Data Science
Project Report

Problem Statement :

Fraud Detection in Banking System

Introduction :

i) The Indian banking sector has experienced significant growth and changes since liberalization of economy in 1991.

ii) Though the banking industry is generally well regulated and supervised, the sector suffers from its own set of challenges when it comes to ethical practices, financial distress and corporate governance.

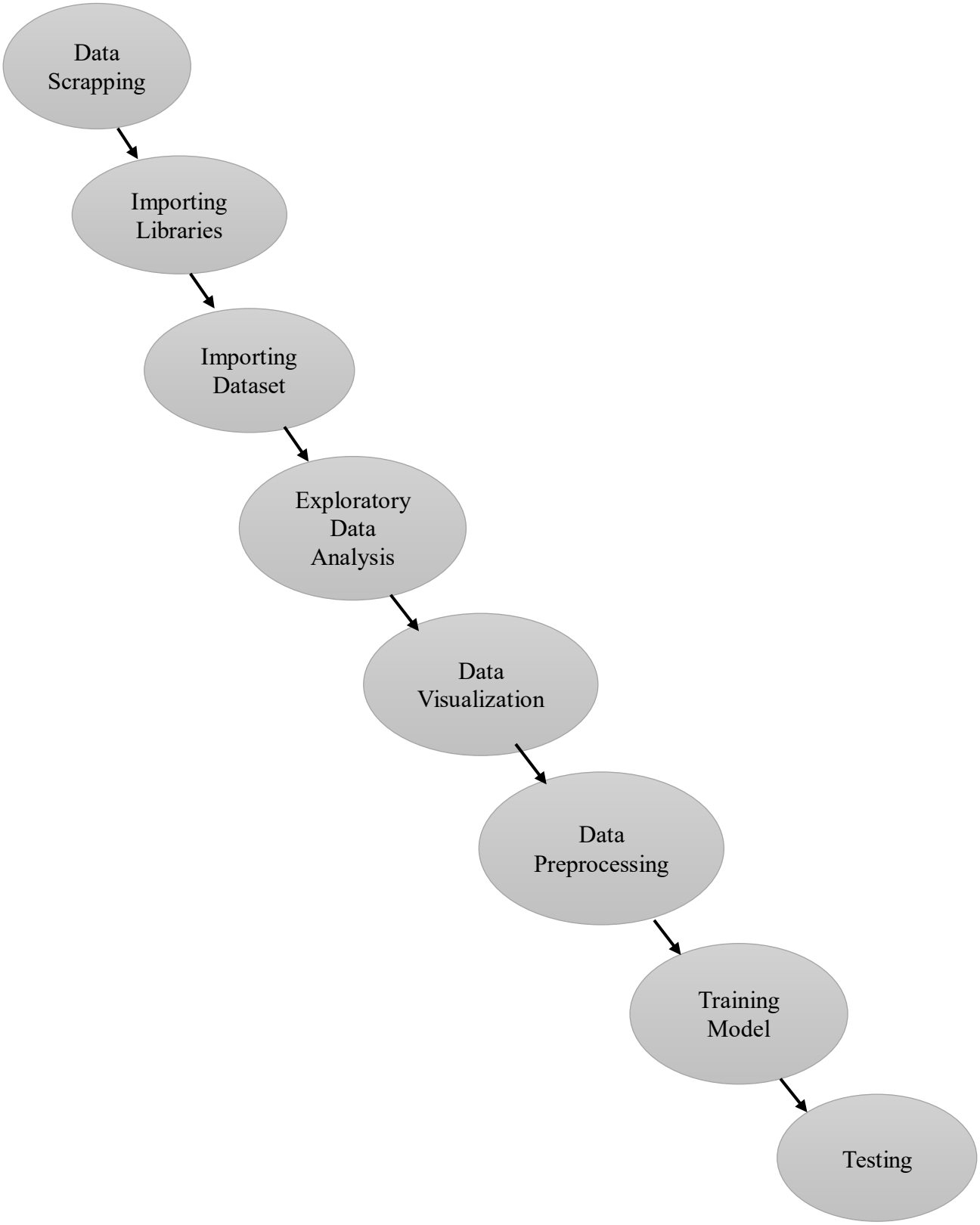
iii) In recent years, instances of financial frauds have regularly been reported in India. Although banking frauds in India have often been treated as cost of doing business, post liberalization the frequency, complexity and cost of banking frauds have increased manifold resulting in a very serious cause of concern.

iv) Hence an attempt need to be made to solve this issue by providing a proper mechanism of detecting frauds.

Objectives :

- To implement machine learning algorithms to detect fraudulent transactions.
- The main objective of this project is to reduce losses due to payment fraud for both merchants and issuing banks and increase revenue opportunities for merchants.

Project Workflow: Steps followed during the project are as follows:



Data Scrapping:

i) Data scraping, in its most general form, refers to a technique in which a computer program extracts data from output generated from another program.

ii) Data scraping is commonly manifest in web scraping, the process of using an application to extract valuable information from a website.

iii) Data scraping, also known as web scraping, is the process of importing information from a website into a spreadsheet or local file saved on your computer.

iv) There are various techniques to scrap data from web. We have scrapped data using BeautifulSoup.

Importing necessary library dependencies:

i)Required machine learning libraries are imported in this step.

ii) The libraries that are imported in this project are as follows:

a) **pandas :** A Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive.

b) **numpy :** To perform a wide variety of mathematical operations on arrays

c) **seaborn :** A Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

d) **matplotlib :** a comprehensive library for creating static, animated, and interactive visualizations in Python.

e) **sklearn :** Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Importing Dataset:

Scrapped Dataset can be imported by reading and loading it from the CSV file.

Data Preprocessing:

i)Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model.

ii) It is the first and crucial step while creating a machine learning model.

iii)Data preprocessing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, greater is the reliance on the produced results. Incomplete, noisy, and inconsistent data are the properties of the huge banking system datasets.

iv)Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise and resolving inconsistencies.

v) As part of data preprocessing, we have

- dropped zipcode and zipMerchant since they have only one unique value
- converting object columns type to categorical for easing the transformation process
- converting categorical values to numeric values

Data Visualization:

i)Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

ii) In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

iii) In this project we have visualized the data with the help of count plot, heatmap, box plot.

Models Implementation:

For detecting fraudulent transactions in banking system data we have used two models:

i)Random Forest Classifier

ii)Support Vector Machine

The purpose behind using different models is to check the relative performance of different models and selecting the best model based on its accuracy score.

Testing:

Sr .No	Model	Score
i)	Random Forest Classifier	99.62
ii)	Support Vector Machine	99.05

On evaluating the models we understood that,

Accuracy score(Random Forest Classifier) > Accuracy score(Support Vector Machine)

Hence, the best model for detecting frauds in banking system is Random Forest Classifier.

References:

i)https://www.miteksystems.com/blog/how-does-machine-learning-help-with-fraud-detection-in-banks

ii)https://medium.com/engineered-publicis-sapient/fraud-detection-in-banking-industry-and-significance-of-machine-learning-dfd31891a0b4