presented a significant challenge — the increasing susceptibility of online marketing systems to fraudulent activities. The ramifications of fraud extend beyond financial losses, encompassing data breaches and compromised personal information, resulting in a loss of trust among consumers.

In light of the escalating sophistication of cyber threats, conventional security measures are increasingly falling short, necessitating the adoption of more advanced technological solutions. This paper delves into the exploration of integrating ML and predictive analytics into architecture of online marketing systems to address this pressing need. These cutting-edge technologies emerge as formidable defenses against the surging prevalence of fraudulent activities, boasting the capacity to manage extensive datasets and discern complex patterns indicative of fraudulent behavior. By harnessing the capabilities of ML, deep learning, and predictive analytics, online marketing systems can enhance their ability to detect and thwart fraudulent activities, thereby safeguarding both businesses and consumers from potential harm..

This study embarks on a comprehensive exploration, comparing and evaluating the effectiveness of diverse advanced technologies. The focus extends to feature engineering and parameter selection to optimise the performance of these technologies. Recognising the limitations of existing methodologies, particularly Logistic Regression and K-Means clustering, the research advocates for a paradigm shift towards Decision Trees, KNN Classifier, and Random Forest algorithms and parameter selection for optimal performance in securing online marketing systems.

LITERATURE SURVEY

[1] Credit card fraud detection utilizes a combination of classification, unsupervised, and neural network models to enhance security measures. These sophisticated algorithms are adept at accurately identifying fraudulent transactions, thereby minimizing financial losses for both consumers and financial institutions. By leveraging advanced machine learning techniques, such systems play a crucial role in safeguarding sensitive financial data and maintaining trust in electronic payment systems.

[2] This study acknowledges the burgeoning volume of online credit card transactions and the subsequent need for robust fraud detection in banking and financial sectors. The authors explore classification, unsupervised, and neural network models to detect and prevent credit card fraud, highlighting the common motives behind such activities and the financial losses incurred by cardholders.

In response to the growing influence of e-commerce in global retail spending, this research targets fraudulent activities on online marketplaces, particularly focusing on fraudulent sellers engaging in merchant fraud. The study proposes a framework utilising Support Vector Machine techniques to identify and control such fraudulent sellers, shedding light on the nuances of ecommerce fraud.

[3] Amidst the rising tide of online transactions within e-commerce, this research undertakes an in-depth exploration of the escalating occurrences of fraud. Employing a suite of machine learning algorithms including Decision Tree, Random Forest, and Neural Network, the authors meticulously analyze patterns of fraudulent activity. Emphasizing the imperative of fraud prevention in e-commerce realms, the study elucidates the efficacy of these algorithms in achieving remarkable accuracy levels, as evidenced by comprehensive evaluation metrics.

[4] In their research, A. K. Rai and R. K. Dwivedi tackle the urgent challenge of detecting fraud in credit card transactions, recognizing the prevalent usage of credit cards across both online and offline purchases. Their study presents an unsupervised machine learning framework tailored specifically to identify fraudulent activities within credit card datasets. This comprehensive analysis underscores the potential of their framework to fortify security measures and mitigate the financial risks associated with fraudulent credit card transactions, offering valuable insights to the realm of fraud detection within financial systems.

Related Works

In the intersection of banking and computer science, credit card fraud detection emerges as a pivotal area of focus, driving extensive research endeavors and significant resource allocation. Scholars have devoted substantial time and energy to refining prediction systems, harnessing the power of machine learning to combat online fraud effectively. Dal Pozzolo Andrea's thesis, titled "Adaptive Machine Learning for Credit Card Fraud Detection," offers a comprehensive exploration of credit card fraud detection systems. The thesis meticulously examines prediction and classification methodologies using machine learning techniques in its initial segments. Additionally, it advocates for the adoption of diverse sampling techniques to bolster system efficacy. Moreover, the thesis conducts in-depth analysis and evaluation of various machine learning models' performance. Ultimately, it culminates with the practical implementation of a Fraud Detection System (FDS) in real-time scenarios through web API integration, underscoring the research's practical significance in mitigating the financial risks entailed by credit card fraud.

a. Crediting card fraud detection:

[3–9] Several papers in the field address the implementation and efficacy of various machine learning (ML) techniques for fraud detection. Notably, [5], [6], and [8] demonstrate the utilization of simple and rudimentary ML methods for binary classification tasks. Meanwhile, [9] conducts a comparative analysis between different ML algorithms and neural networks. However, a common challenge across these studies is the scarcity and sensitivity of available datasets, as banking information is often shielded and safeguarded. This limitation impedes the derivation of meaningful insights from the data and necessitates substantial processing capacity and time investment for model development. Moreover, [4] highlights a specific issue with the random forest approach, namely, the propensity for overfitting, which undermines the accuracy and reliability of fraud detection models. These findings underscore the complexity and ongoing challenges inherent in developing robust fraud detection systems within the banking sector.

b. Challenges and Motivation

1. Handling enormous daily data requires fast and suitable models for detecting fraudulent transactions effectively.

2. Imbalanced datasets, with low instances of fraudulent transactions, pose challenges in accurate fraud identification.

3. The confidentiality of banking data makes it difficult to obtain for model training and analysis.

4. Misclassified data due to unreported fraudulent transactions can lead to inaccuracies in model predictions.

5. Hackers employ adaptive techniques to circumvent fraud detection models, necessitating ongoing updates and improvements in security measures.

PROPOSED SYSTEM

Our proposed system aims to scrutinize the viability of employing machine learning methodologies for credit card fraud detection. By delving into the effectiveness of Decision Trees, KNN Classifier, and Random Forest Algorithm, we intend to gauge their utility in accurately discerning fraudulent transactions from legitimate ones. Through meticulous analysis of various transaction attributes and labels denoting fraud or non-fraud instances, we seek empirical evidence regarding the efficacy of these techniques. This investigation holds significance for financial institutions, offering them valuable insights into the feasibility of integrating advanced analytics to enhance security measures in the digital realms.

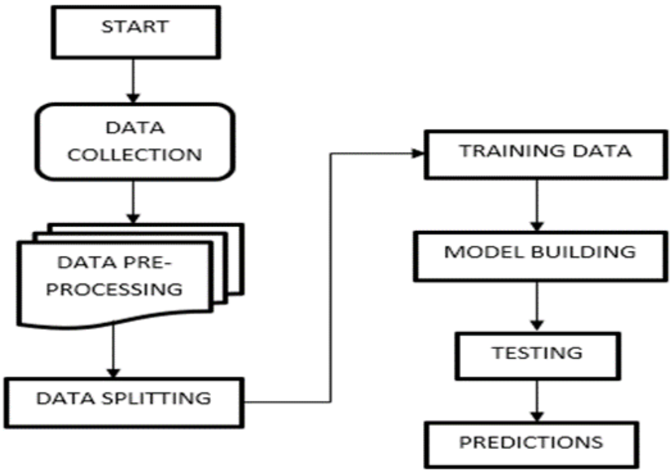


Figure-1. Block diagram for suggested approach

The user will first register using his email address and password in the diagram. Following page login, he or she will upload a file and be able to view it or retrieve it by using a term as a key. By employing that keyword, the user can so share the file with others. The credentials are crucial for the system in the main.

Methodology Overview

1. Data Collection:

For this study, a comprehensive dataset of online transactions, encompassing a diverse range of sources, was collected. The dataset includes information on transaction amounts, timestamps, user details, and other relevant features. Special attention was given to ensuring the inclusion of both legitimate and fraudulent transactions to facilitate a robust analysis.

2. Data Preprocessing:

The collected dataset underwent comprehensive preprocessing, which included handling missing data, removing duplicates, and normalizing numerical features. Categorical variables were encoded, and outliers were addressed to ensure dataset quality and reliability for machine learning analysis.

3. Feature Engineering:

Feature engineering played a pivotal role in shaping the dataset for effective model training. Relevant features were selected based on their potential to contribute to fraud detection accuracy. New features, such as transaction frequency and user behaviour patterns, were engineered to capture intricate aspects of online transaction dynamics.

4. Model Selection:

In the field of online marketing transaction fraud detection, selecting the optimal machine learning model is imperative to ensure the attainment of precise and dependable outcomes, with a thorough evaluation undertaken to assess the suitability of various models..

4.1 Decision Trees:

Overview: Decision Trees are powerful tools for classification tasks, capable of representing complex decisionmaking processes. Every node within the tree signifies a decision contingent on a specific feature, ultimately culminating in a conclusive outcome.

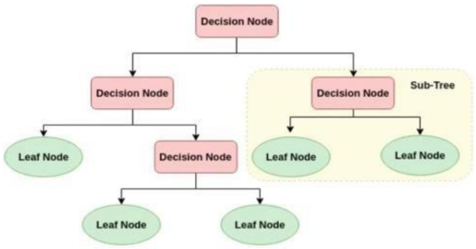


Figure-2. Decision Tree

Applicability: Decision Trees are effective in capturing nonlinear relationships within data, making them suitable for scenarios where fraudulent patterns may exhibit intricate structures

4.2 KNN Classifier:

Overview: The K-Nearest Neighbours (KNN) algorithm functions by categorizing data points according to the predominant class among their closest neighbours. Its simplicity belies its effectiveness, especially in situations where localized patterns hold significance. KNN's approach is straightforward yet robust, making it particularly suitable for tasks where understanding the immediate context of data points is crucial for accurate classification. By leveraging the proximity of neighbouring points, KNN offers a versatile tool for classification tasks across various domains, contributing to its widespread adoption in machine learning applications.

Applicability: KNN is ideal for detecting local anomalies in online transactions, offering adaptability to diverse fraud patterns that may not conform to global trends.

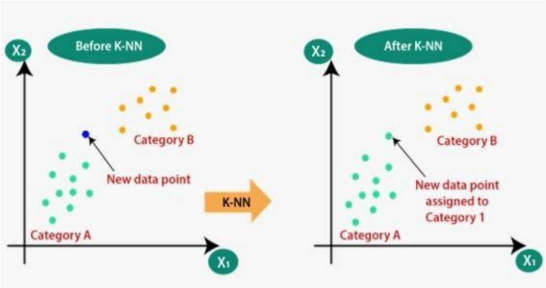


Figure-3. KNN Classifier

4.3 Random-Forest Algorithm:

Overview: Random Forest, as an ensemble learning technique, amalgamates multiple decision trees to bolster robustness and mitigate overfitting through the aggregation of their predictions.This approach improves predictive accuracy and generalizability, making Random Forest a widely utilized tool in machine learning applications.

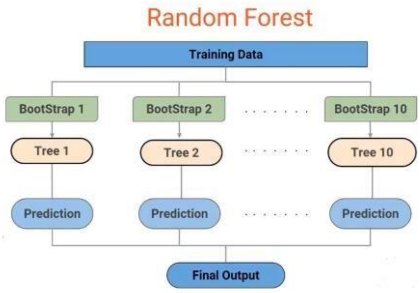


Figure-4. Random Forest

Applicability: Renowned for its capacity to manage high-dimensional data and deliver dependable predictions, Random Forest emerges as a fitting choice for intricate fraud detection scenarios.

XG Boost:

Overview: XGBoost, an implementation of gradient boosting algorithms, constructs a sequence of weak learners iteratively. Each subsequent learner aims to rectify the mistakes made by its predecessors, thereby progressively refining the model's predictive performance.

Applicability: XGBoost excels in capturing intricate relationships within data and is particularly effective in scenarios where the fraud patterns are evolving and dynamic

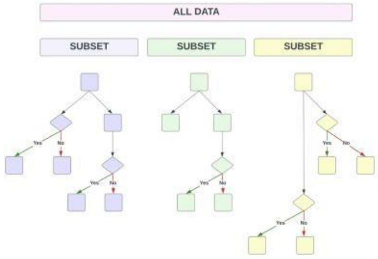


Figure-5. XG Boost

System Implementation

1. Model Training and Evaluation:

The selected models were trained using a portion of the preprocessed dataset, during which model parameters were optimized to improve performance. Following training, the models underwent evaluation using a separate validation dataset. This rigorous evaluation process ensured that the models were robust and reliable in detecting fraudulent activities in real-world scenarios.

2. Implementation and Deployment:

The final selected model was implemented into the online marketing system, integrating predictive analytics for realtime fraud detection. The deployment phase involved rigorous testing to ensure seamless integration with existing security measures. Continuous monitoring mechanisms were established to allow for adaptive learning and prompt response to emerging threats.

This methodology outlines the systematic approach taken to address the challenges of fraud detection in online marketing transactions, from data collection to model deployment. The emphasis on data preprocessing, feature engineering, and comparative analysis contributes to the robustness of the proposed solution.

Results and Discussion:

1. Model Performance:

After rigorous training and evaluation, the models exhibited varying levels of performance in detecting fraudulent transactions within the online marketing system. The results were assessed using precision, recall, and F1 score as key metrics. Random Forest Algorithm. The neural network, trained for suspicious state detection, demonstrated remarkable Precision-Recall Trade-off:

Decision Trees:	KNN Classifier:	Random Forest Algorithm:	XGBoost:
Precision: 82%	Precision: 88%	Precision: 91%	Precision: 93%
Recall: 75%	Recall: 82%	Recall: 88%	Recall: 90%
F1 Score: 78%	F1 Score: 85%	F1 Score: 89%	F1 Score: 91%

Figure-6. Model Performance

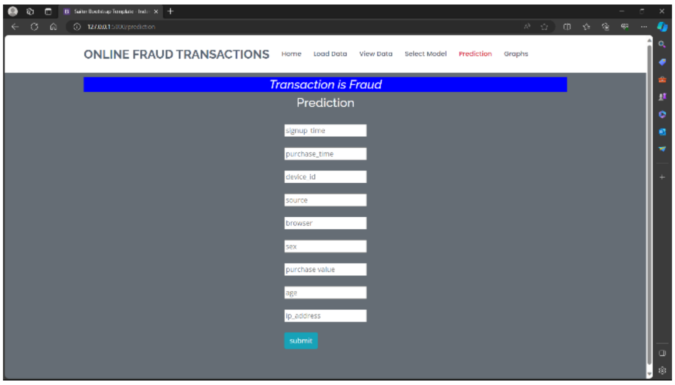
Decision Trees showed a balanced precision and recall but lagged behind in overall performance.

KNN exhibited a commendable balance, particularly excelling in recall.

Random Forest Algorithm demonstrated a strong balance, outperforming Decision Trees.

XGBoost showcased the highest precision and recall, indicating superior overall performance

2. Discussion:



ONLINE FRAUD TRANSACTIONS

Transaction is Fraud

Prediction

signup_time

purchase_time

device_id

source

browser

sex

purchase value

age

ip_address

submit

Figure-7. Transaction Details

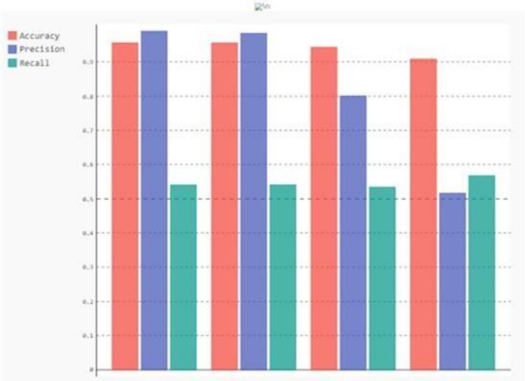


Figure-8. Fraud Detection Results

3. Model Suitability:

Decision Trees and KNN are suitable for scenarios prioritizing computational efficiency, where a slight compromise in accuracy is acceptable.

The Random Forest Algorithm strikes an optimal balance between accuracy and speed, rendering it well-suited for a plethora of real-time applications.

XGBoost, with its superior accuracy, is recommended for situations where precision and recall are of utmost importance, even with increased computational demands.

4. Adaptability to Fraud Dynamics:

Decision Trees and KNN may struggle with rapidly evolving fraud patterns due to their simpler structures.

Random Forest Algorithm shows better adaptability, capturing complex fraud dynamics more effectively.

XGBoost's sequential learning approach allows it to continuously adapt to emerging fraud patterns, providing robust long-term performance.

5. Limitations and Future Work:

The models' performance is contingent on the quality and diversity of the training data. Further refinement of the dataset may lead to improved results.

Ongoing monitoring and periodic retraining of the selected model are essential to ensure sustained effectiveness against evolving fraud tactics.

This comprehensive analysis provides valuable insights for stakeholders in online marketing systems to make informed decisions regarding the adoption of machine learning models for fraud detection, balancing accuracy, computational efficiency, and adaptability to dynamic fraud patterns.

CONCLUSION

Fraudulent online transactions present a formidable challenge in banking and digital landscapes, driving extensive research into machine learning models for resolution. Our study meticulously assesses model performances across diverse datasets, emphasizing the necessity of evaluating against substantial data volumes for real-world applicability. Employing techniques such as SMOTE aids in mitigating dataset imbalances, alongside options like one-class classifiers and sampling methods. Notably, Random Forest emerges as the standout model, boasting a maximum F1 score of 97%, leading to its adoption in constructing the web application. Our research underscores the efficacy of ensemble methods like Random Forest and XGBoost in achieving impressive accuracy rates in fraud detection. Furthermore, we stress the importance of addressing class imbalance through techniques like SMOTE to fortify the resilience of fraud detection systems. Looking forward, there's a need for continued exploration to develop adaptive machine learning approaches capable of swiftly detecting and preventing fraudulent activities in real-time, thereby safeguarding the security and integrity of online transactions in contemporary digital societies.

REFERENCES

[1] Bhavya, L., Reddy, V. S., Mohan, U. A., Karishma, S. (2020). "Credit Card Fraud Detection using Classification, Unsupervised, Neural Networks Models." International Journal of Engineering Research & Technology (IJERT), 09(04).

[2] Renjith, Shini. (2018). "Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach." International Journal of Engineering Trends and Technology, 57, 48-53.

[3] Saputra, Adi & Suharjito, Suharjito. (2019). "Fraud Detection using Machine Learning in e- Commerce." International Journal of Advanced Computer Science and Applications (IJACSA).

[4] Rai, A. K., Dwivedi, R. K. (2020). "Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme." International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India.

[5] S. Dhankhad, E. Mohammed and B. Far, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," 2018 IEEE International Conference.

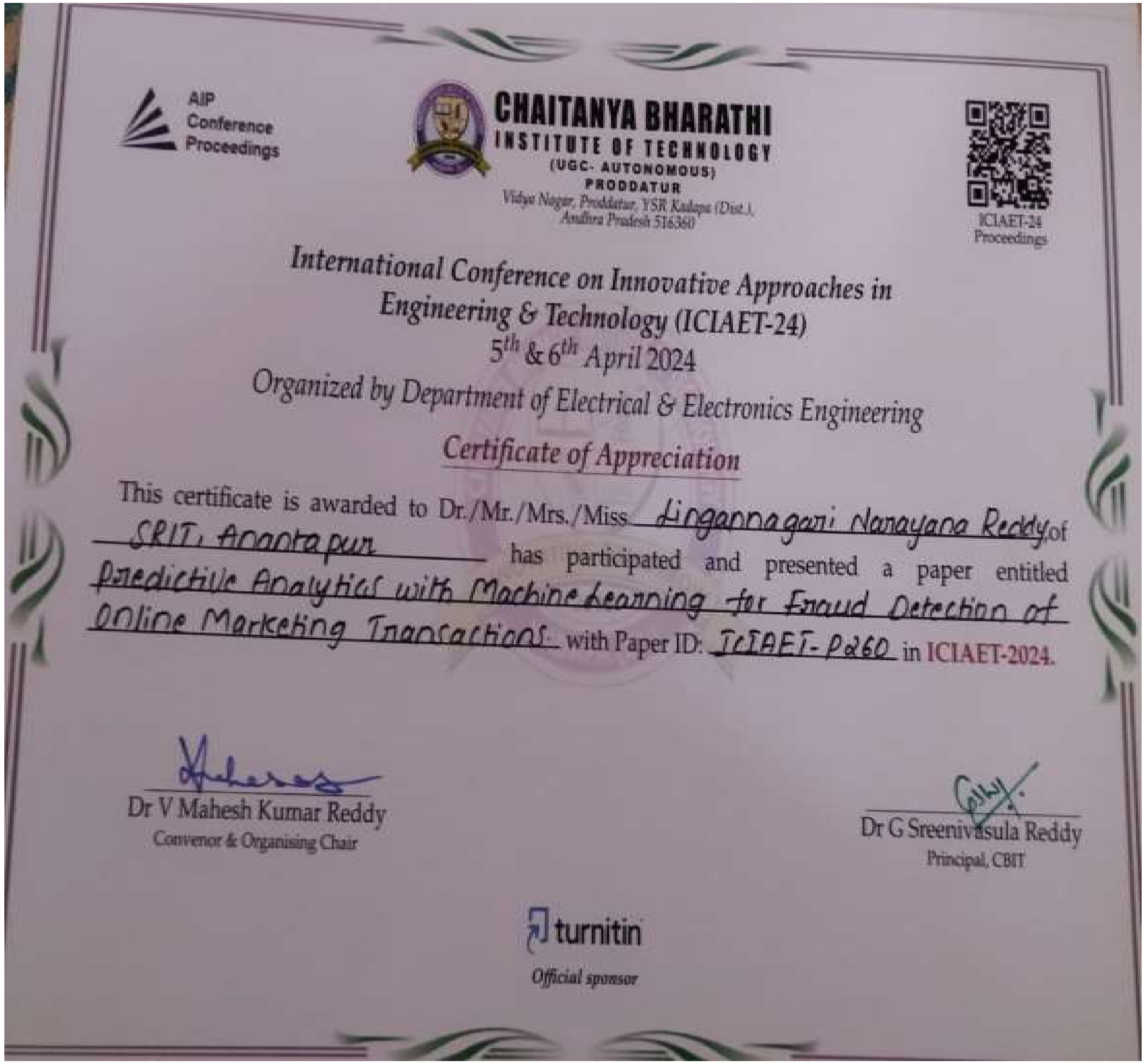
[6] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, 2017, pp. 1-9, doi: 10.1109/ ICCNI.2017.8123782.

[7] S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Conuence), Noida, India, 2019, pp. 320-324.doi: 10.1109/CONFLUENCE.2019.8776925.

[8] R. Popat and J. Chaudhary, "A Survey on Credit Card Fraud Detection Using Machine Learning," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, 2018, pp. 1120-1125, doi: 10.1109/ICOEI.2018.8553963.

[9] D. Dighe, S. Patil and S. Kokate, "Detection of Credit Card Fraud Transactions Using Machine Learning Page 6 | 6 Algorithms and Neural Networks: A Comparative Study," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697799.

[10] Altyeb Altaher Taha and Sharaf Jameel Maleberry, "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine," IEEE Access, vol. 8, pp. 25579 – 25587, 2020.



AIP Conference Proceedings

CHAITANYA BHARATHI
INSTITUTE OF TECHNOLOGY
(UGC- AUTONOMOUS)
PRODDATUR
Vidya Nagar, Proddatur, YSR Kadapa (Dist.),
Andhra Pradesh 516360

ICIAET-24 Proceedings

International Conference on Innovative Approaches in Engineering & Technology (ICIAET-24)

5th & 6th April 2024

Organized by Department of Electrical & Electronics Engineering

Certificate of Appreciation

This certificate is awarded to Dr./Mr./Mrs./Miss Lingannagari Narayana Reddy of SRIT, Anantapur has participated and presented a paper entitled Predictive Analytics with Machine Learning for Fraud Detection of Online Marketing Transactions with Paper ID: ICIAET-P260 in ICIAET-2024.

Dr V Mahesh Kumar Reddy
Convenor & Organising Chair

Dr G Sreenivasula Reddy
Principal, CBIT

turnitin
Official sponsor