# Intelligent E-Commerce Analytics: Fraud Detection and Insights through Machine Learning

***Abstract- The project's foundation includes sentiment analysis, machine learning models based on transaction fraud detection, and restaurant evaluations pertaining to e-commerce. The goal of the project is to use forecasts to enhance company decision-making by utilizing data sources such as consumer feedback and transactions. Methods under consideration as core methods include Naive Bayes, Random Forests, Regressions, KNN, K-Means, and the Apriori algorithm to identify association rules for consumption. Although the KNN approach offers a nonlinear forecast, Random Forest and Naive Bayes are appropriate for sentiment grouping and, consequently, ratings and product attributes. A K-Means is an unsupervised approach, it was able to detect high-risk transactions and is required in fraud discovery. Major findings reveal that although it has higher accuracy for classification and prediction jobs, difficulties like interpretability in Random Forest and comparatively high computational cost in KNN are some of the drawbacks that need to be improved. Future research could involve integrating AI and improving estimation to detect fraud more accurately based on features. The study's conclusions can be applied to improve cost analysis, aberration control, and customer happiness in dining establishments and online business settings.***

*Keywords- Classification Method, Regression Method, KNN, Random Forest, Unsupervised algorithms, Naive Bayes, K-means clustering, text analysis, Apriori Algorithm.*

## I. INTRODUCTION

### *A. Motivation of the work*

Understanding consumer behavior and preventing fraud are crucial components of improving profitability and cost control in the competitive business world of today. The project's primary goal is to apply various machine learning algorithm and data mining techniques towards addressing some few difficulties which may affect commercial organizations, including sentiment analysis, fraud detection, cost calculation, and customer rating concerns. Thus, this work is invaluable for improving fraud detection and analysis of customer behavior to improve customers' financial situation and improve their overall satisfaction.

*Classification Method*

A supervised learning technique called classification uses labelled data to teach an algorithm how to categorise unknown data into predetermined classes. Like how product categories are classified, Naïve Bayes along with random forest which is utilized within such learning to determine sentiment categories of customer reviews.

*Regression Method:* The regression method is another machine learning approach used to categorize dependent variables according to their input factors. Linear Regression and KNN models are used in this project to predict costs for each considerable numerical data from customer ratings and product-related features.

*Text Analytics:* The project focuses on text analysis which is crucial where customer feedback is assessed with a view of classifying it as either positive, negative, or even neutral. By using machine learning models, the changes in customer feedback can be directly identified and businesses can address the issues at the soonest time possible and improve the customer experience in the process.

*Unsupervised Machine Learning:* Cluster analysis involves sorting similar data based on a distance or similarity measure. One of the methods used in unsupervised learning, K-Means Clustering.

*Rule Mining Association:* During information discovery, rule mining association is a technique which can be drawn to specifically find correlations between variables in the dataset at hand. This is often done with the help of the Apriori Algorithm that finds frequent itemsets relevant to a certain support.

### *B. Research Objectives*

*Research Objective 1: To assess customer ratings and related characteristics regarding entities and put them into certain categories.*

The purpose of evaluating customer ratings and similar attributes is to segment entities about customers' preferences, opinions, and actions. Business organizations can discover patterns that associate ratings generated by customers with attributes of products or services including quality, services offered, and prices.

*Research Objective 2: To predict cost values using customer ratings as well as any other numerical value that may be associated with an item.*

The objective involves making estimates of cost using data such as customer ratings together with other quantitative data like price, quantity, or features.

*Research Objective 3: To analyse sentiment and categorise the customer reviews to gain an understanding of them.*

The determination for this inspection objective is to distinguish consumer feedback into three categories: neutral, negative, and positive attitudes using sentiment analysis techniques. This approach is helpful since it may help businesses achieve greater insight into the perceived value their products or services have to customers.

*Research Objective 4: To cluster transactions to discover possible anomalies in terms of temporal patterns.*

The purpose of this objective relates to the identification of time-related patterns in transaction grouping to identify symptoms of anomalous behaviour. Holders of business can therefore use the temporal analysis of transactions to detect the outliers, do fraud detection, and more importantly comprehend the customer purchasing

pattern over the period, thus enhancing the decision and risk management.

*Research Objective 5: To identify correlations within or between the attributes of transactions for improved fraud identification.*

One of the essential objectives is because it focuses on intra and inter-attribute dependencies of attributes like location, amount, time, and customer history to find out co-relation which may indicate fraud.

*C. Research Questions*

- *Research Question 1: What are the main variables involved in the determination of the nature of customer ratings?*
- *Research Question 2: How can regression modeling improve cost prediction results?*
- *Research Question 3: What influences the difference in the sentiment of customer feedback concerning different product categories?*
- *Research Question 4: Which clusters suggest high-risk or suspicious behavior?*
- *Research Question 5: Which transaction attributes have been identified as relevant to the presence of fraud most of the time?*

*D. Overview of the sections*

The project revolves around three data sets such as customer ratings, transaction fraud detection, and sentiment analysis to compare some of the major machine learning approaches.

*Introduction:* One of the topics of this project is the introduction, which is followed by the study's goals and research questions. It outlines the paper's main goal, which is to help with decision-making by finding trends and imputations for values associated with customer reviews, ratings, and transaction data.

*Related Work:* The previous research in the areas of customer sentiment analysis, prediction models, and transaction data clustering is examined and critically addressed based on the associated work section. A review of previous uses of the datasets and approaches applied to them focuses on the identified improvements and novelty of the present work.

*Data Mining Methodology:* The methodology adopted in data mining is described together with the use of a life cycle such as CRISP-DM in the research.

*Evaluation:* The methodology section explains how the effectiveness of the models and methodologies used was assessed.

*Conclusions and Future Work:* The study's findings are summarised in the conclusions, which also address the research issues that were examined. The section also provides brief conclusions of the main findings of the study in relation to customer knowledge enhancement, better predictive techniques, and fraud-identifying approach.

## II. RELATED WORK

*A. Critical Evaluation of key related works*

The paper conducts a critical evaluation based on relevant brochure about some operations which include predictive analysing models, like K-Nearest Neighbours, Naïve Bayes, Linear Regression, Random Forest for transactional data analysis, cost prediction, and consumer sentiment detection. The two classification models are established for analysing purposes like Naive Bayes, and Random Forest [8]. Naive Bayes is simple and efficient, it can be very useful for sentiment analysis in customer reviews, particularly when data is of high dimensionality. Naive Bayes considers each feature to be independent of others thereby, enabling it to process even large-scale sets in real quick time. The essential limitation is in the undertaking of quality freedom, that might not be granted for real-world customer review datasets.

The ensemble learning method known as the Random Forest application aggregating outputs of several decision trees is believed to be among the most powerful models for classification tasks, especially with structured data focused on customer ratings or transaction attributes [9]. Random Forest models can be very expensive as the number of decision trees grows and are usually considered less interpretable compared to simpler models, which may reduce their utility in some business contexts where model interpretability is paramount. Linear Regression has been successfully applied for cost value predictions based on customer ratings based on the regression analysis for analysing customer sentiment [10].

Regression Model Using K-Nearest Neighbours (KNN) Another popular regression model that works well for predicting continuous values in situations where feature interactions are intricate and nonlinear is K-Nearest Neighbours. [7]. KNN is excellent at identifying any kind of local pattern since it operates by averaging the output of the closest proximate in the feature space. KNN is a non-parametric technique and thereby Linear Regression, assumes less about the data [11]. As the dimensions of the features increase, generally the performance decreases with the increase in sparsity of the dataset.

*B. Limitations*

The applications of every model whether it is a classification model or a regression framework. Machine learning techniques like Random Forest, Linear Regression, K-Nearest Neighbours, Navie Bayes have limitations when it comes to copy observation and cost estimation prediction.

*Naive Bayes*

Due to its effectiveness as a classification model, Naive Bayes is used for a variety of text classification tasks, such as sentiment analysis of customer reviews. The main drawback of Naive Bayes is the independence assumption on different features [8]. The Naive Bayes classifier does not take these interdependencies into account, sometimes with suboptimal performances, especially in those cases when feedback provided by a customer includes Den lymphatic expressions. The Naive Bayes classifier has no way of accounting for this interdependency and will misclassify examples.

*Random Forest Classifier*

To increase forecast accuracy, the Random Forest ensemble learning technique builds several decision trees and then aggregates their output. When Random Forest is used to train on datasets with a lot of trees or characteristics, it can be very sluggish. In terms of cost prediction or fraud detection in real time, this can provide a significant bottleneck [12]. One of

the main issues with random forests is that they are frequently viewed as a "black box" model, making it challenging to interpret the outcomes. For example, this would be something very important to show in fraud detection or customer behavior analysis, making a prediction.

*Linear Regression*

Linear regression is usually the most standard approach for continuous variable prediction, whether it is a cost prediction or sales forecasting. The presumption of a linear relationship between the independent and dependent variables is one of the primary drawbacks of linear regression. ***Linear Regression*** can also face serious problems of multicollinearity when different independent variables start showing high correlations among themselves [13]. Another limitation is that it is sensitive to outliers, whereby a few results might skew it much in the case of datasets with erroneous or extreme values.

*K Nearest Neighbors*

Regression and classification tasks can be handled using the non-parametric KNN model. One issue is known as the "curse of dimensionality," which occurs when a dataset's feature space grows. This causes the distances between data points to become less significant, making it more difficult for algorithms to identify the most pertinent neighbors [14]. KNN is computationally expensive during the prediction phase and distance needs to be calculated from the test point to all other training points within a dataset. KNN is sensitive to irrelevant features, which may distort the distance calculation and affect model accuracy.

*C. Previous uses of the datasets and methods*

Customer ratings, transactional fraud detection, and sentiment analysis datasets explored in this report and widely used within different domains to grasp consumer behavior, predict costs, and detect fraudulent activities. In both commercial and research settings, alternative machine learning models that use Naive Bayes, Random Forest, Linear Regression, and K-Nearest Neighbours show noteworthy results [15]. The two variables are very important in providing businesses with information on areas they need to improve their services [1]. Prior research has used customer review data to create sentiment analysis models that categorize reviews as neutral, negative, or positive.

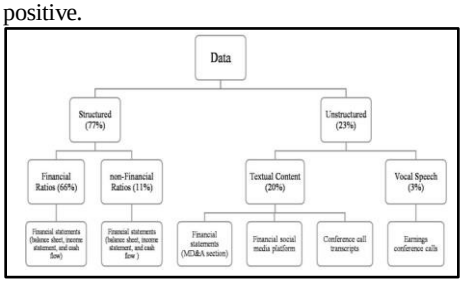



Fig. 01: Flowchart on the categorisation of the data types

Sentiment analysis, particularly through Naive Bayes and Random Forest, is one of the most common applications in product or service feedback analysis. Examples include the use of sentiment analysis in the comprehension of customer satisfaction to improve product offerings. Naive Bayes is famous not only for its simplicity and efficiency in text classification but also for its use for classifying reviews based on sentiment [4]. On the contrary, the Random Forest algorithm has been applied in the literature to high-dimensional data by integrating the outputs of distinct trees [5]. However, because of the complexity and the difficulty of interpreting results obtained from Random Forest models, it tends to remain less preferred when businesses require clear explanations for the trends within customer sentiment. Analyses of customer ratings have also included predicting the customers' behavior or future purchases considering the given ratings and other characteristics.

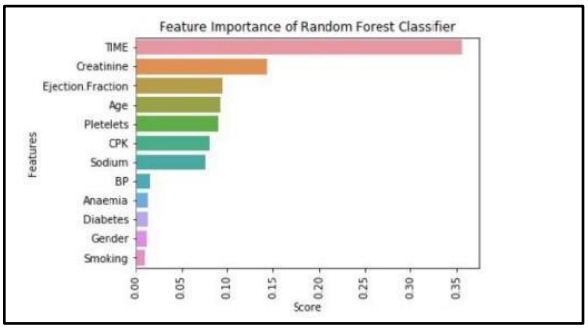



Fig. 02 Random Classifier of Fraud Detection

The integration of ratings and product features including quality, price, and service has been analyzed using classification and regression analysis [3]. Prior research has applied Linear Regression and KNN for estimating sales or costs related to customer purchasing behavior. Linear Regression has been applied to forecast continuous variables, including total sales in the future and cost of sales, given that the nature of relationships between buyer ratings and product attributes is presumed to be linear [2]. KNN has been used to predict other patterns such as the next product that will be purchased by a client, through the relative distances between the data points [6]. Fraud detection is perhaps the most important use of Machine learning for financial and e-commerce institutions. To identify fraud, a few derived studies applied clustering techniques, such as K-Means and Hierarchical Clustering, on the transactional data.

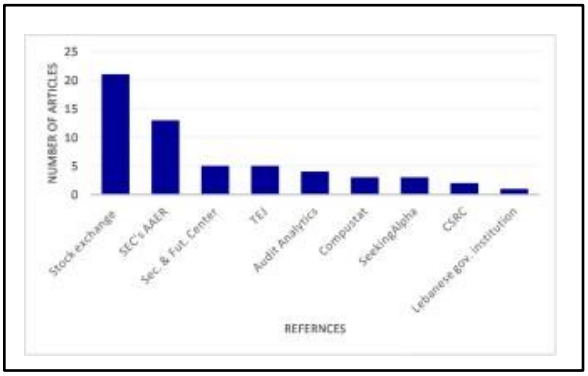



Fig. 03: Status regarding fraud detection status of Companies

Transaction data normally consist of such parameters as transaction amount, time, and place characteristics; customer behavior may include a temporal

aspect that reflects fraudulent transactions. In the banking sector, for instance, K-Means clustering has been used to classify credit card transactions into clusters that aid in the detection of credit card fraud by identifying the groups with fraud incidents. These methods were found to be useful in many scenarios, but they have some weaknesses, including the assumptions made about the distribution of data and the interpretability of results, and are the continuing subject of research to improve on them.

## III. DATA MINING METHODOLOGY

### *A. Adapted Approach*

CRISP-DM approach has been used for the analysis of data mining by following the research question. The six phases of this methodology business understanding, data understanding, data preparation, modelling, evaluation, and deployment are guided step-by-step. In the thorough process of implementing different data mining approaches, these stages are significant and beneficial. Five different methods have been integrated into the analysis to answer the research objectives such as classification, regression, text analysis, unsupervised machine learning, and association rule mining.

The classification phase involves predictive models, like Random Forest and Naive Bayes, being applied to data that has been categorized with predefined labels. Regression techniques, like Linear Regression and K-nearest neighbors, have been employed to model continuous variables and establish relationships within the data [16]. Text analysis has been done on unstructured textual data to extract understanding from it, thereby enabling feature extraction and sentiment analysis. Various unsupervised machine learning methods have been employed, including clustering, to identify trends and classify related data points. Association rule mining has been used afterward to find interesting relationships among the variables that can show the important co-occurrence patterns within a dataset. The CRISP-DM framework at its base, provided a sound method to answer the research questions and establish holistic insights by using a mix of various analysis techniques.

### *B. Datasets*

*1st dataset: Amazon Books Reviews*

The Amazon Books Reviews dataset is credited to Kaggle and presents customer reviews for various books listed on Amazon. Indeed, this dataset is made up of unstructured textual materials, including full reviews of the books, ratings, and metadata about the books. Some of the important attributes in the dataset include review text, rating, book title, and author. The use of this dataset was majorly associated with text analysis preprocessing, sentiment analysis, and topic modeling.

The different varieties of preprocessing text data by tokenization, stopword removal, and lemmatization were applied for meaningful pattern extraction and review sentiments. Customer sentiment prediction and thematic consumption of customer feedback have been done on review text by employing this project. Conjoint application of NLP algorithms would help gain insight into consumer preferences and trends within the book industry.

*2nd dataset: Zomato Restaurant Data*

The Zomato Restaurant Data is available in Kaggle; the dataset shows minute details of various restaurants listed in Zomato based on their ratings, location, type of cuisine, average cost, number of votes received, etc. Regression and classification analysis have been performed on the following dataset.

Predicting the rating category that is, rating bins such as 1-2, 2-3, 3-4, and 4-5 based on characteristics such restaurant kind, location, and average cost for two people is the classification challenge. The regression task was to predict a rating of restaurants using continuous variables of the same set of features. To work with these models, preprocessing procedures have involved addressing missing data, encoding categorical variables, and scaling numerical characteristics. Preprocessing steps have included scaling numerical characteristics, encoded categorical variables, and addressed missing data to operate with these models.

*3rd Dataset: Online Payment Fraud Detection*

To determine customer ratings, create predictions, and examine the factors influencing restaurant ratings and price, this dataset has been exposed to several classifiers, including Random Forest and Naive Bayes, as well as regression approaches, including Linear Regression and K-Nearest Neighbors. Third Dataset: Fraud Detection in Online Payments Additionally, the Kaggle dataset for Online Payment Fraud Detection was obtained. It contains transactional information on online payments, including user activity, payment type, and transaction amount. The labels of transactions that indicate whether they are fraudulent or not are the goal variable here. This dataset has been used to carry out unsupervised machine learning techniques, such as clustering, on transactions to track down unusual patterns that show fraud.

Association Rule Mining has also been carried out to extract the hidden relationships between various features, such as payment method and fraud likelihood. To get the data ready for analysis, data preprocessing procedures included encoding categorical variables, managing missing values, and feature scaling. The insights drawn from unsupervised learning models and association rules have been used to identify key factors in fraud detection within online payment systems.

### *B. Data Preprocessing and transformation*

*Classification and Regression*

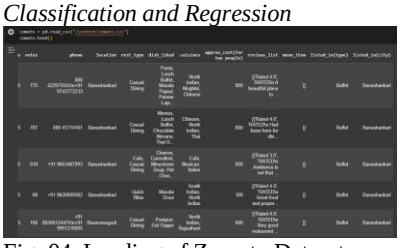


Fig. 04: Loading of Zomato Dataset

Loading of the zomato dataset depicts a sneak peek into the data columns, which range from URL, name, online_order, and many others. This provides a general understanding of the data's structure as well as the specifics of its properties.

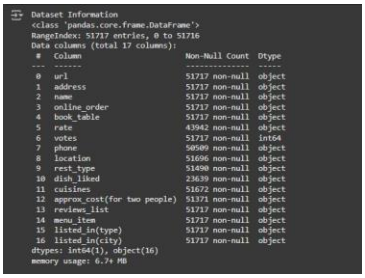



Fig. 05: Dataset Information

The Dataset Information includes several non-null entries, data types, and memory usage. The dataset information will be important in building an idea of the characteristics within this dataset and, probable problems with the data.


Fig. 06: Checking for Missing values

Checking for ***missing values*** shows common problems with real-world data sets. The result of this analysis will tell us which columns have missing values and may need further attention or imputation. The columns of the dataset contain missing values the columns such as rate, phones, location, rest_type, dish liked, and approx_cost from the dataset.

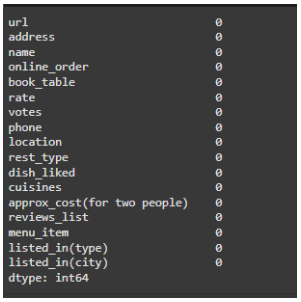



Fig. 07: After Checking Missing Values

The missing or null values have been eliminated from the dataset after the missing values have been checked. The accuracy of model performance is impacted by the existence of missing values in the dataset.

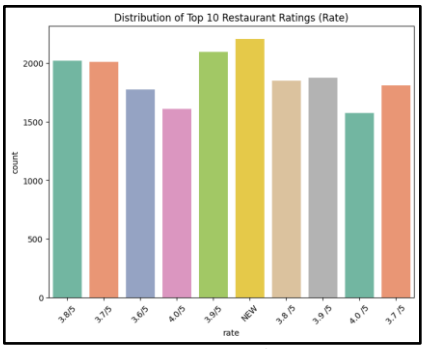



Fig. 08: Distribution of Top 10 Restaurant Ratings

Distribution of Top 10 Restaurant Ratings shows various rates for different restaurants. This kind of information helps in understanding the competition that exists between restaurants and which ones perform better than others.

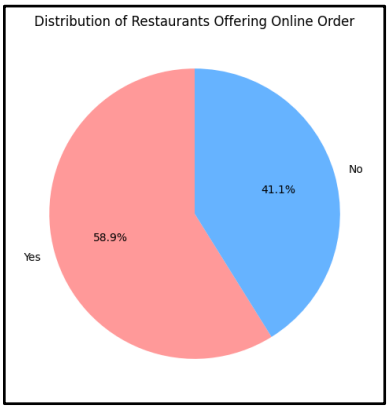



Fig. 09: Distribution of Restaurants Offering Online Order

The distribution of restaurants offering online ordering services is 58.9% is yes and 41.1% is no. That would mean more and more of them have joined digital ordering platforms, which again hints at business strategy and user experience optimisation.

*Text Analysis*

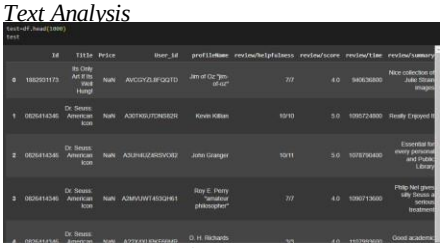


Fig. 10: Loading of book rating dataset

Title, price, user ID, profile name, review and helpfulness, review score, and review summary are just a few of the many attributes that are included in the loading of the book rating dataset. The data could clarify the structure and content of the book rating information.

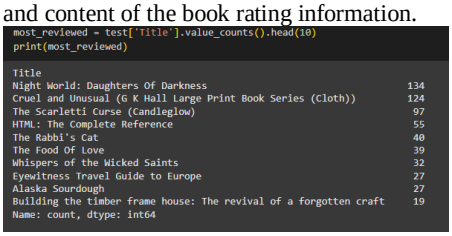



Fig. 11: Most Reviewed Title

The Most Reviewed Title showing books which occupy the top position according to the number of reviews. This could be helpful in determining the most popular or influential books present within the dataset.

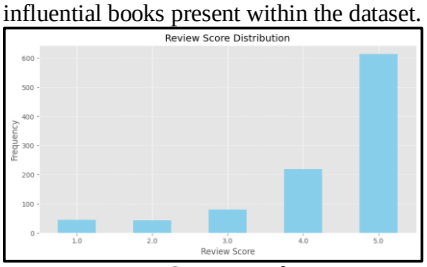



Fig. 12: Review Score Distribution

Review Score Distribution showa the reflected frequency of the review scores. During analysis it gets an idea of sentiment and rating distribution in general in these reviews.

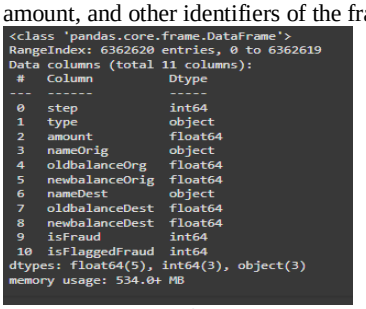
*Clustering and Apriori Algorithm*

```
[2] data = pd.read_csv('onlinefraud.csv')
    data.head()
```

| | step | type | amount | nameOrig | oldbalanceOrg | newbalanceOrig | nameDest | oldbalanceDest | newbalanceDest | isFraud | isFlaggedFraud |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | PAYMENT | 9839.64 | C1231006815 | 170136.0 | 160296.36 | M1979787155 | 0.0 | 0.0 | 0 | 0 |
| 1 | 1 | PAYMENT | 1864.28 | C1666544295 | 21249.0 | 19384.72 | M2044282225 | 0.0 | 0.0 | 0 | 0 |
| 2 | 1 | TRANSFER | 181.00 | C1305486145 | 181.0 | 0.00 | C553264065 | 0.0 | 0.0 | 1 | 0 |
| 3 | 1 | CASH_OUT | 181.00 | C840083671 | 181.0 | 0.00 | C38997010 | 21182.0 | 0.0 | 1 | 0 |
| 4 | 1 | PAYMENT | 11668.14 | C2048537720 | 41554.0 | 29885.86 | M1230701703 | 0.0 | 0.0 | 0 | 0 |

Fig. 13: Loading of Fraud Dataset

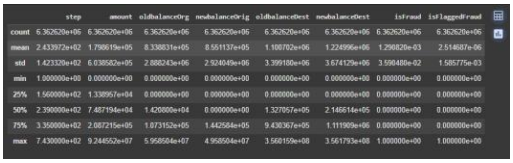
Loading of Fraud Dataset shows the structure of the data. It contains information related to step, type, amount, and other identifiers of the fraud transactions.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6362620 entries, 0 to 6362619
Data columns (total 11 columns):
 #   Column          Dtype
---  ------          -----
 0   step            int64
 1   type            object
 2   amount          float64
 3   nameOrig        object
 4   oldbalanceOrg   float64
 5   newbalanceOrig  float64
 6   nameDest        object
 7   oldbalanceDest  float64
 8   newbalanceDest  float64
 9   isFraud         int64
 10  isFlaggedFraud  int64
dtypes: float64(5), int64(3), object(3)
memory usage: 534.0+ MB
```

Fig. 14: Dataset Information

The Dataset Information of the fraud dataset, including the columns of data, the data types, and memory usage. This could help in understanding the nature and structure of the fraud data.



| | step | amount | oldbalanceOrg | newbalanceOrig | oldbalanceDest | newbalanceDest | isFraud | isFlaggedFraud |
|---|---|---|---|---|---|---|---|---|
| count | 6.362620e+06 | 6.362620e+06 | 6.362620e+06 | 6.362620e+06 | 6.362620e+06 | 6.362620e+06 | 6.362620e+06 | 6.362620e+06 |
| mean | 2.433972e+02 | 1.798619e+05 | 8.338831e+05 | 8.551137e+05 | 1.100702e+06 | 1.224996e+06 | 1.290820e-03 | 2.514687e-06 |
| std | 1.423320e+02 | 6.038582e+05 | 2.888243e+06 | 2.924049e+06 | 3.399180e+06 | 3.674129e+06 | 3.590480e-02 | 1.585775e-03 |
| min | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| 25% | 1.560000e+02 | 1.338957e+04 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| 50% | 2.390000e+02 | 7.487194e+04 | 1.420800e+04 | 0.000000e+00 | 1.327057e+05 | 2.146614e+05 | 0.000000e+00 | 0.000000e+00 |
| 75% | 3.350000e+02 | 2.087215e+05 | 1.073152e+05 | 1.442584e+05 | 9.430367e+05 | 1.111909e+06 | 0.000000e+00 | 0.000000e+00 |
| max | 7.430000e+02 | 9.244552e+07 | 5.958504e+07 | 4.958504e+07 | 3.560159e+08 | 3.561793e+08 | 1.000000e+00 | 1.000000e+00 |

Fig. 15: Descriptive Statistics

The descriptive statistic for every important variable in the dataset, including count, mean, standard deviation, minimum, and percentiles. The information will help in finding the general distribution and characteristics of the data.

## IV. EVALUATION

### *A. Evaluation of methodology*

*Classification Models*

Random Forest: Several decision trees are used in the Random Forest ensemble learning model to improve prediction accuracy without overfitting. Therefore, ***Random Forest*** appears preferable for usage because it can better recognize various properties of data by means of combining results obtained through several decision trees [18].

Naive Bayes: Based on the Bayes theorem, this model is quick and accurate and works well with text or categorical data because its predictors are not dependent on the target predictors [19]. Naive Bayes tends to work well particularly when using the assumption of independence is reasonable best.

*Regression Models*

Linear Regression: The research design involves straight-line relationships among variables or in some cases simply to find the minimum of the sum of squared residual [17]. Linear Regression is used when dependent variables are continuous; it offers interpretations of the importance of these variables to the dependent variable.

K-Nearest Neighbors (KNN): The KNN uses the average cases of the K-nearest data as its prediction, much like any other non-parametric model. KNN is great when using regression problems since the nature of problems changes and it can be useful in cases when the connection between variables is not linear.

*Unsupervised Machine Learning and Association Rule Mining*

K-means Clustering: K-means clustering classified the data into three clusters. The first step in the choice of the methodology was data scaling through StandardScaler to bring the dataset to a normal scale. The quality of cluster formation was assessed based on the losses Within-Cluster Sum of Squared Errors (WSSSE) for compactness, and the Davies-Bouldin index, according to the overall performance [20].

Principal Component Analysis (PCA): The PCA method was used to transform clusters into two dimensional for easy visualization to check on the actual group formation in a reduced feature space. As results showed, this dimensionality reduction technique was useful for increasing interpretability without sacrificing much information.

Association Rule Mining with Apriori: Designed for finding the most often occurring combinations of items in transactional databases. Using the Apriori algorithm, the measure of association between categorical variables was observed for transaction types. Frequent itemsets determined with minimum support of 0.01 to put into consideration the different common fraud type and transactions.

### *B. Choice of performance measures*

*Classification Models*

Random Forest and Naïve Bayes are primarily utilised for classification issues, which means that they are employed when the variable that needs to be predicted falls into a specific group. Usually, the following metrics are used to assess these models' performance: The easiest metric to compute is accuracy, which is the ratio of accurate predictions to actual predictions.

F1-Score, Precision, and Recall: The latter is more useful when dealing with data sets that fall into unbalanced classifications. Recall shows how successfully all positive data are detected, Precision determines the accuracy rate of real positive predictions, and the F1-score strikes a balance between Precision and Recall values.

Confusion Matrix: Compared to the accuracy score, the breakdown of true positive, false positive, true negative, and

false negative data provides a superior assessment for categorizing issues.

ROC-AUC: The Receiving Operating Characteristic curve and its globalization, the AUC is used when comparing models at various thresholds.

*Evaluation Metrics*

$R^2$: A measure of model fitness is the amount of variation in the dependent variable that can be accounted for by the independent variables.

Because it is expressed in the same unit as the target variable, the standard error of mean, or RMSE (Root Mean Squared Error), is a highly interpretable measure of the mean square error or the typical magnitude of the variance from the mean of the data set used in the model.

MAPE (Mean Absolute Percentage Error): MAPE is expressed in percentage defines prediction accuracy and makes the comparison with other models as a way of measuring how far off from the correct values the predictions are, in percentage.

*C. Results*

```
Random Forest Classifier:
Accuracy: 0.8937536485697607
Classification Report:
              precision    recall  f1-score   support

         2-3       0.77      0.71      0.74       207
         3-4       0.90      0.91      0.91      1794
         4-5       0.90      0.90      0.90      1425

    accuracy                           0.89      3426
   macro avg       0.86      0.84      0.85      3426
weighted avg       0.89      0.89      0.89      3426
```

Fig. 16: Random Forest Classifier

The Random Forest Classifier's classification report, broken down into the many classes used in the classification, precision, recall, f1-score, and support. These can all provide insight into how well the model is performing in predicting the right target classes.

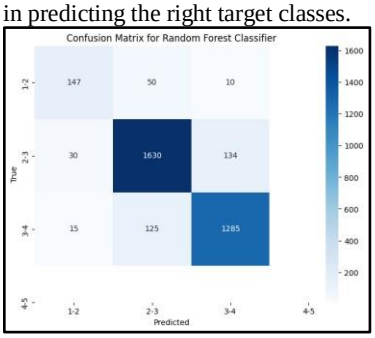



Fig. 17: Matrix of Random Forest Classifier Confusion

The Random Forest Classifier's Confusion Matrix runs the model on both correctly and poorly classified test dataset instances. It is a matrix that will help in determining the specific reasons behind the model's poor or good performance.

```
Naive Bayes Classifier:
Accuracy: 0.6199649737302977
Classification Report:
              precision    recall  f1-score   support

         2-3       0.00      0.00      0.00       207
         3-4       0.59      0.94      0.72      1794
         4-5       0.81      0.30      0.44      1425

    accuracy                           0.62      3426
   macro avg       0.46      0.42      0.39      3426
weighted avg       0.64      0.62      0.56      3426
```

Fig. 18: Classifier Using Naive Bayes

To prepare, the Naive Bayes Classifier report displays the same performance indicators as the Random Forest Classifier report. This makes it possible to compare the two model's capabilities directly.

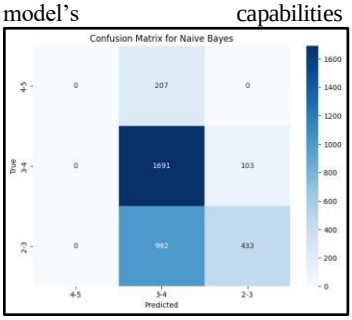



Fig. 19: Naive Bayes Confusion Matrix

The Naive Bayes Classifier's performance is displayed in the Confusion Matrix. The relative advantages and disadvantages of the two theories may be suggested by comparing the confusion matrices.

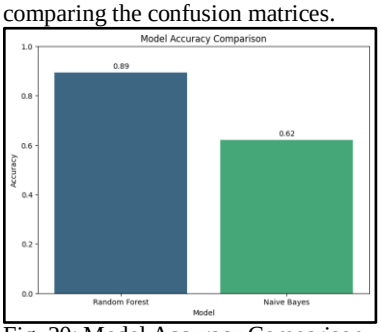



Fig. 20: Model Accuracy Comparison

Model Accuracy Comparison displays the models' overall accuracy. A comparison is made between Random Forest and Naive Bayes. This may lead to easier detection as to which model has outperformed the other and make an educated decision on which model best fits based on accuracy rate.

```
Linear Regression Model:
R² Score: 0.0862
Root Mean Squared Error (RMSE): 0.3998
Mean Absolute Percentage Error (MAPE): 0.0817

K-Nearest Neighbors Regression Model:
R² Score: 0.5388
Root Mean Squared Error (RMSE): 0.2840
Mean Absolute Percentage Error (MAPE): 0.0465
```

Fig. 21: Regression Models performance values

Comparison of two regression models such as the Linear Regression and the other one is K-Nearest Neighbours. After determining the data's authenticity, it separates the data into training and testing sets. Three performance indicators, including Mean Absolute Percentage Error, Root Mean Squared Error, and *R*2 Score, are then used to train and assess the models using test data.

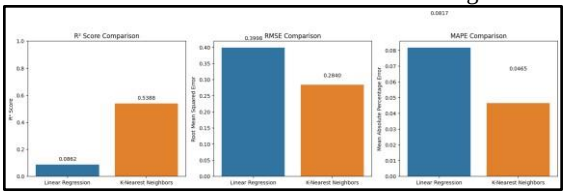



Fig. 22: Regression Model Comparison

Three distinguishing aspects of the K-Nearest Neighbors and Linear Regression models are compared: Mean Absolute Percentage Error [MAPE], Root Mean Squared Error [RMSE], and *R*2 Score. In the three sub-plots within each bar chart, raw values are shown with the scores noted above each bar.

*Text Analysis*

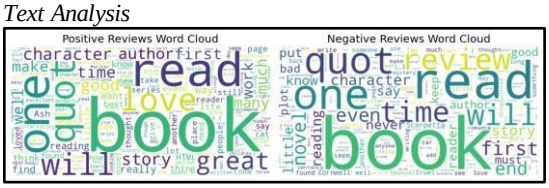



Fig 23: Word cloud for reviews, both good and bad

The most frequently used words in both positive and negative reviews are displayed in the Word Cloud for Positive and Negative Reviews. Review texts will provide insight into the sentiment that customers have expressed.

| | review/text | sentiment_textblob | sentiment_class_textblob |
|---|---|---|---|
| 0 | This is only for Julie Strain fans. It's a col... | 0.088889 | positive |
| 1 | I don't care much for Dr. Seuss but after read... | 0.214121 | positive |
| 2 | If people become the books they read and if "t... | 0.266271 | positive |
| 3 | Theodore Seuss Geisel (1904-1991), aka "D... | 0.204355 | positive |
| 4 | Philip Nel - Dr. Seuss: American IconThis is b... | 0.169149 | positive |

Fig 24: Sentiment Analysis

Sentiment analysis results indicate whether the review content is positive or negative. The general trend of the reviews summarized and to proceed further in understanding the hidden trend.

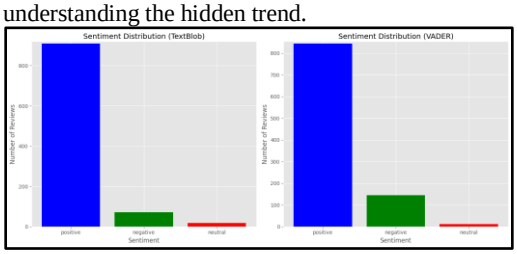



Fig 25: Plot of Sentiment Analysis

The distribution of sentiment scores for both text blob and VADER sentiment analysis techniques. It shall probably give an idea about the consistency and distinction between these two areas of sentiment analysis.

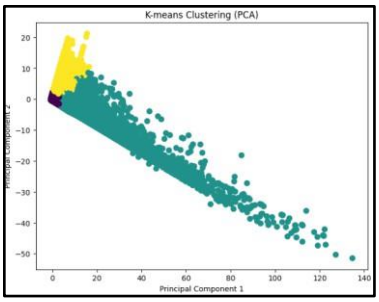



Fig. 26: K-means clustering (PCA)

Principal component analysis and data clustering are displayed using K-means Clustering (PCA). The presence of subgroups or segments within the data could be useful in further analysis and subsequent decisions.

Frequent Itemsets:

| | support | itemsets |
|---|---|---|
| 0 | 0.359753 | (type_PAYMENT) |
| 1 | 0.084519 | (type_TRANSFER) |
| 2 | 0.339765 | (type_CASH_OUT) |

Fig. 27: Frequency Itemsets

The Frequency Itemsets represent the support for different transaction types, providing a quantitative view toward the assessment of the relative frequency of diverse transaction categories within the dataset. The information can be useful in assessing the prevailing trends in the transactions where an analysis needs to be done regarding striking trends or imbalances.

*D. Discussion on sampling methods*

*Random Sampling:* Random sampling is the easiest and perhaps the most common of all techniques. The process where all instances of the data set are chosen confidently with each instance having the same likelihood of the selection. This method applies well when the data does not follow a specific order that leads to the rendering of prejudice. Low performance of the model if the data are unbalanced or if some extreme values influence the sample.

*Stratified Sampling:* The data is divided into distinct subgroups or layers of categories; the research uses stratified sampling. This method also guarantees that subgroups are also represented in the sample which subsequently increases the reliability of models, particularly in the context of classification models where there are many more data points in one class than in the other. Online credit card fraud detection and stratified sampling help address the issue of representation both in fraudulent and non-fraudulent cases.

*Cluster Sampling:* Cluster sampling entails partitioning of data into clusters and then sampling on some of the clusters in the data. This method is drawn on situations where data has been partitioned into groups such as areas. Although it can work effectively, its advantage is less apparent if the cluster samples are not sufficiently diverse from other groups.

*Systematic Sampling*

The Data points are taken in an interval form such as every ten observations. This method is convenient when data is ordered and possesses certain sources of biases in case of the existence of periodicity.

## V. CONCLUSIONS AND FUTURE WORK

*A. Summary of findings and limitations*

Different data sets of restaurant rating, books, and review and fraud transactions are exampled for classification, regression and unsupervised learning. It overlaps data preprocessing techniques like missing values handling, or distribution visualizations. Metrics like accuracy, precision, and recall are used to evaluate classification models like Random Forest and Naive Bayes, whereas R2 reflects regression models like KNN and Linear Regression. Other forms of conventional methods such as K-means clustering method and Apriori for association rule mining are also experimented. The outcomes returned the findings concerning model efficiency, clubs and the sentiment portrayed by respondents, which are helpful in decision-making.

*Limitations*

- In-exactness in data information and, thus, the risk of prejudiced dataset and, in turn, biased models due to missing or inconsistent data.
- Limited use of some specific algorithms, which may not be able to discover all the patterns in big data.

- No attention paid to factors outside the tests, constraints to extend the findings to other groups and populations.
- Little comparison with even more advanced or mixed types of interventions.

*B. Responses to the research enquiries*

*Responses to the research enquiries 1:* The customer-perceived variables for ratings including the quality and price of the product, features, age, gender, and class of buyers, and their experience on purchase which includes qualify and punctuality of the service provided. All these factors influence the ways that customers may rate a certain product or service.

*Answer to Research question 2:* Linear Regression and KNN are proficient in reducing the cost prediction error by identifying relations between customers' ratings and quantitative features of a product. Linear Regression is effective when the nature of the relationship is linear, but with surfaces, while KNN provides more accurate predictions in variate environments.

*Answer to Research question 3:* The result further showed that product attributes, customer expectations, and market positioning significantly impact customer satisfaction across categories. The results based on the analysis revealed better quality or cheaper items get positive feedback, while poor quality items, irrespective of the category, got negative feedback.

*Answer to Research question 4:* Transactions that are conducted in groups are easily identified by their high risk or demand for suspicious characteristics such as distorted spending patterns, odd time frames, or higher frequencies than normally expected. These clusters as distinguished from similar sets of other data analytics techniques such as Kunga-Means can assist in fraud detection.

*Answer to Research question 5:* Transaction characteristics that are important for fraud identification include; amount, geographical location, time of the transaction, and interaction frequency of the transaction and the customer. This is the case as these variables assist in outlining suspicious patterns normally when they are outside the average or customers' norms.

*C. Key Implications*

The Apriori technique demonstrates that a variety of machine learning models, such as Naive Bayes, Random Forest, Linear Regression, and KNN models, may be used to evaluate consumer sentiment, predict expenses, and detect fraud. Some of the drawbacks like ambiguity of interpretation of models like Random Forest and time complexity of KNN make the need for more interpretable and efficient such models. Reducing class imbalance in fraud detection and integration of multi-models for improved results will again increase the probability of accurate results and decisions made by the business organization.

*D. Future Work*

*Improving Interpretability:* The Random Forest algorithm practices a good amount of prediction accuracy, but it is not so suitable from the business perspective. Random Forest has several uses, such as sentiment analysis and fraud detection; Random Forest techniques are used to explain the results of these applications.

*Handling Data Imbalance:* According to fraud detection, especially in imbalanced datasets is observed to be highly problematical because the fact is that the incidence of actual fraudulent transactions is usually extremely low in comparison to legitimate ones.

*Real-time Processing:* Future work can improve the current algorithm and apply it in real-time, especially for those applications like fraud detection that need large amounts of computing.

## VI. REFERENCES


[1] Ashtiani, M.N. and Raahemi, B., 2021. Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review. *Ieee Access*, *10*, pp.72504-72525.

[2] Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V. and Nappi, M., 2021. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE access*, *9*, pp.39707-39716.

[3] Shi, Y., Sun, F., Zuo, H. and Peng, F., 2023. Analysis of learning behavior characteristics and prediction of learning effect for improving college students' information literacy based on machine learning. IEEE Access, 11, pp.50447-50461.

[4] Bujang, S.D.A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H. and Ghani, N.A.M., 2021. Multiclass prediction model for student grade prediction using machine learning. *Ieee Access*, *9*, pp.95608-95621.

[5] Asthana, P., Mishra, S., Gupta, N., Derawi, M. and Kumar, A., 2023. Prediction of student's performance with learning coefficients using regression based machine learning models. *IEEE Access*.

[6] Vispute, S.R. and Saini, M.L., 2022. Performance Analysis of Soil Health Classifiers Using Data Analytics Tools and Techniques for Best Model and Tool Selection. *Int. J. Online Biomed. Eng.*, *18*(10), pp.169-189.

[7] Sharma, M., Joshi, S., Sharma, S., Singh, A. and Gupta, R., 2021, September. Data mining classification techniques to assign individual personality type and predict job profile. In *2021 9th international conference on reliability, infocom technologies and optimization (trends and future directions)(icrito)* (pp. 1-5). IEEE.

[8] Pal, M. and Parija, S., 2021, March. Prediction of heart diseases using random forest. In *Journal of Physics: Conference Series* (Vol. 1817, No. 1, p. 012009). IOP Publishing.

[9] Wisanwanichthan, T. and Thammawichai, M., 2021. A double-layered hybrid approach for network intrusion detection system using combined naive bayes and SVM. *Ieee Access*, *9*, pp.138432-138450.

[10] Al Ahasan, M.A., Hu, M. and Shahriar, N., 2023, May. Ofmcdm/irf: A phishing website detection model based on optimized fuzzy multi-criteria decision-making and improved random forest. In *2023 Silicon Valley Cybersecurity Conference (SVCC)* (pp. 1-8). IEEE.

[11] Jindal, H., Agrawal, S., Khera, R., Jain, R. and Nagrath, P., 2021. Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[12] Chung, J. and Lee, K., 2023. Credit card fraud detection: an improved strategy for high recall using KNN, LDA, and linear regression. *Sensors*, *23*(18), p.7788.
[13] Usman, A.U., Abdullahi, S.B., Liping, Y., Alghofaily, B., Almasoud, A.S. and Rehman, A., 2024. Financial Fraud Detection Using Value-at-Risk with Machine Learning in Skewed Data. *IEEE Access*.
[14] Nayyer, N., Javaid, N., Akbar, M., Aldegheishem, A., Alrajeh, N. and Jamil, M., 2023. A new framework for fraud detection in bitcoin transactions through ensemble stacking model in smart cities. *IEEE Access*.
[15] Mienye, I.D. and Sun, Y., 2023. A deep learning ensemble with data resampling for credit card fraud detection. *IEEE Access*, *11*, pp.30628-30638.
[16] Phyo, P.P. and Jeenanunta, C., 2021. Daily load forecasting based on a combination of classification and regression tree and deep belief network. *IEEE Access*, *9*, pp.152226-152242.
[17] Noothout, J.M., De Vos, B.D., Wolterink, J.M., Postma, E.M., Smeets, P.A., Takx, R.A., Leiner, T., Viergever, M.A. and Išgum, I., 2020. Deep learning-based regression and classification for automatic landmark localization in medical images. *IEEE transactions on medical imaging*, *39*(12), pp.4011-4022.
[18] Liu, K., Hu, X., Zhou, H., Tong, L., Widanage, W.D. and Marco, J., 2021. Feature analyses and modeling of lithium-ion battery manufacturing based on random forest classification. *IEEE/ASME Transactions on Mechatronics*, *26*(6), pp.2944-2955.
[19] Rupapara, V., Rustam, F., Shahzad, H.F., Mehmood, A., Ashraf, I. and Choi, G.S., 2021. Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access*, *9*, pp.78621-78634.
[20] Iqbal, M., Iqbal, A., Alshammari, A., Ali, I., Maghrabi, L.A. and Usman, N., 2024. Sell or HODL Cryptos: Cryptocurrency Short-to-Long Term Projection Using Simultaneous Classification-Regression Deep Learning Framework. *IEEE Access*.