Analysis of Machine Learning Models in Banking, E-Commerce, and Real Estate Sectors

Preena Darshini .
School of Computing
National College of Ireland
Dublin, Ireland
x22238590@student.ncirl.ie

***Abstract*—This report provides a comprehensive analysis of three different and unique datasets: Bank Customer Churn, Amazon Reviews, and California Housing, using various machine learning models. This project aims to show the efficacy of Logistic Regression and K-Medoids on bank customer churn, K-Means for clustering Amazon Reviews, and Multiple Linear Regression and Random Forest for predicting California housing prices. Each model's performance and diagnostic tests will be assessed.**

Keywords—Logistic Regression, K-Medoids, K-Means, Multiple Linear Regression, Random Forest

I. INTRODUCTION

The main purpose of this project is to create a portfolio to analyse different machine learning models on different datasets representing unique sectors. This analysis involves implementing different machine learning models for the banking, e-commerce, and real estate sectors. Classification, clustering, and regression models are implemented. Insights from this analysis will not only contribute to understandings of how different models perform but also provide valuable information and insights for both academic research and practical applications in data science. Knowledge Discovery in Databases (KDD) [13] is the data mining methodology used in this project portfolio. Each sector comes with its own set of challenges. For the banking sector, one of the challenges lies in predicting customer churn accurately. For e-commerce, large amounts of textual data like reviews, must be categorized into groups. In the housing sector, the challenge lies in predicting the sale price of a property, which is dependent on various factors like physical state of the property to the surrounding conditions of the area in which the property is located.

II. RELATED WORK

In the realms of banking, e-commerce, and real estate, numerous studies have leveraged various machine learning models to predict bank customer churn, classify reviews, and predict housing prices.

A) Dataset 1: Banking customer churn

According to [1], the data was balanced using SMOTE (Synthetic Minority Over-sampling Technique). It assesses the effectiveness of different machine learning models, especially classification models, including logistic regression and random forest. It focuses on the importance of preprocessing data and concludes that Random Forest performed the best after data balancing.

According to [2], it uses a Logit Leaf Model (LLM) which is a hybrid algorithm. LLM combines both decision trees and logistic regression. LLM was used for enhanced accuracy in prediction. It outperformed traditional logistic regression and decision trees. It provided valuable insights and information for strategies for retaining customers.

In [3], the research focused on comparing the performance of logistic regression, random forest, and SVM (Support Vector Machine). It comes to the conclusion that SVM outperformed other models for churn prediction in the banking sector.

[4] explores the use of deep learning in predicting customer churn. It utilizes the dataset from Pasargad Bank in Iran. This study uses a Bi-directional Long Short-Term Memory (Bi-LSTM) neural network. It concludes that deep learning predicts customer churn in retail banking accurately.

To summarize, all the above studies used Logistic Regression and Random Forest to predict churn. It could require extensive preprocessing and many struggle due to imbalance in datasets.

In this dataset, Logistic Regression will be applied to predict customer churn and K-Medoids will be applied to segment customers. K-medoids is usually robust to outliers compared to K-Means.

B) Dataset 2: Amazon Reviews

[5] emphasizes the use of product reviews and Amazon's review voting system. It implies that the number of product reviews and the percentage of helpful votes have an inverse relationship.

[6] combines collaborative filtering and K-means clustering for recommending products. It concludes that using popularity-based, model-based, and description-based recommendation systems improve the user experience significantly.

In [7], the effectiveness of clustering algorithms, especially K-means and Peak-searching in sectioning product reviews based on topics is discussed. It finds that K-means performs better than Peak-searching.

[8] explores the application of K-means and Salp Swarm Algorithm to analyse customer reviews. It ranks customer reviews from different shopping websites. This study highlights the use of big data in mobile commerce, especially to understand customer preferences.

It can be summarized that various studies above used clustering algorithms like Peak- searching and K-means to categorize reviews. K-means can be sensitive to outliers however, it has good computational efficiency, which makes it suitable for large datasets.

C) Dataset 3: Housing data

[9] uses boosting ensemble regression trees and Gaussian process regression. It was then optimized using Bayesian techniques. It utilizes housing data from Taiwan and concludes that the model that performs the best is the boosting ensemble regression trees, which indicates that Bayesian optimization played a significant role in increasing model performance.

In [10], different regression methods were used to predict house price. It concludes that the Gradient Boosting Regressor was the most accurate post hyperparameter optimization. It emphasizes more on using statistical analysis methods in the selection of the model.

[11] explores the efficiency of Random Forest, XGBoost, SVR, and ANN models to predict house prices. It determines that ANN models are the best at predicting prices.

Finally, [12] research highlights the benefits of using Gradient Boosting to handle sophisticated datasets over Linear Regression. It aims at providing useful insights to both customers and real estate professionals.

Finally, from the above studies, Linear Regression provides a clear and easy to interpret model. It may not capture all the complexities. Random Forest handles non-linear data better.

III. Data Mining Methodology

Knowledge Discovery in Databases (KDD) is the process that uses data mining as one of its key components. It includes various steps: data selection, data preprocessing, data transformation, data mining, evaluation, and finally knowledge presentation [13].

Data Selection: The relevant datasets from data sources are collected.

Data pre-processing: Preparing the data and cleaning it so data mining can be performed.

Transformation: This step involves implementing various transformations like log, square root, and so on.

Data mining: This step involves selecting the suitable machine learning algorithms to implement on the datasets.

Evaluation: In this step, results are evaluated and analysed to identify trends and patterns.

Knowledge representation: This is the final step of KDD process. It involves presenting useful information.

KDD has both advantages and disadvantages. Some of the advantages include: gaining valuable information from datasets, and identifying trends and patterns. Some of the disadvantages include: complexity, it can be a time-consuming process, privacy concerns when dealing with sensitive data. In this analysis using R programming language in R Studio, all KDD steps have been implemented.

IV. Implementation

A) Dataset 1: Bank Customer Churn

This dataset [14] was sourced from Kaggle and contains information about customers at ABC Multistate Bank. It contains 12 columns and 10,000 rows.

TABLE I. VARIABLES AND DATA TYPES FOR DATASET 1

Variable Name	Type
customer_id	Numerical, Discrete
credit_score	Numerical, Discrete
country	Categorical
gender	Categorical
age	Numerical, Discrete
tenure	Numerical, Discrete
balance	Numerical, Continuous
products_number	Numerical, Discrete
credit_card	Categorical
active_member	Categorical
estimated_salary	Numerical, Discrete
churn	Categorical

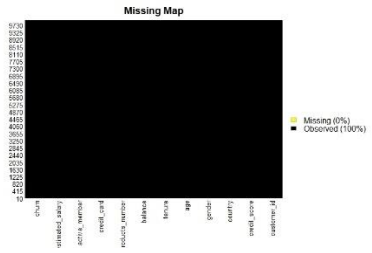
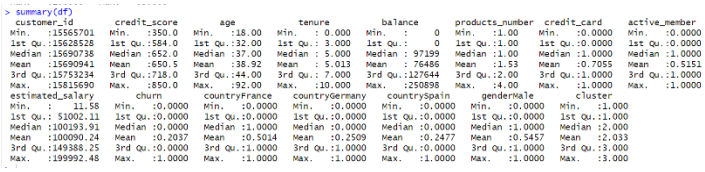


Fig. 1: Amelia Missmap for Dataset 1

a) Logistic Regression

Step 1: The dataset [14] is selected from Kaggle as mentioned above. All the necessary libraries for preprocessing the data and for modeling are loaded. The dataset is in .csv format. The Bank Customer Churn.csv data is read and stored in a dataframe.



```
> summary(df)
  customer_id      credit_score        age             tenure          balance       products_number  credit_card      active_member
 Min.   :15565701   Min.   :350.0   Min.   :18.00   Min.   : 0.000   Min.   :     0   Min.   :1.00   Min.   :0.0000   Min.   :0.0000
 1st Qu.:15628528   1st Qu.:584.0   1st Qu.:32.00   1st Qu.: 3.000   1st Qu.:     0   1st Qu.:1.00   1st Qu.:0.0000   1st Qu.:0.0000
 Median :15690738   Median :652.0   Median :37.00   Median : 5.000   Median : 97199   Median :1.00   Median :1.0000   Median :1.0000
 Mean   :15690941   Mean   :650.5   Mean   :38.92   Mean   : 5.013   Mean   : 76486   Mean   :1.53   Mean   :0.7055   Mean   :0.5151
 3rd Qu.:15753234   3rd Qu.:718.0   3rd Qu.:44.00   3rd Qu.: 7.000   3rd Qu.:127644   3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :15815690   Max.   :850.0   Max.   :92.00   Max.   :10.000   Max.   :250898   Max.   :4.00   Max.   :1.0000   Max.   :1.0000
 estimated_salary       churn        countryFrance    countryGermany    countrySpain      genderMale        cluster
 Min.   :    11.58   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :1.000
 1st Qu.: 51002.11   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000
 Median :100193.91   Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000   Median :1.0000   Median :2.000
 Mean   :100090.24   Mean   :0.2037   Mean   :0.5014   Mean   :0.2509   Mean   :0.2477   Mean   :0.5457   Mean   :2.033
 3rd Qu.:149388.25   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :199992.48   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :3.000
```

Fig.2: Statistical Summary of Dataset 1

Step 2: The basic data exploration is done. The dimensions of the dataset, number of rows and columns, the column names, and the first few rows of the dataset is printed, the statistical summary and structure of dataset is viewed as shown in Fig. 2 and Table I, respectively. The missmap from Amelia package in R Studio is used to identify columns that have any missing value, as seen in Fig. 1. There are no missing values.

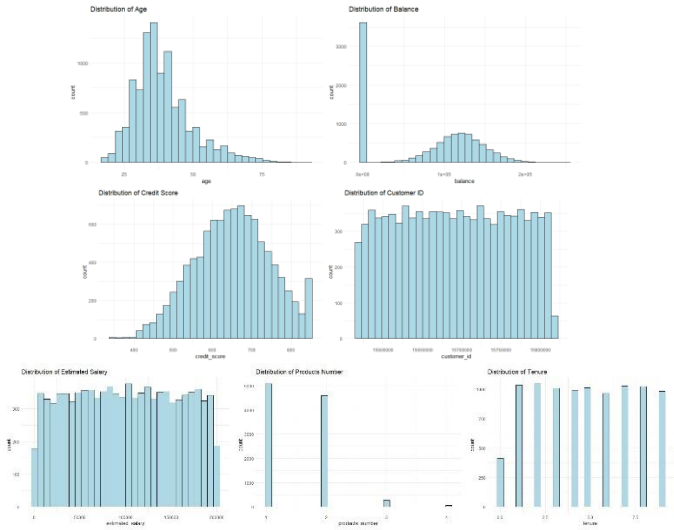


Fig. 3: Histograms of numerical variables

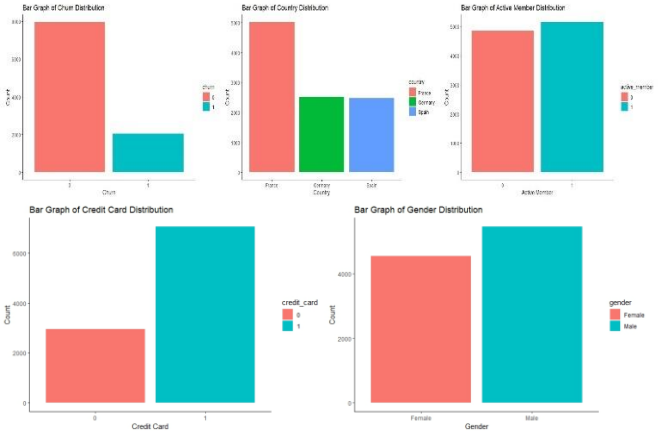


Fig. 4: Bar Plots for Categorical Variables

Step 3: EDA (Exploratory Data Analysis) is done here. Histograms for numerical variables like customer ID, credit score, age, tenure, balance, products number, and estimated salary are plotted to understand each variable distribution as seen in Fig. 3. Bar plots are plotted for categorical variables like country, gender, credit card possession, and churn, as seen in Fig. 4. Bivariate analysis of all variables against churn is done using scatter plots. Log transformation is applied to products number to normalize the data. The numerical variables are checked for outliers. The outliers are not wrong or incorrect information, so it will be used in the analysis and not imputed or deleted. Credit scores and ages are grouped into categories. Binning process reduces minor observation errors. Initially, without binning the model was performing poorly with low Kappa value. Binning was employed to increase model performance. Log transformations were applied to products number to normalize it. PCA for dimensionality reduction was initially attempted. However, the computational time was high, and the workload on the system increased, thereby, slowing the system down. Hence, its application was withdrawn.

Step 4: The dataset is ready for modeling. It is split into training and testing data. The Logistic Regression Model is built using the 'glm' function. The dependent variable is churn.

```
Confusion Matrix and Statistics

          0    1
  0 2310  449
  1   78  162

               Accuracy : 0.8243
                 95% CI : (0.8102, 0.8377)
    No Information Rate : 0.7963
    P-Value [Acc > NIR] : 5.914e-05

                  Kappa : 0.3003

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9673
            Specificity : 0.2651
         Pos Pred Value : 0.8373
         Neg Pred Value : 0.6750
             Prevalence : 0.7963
         Detection Rate : 0.7703
   Detection Prevalence : 0.9200
      Balanced Accuracy : 0.6162

       'Positive' Class : 0
```

Fig. 5: Confusion Matrix

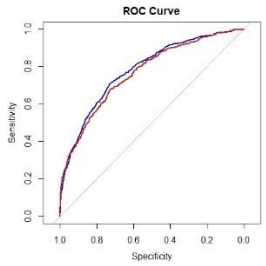


Fig 6: ROC curves for train and test data

Step 5: Model Evaluation is done. The performance of the model is evaluated on both training and test data using the ROC (Receiver Operating Charateristic) curve and AUC (Area Under the Curve). The AUC value of training data was 0.7796, and the AUC value of test data was 0.7650. The Confusion Matrix is created as shown in Fig. 5. It tells about the predictive accuracy of the model. The accuracy was 82.4%, which is the proportion of total predictions that were correct. Kappa value of 0.30 which is average. It suggests that there is a moderate agreement between predicted and actual values. The model performs better than random guessing. The ROC curves for both training and test data are plotted for comparison, as shown in Fig. 6. Train ROC is plotted in blue, while test ROC is plotted in red.

b) K-Medoids

Step 1: The dataset [14] mentioned above is used. All the necessary libraries for preprocessing the data and for modeling are loaded. The dataset is in .csv format. The Bank Customer Churn.csv data is read and stored in a dataframe.

Step 2: Basic data exploration, like dataset's dimensions, structure, and statistical summary, is done as shown in Table I and Fig. 2, respectively. Missing data, if any is checked. There are no missing values, as seen in Fig. 1.

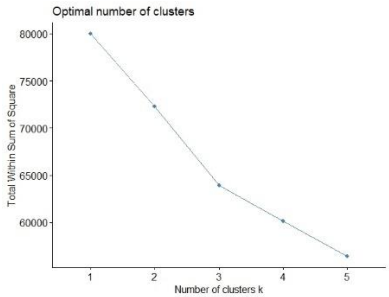


Fig. 7: Optimal number of clusters using Elbow Plot

Step 3: Data preprocessing steps are carried out next. Non-binary variables (gender and country) are converted into factor type and dummy variables are created for the same. All categorical variables are formatted into numerical type. All numerical features are scaled, as it is important for clustering algorithms so that all features contribute equally to the model. Elbow method is applied to check for an optimal number of clusters.The K value is chosen as 3 based on the plot as shown in Fig. 7.

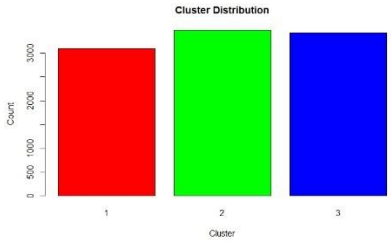


Fig. 8: Cluster Distribution

```
> print(cluster_means)
  Group.1 credit_score      age   balance estimated_salary
1       1     642.5166 40.22843 110978.52         100077.5
2       2     646.2681 38.49483 104114.74         100834.6
3       3     662.0981 38.17489  17244.11          99345.5
```

Fig. 9: Cluster means

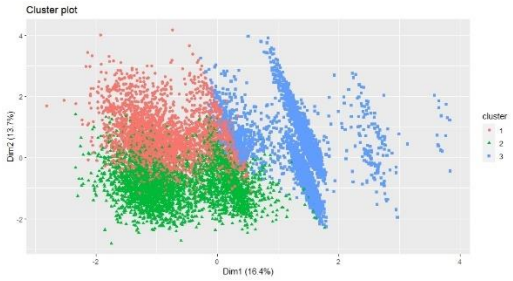


Fig. 10: Cluster Scatterplot

Step 4: Unsupervised clustering algorithm K-Medoids, which uses Partitioning Around Medoids (PAM) algorithm to check for the optimal number of clusters, is applied to the scaled data. A bar plot representing cluster distribution is shown in Fig. 8. Clustering is the data mining process here.

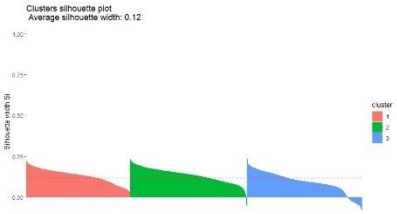


Fig. 11: Clusters silhouette plot

```
> fviz_silhouette(silhouette_score)
  cluster size ave.sil.width
1       1 3095          0.13
2       2 3480          0.13
3       3 3425          0.09
```

Fig. 12: Silhouette scores

Step 5: The centers of clusters, also called medoids, are extracted. From the cluster means values in Fig. 9, it can be inferred that cluster 3 has the highest credit score and the lowest balance indicating it could be new customers . Cluster 1 has higher balances and likely indicates older customers than cluster 2. A scatter plot is used to visualize the cluster center, as shown in Fig. 10. Finally, Silhouette Width scores are calculated as shown in Figs. 11 and 12. It measures how similar the object is to its own cluster. These clustering results give insights into customer segmentation with respect to bank churn analysis.

B) Dataset 2: Amazon Reviews

This dataset [15] was sourced from Kaggle and contains information about different Amazon products and product reviews. It originally contained 10 columns and 568454 rows. Only 14999 rows have been extracted from this dataset for easier computation.

TABLE II. VARIABLES AND DATA TYPES FOR DATASET 2

Variable	Type
Id	Numerical, Discrete
ProductId	Categorical
UserId	Categorical
ProfileName	Categorical
HelpfulnessNumerator	Numerical, Discrete
HelpfulnessDenominator	Numerical, Discrete
Score	Categorical
Time	Numerical, Continuous
Summary	Categorical (Textual Data)
Text	Categorical (Textual Data)

a) K-Means

Step 1: The dataset [15] is selected from Kaggle as mentioned above. All the necessary libraries for pre-processing the data and for modeling are loaded. The dataset is in .csv format. The read.csv function is used to load the filterReviews1.csv dataset in R Studio.

```
> summary(df)
       Id          ProductId           UserId          ProfileName        HelpfulnessNumerator
 Min.   :    1   Length:14999       Length:14999       Length:14999       Min.   :  0.000
 1st Qu.: 3750   Class :character   Class :character   Class :character   1st Qu.:  0.000
 Median : 7500   Mode  :character   Mode  :character   Mode  :character   Median :  0.000
 Mean   : 7500                                                           Mean   :  1.547
 3rd Qu.:11250                                                           3rd Qu.:  2.000
 Max.   :14999                                                           Max.   :202.000
 HelpfulnessDenominator     Score            Time              Summary              Text
 Min.   :  0.000        Min.   :1.000   Min.   :9.617e+08   Length:14999       Length:14999
 1st Qu.:  0.000        1st Qu.:4.000   1st Qu.:1.272e+09   Class :character   Class :character
 Median :  1.000        Median :5.000   Median :1.307e+09   Mode  :character   Mode  :character
 Mean   :  2.011        Mean   :4.143   Mean   :1.295e+09
 3rd Qu.:  2.000        3rd Qu.:5.000   3rd Qu.:1.330e+09
 Max.   :219.000        Max.   :5.000   Max.   :1.351e+09
```

Fig. 13: Statistical Summary for Dataset 2

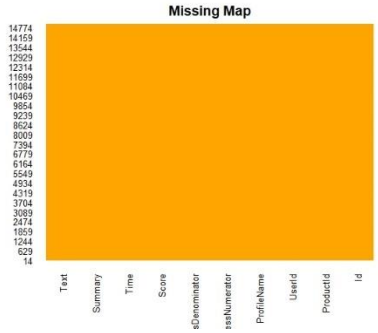


Fig. 14: Amelia Missmap

Step 2: Data preprocessing is done where dimensions, column names, structure, and summary of the dataset is explored, as seen in Table II and Fig. 13. Amelia missmap is used to visualize and check for missing values, if any, as shown in Fig. 14. New columns are created i.e., feature engineering is done.

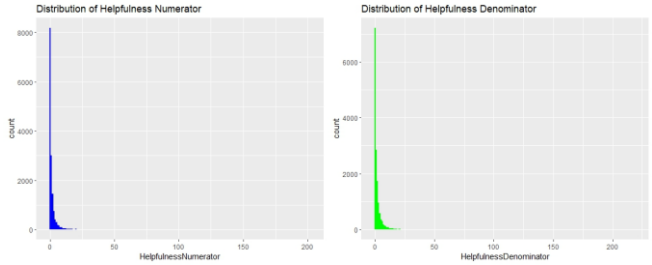


Fig. 15: Histograms of Helpfulness Numerator and Helpfulness Denominator

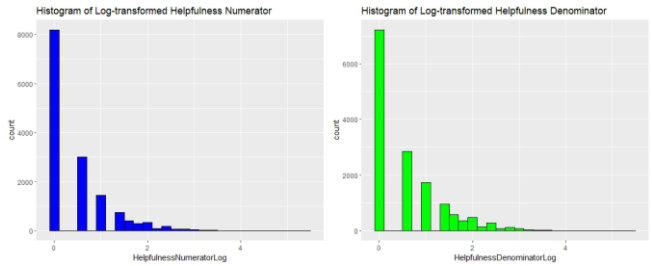


Fig. 16: Histograms of Log transformed Helpfulness Numerator and Helpfulness Denominator

HelpfulnessRatio, HelpfulnessNumeratorLog and HelpfulnessDenominatorLog are derived from the existing data and visualized as shown in Figs. 15 and 16. Outliers are identified using box plots and IQR method and printed. Outliers are not incorrect data so it is included in this analysis.

Step 3: In the transformation phase, HelpfulnessNumeratorLog and HelpfulnessDenominatorLog are normalized using the 'scale' function. This step is essential for a distance-based algorithm like K-means clustering.

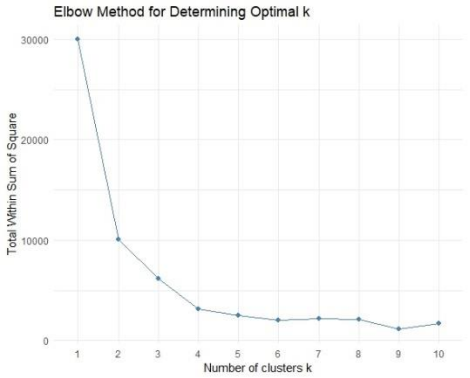


Fig. 17: Elbow Method Plot

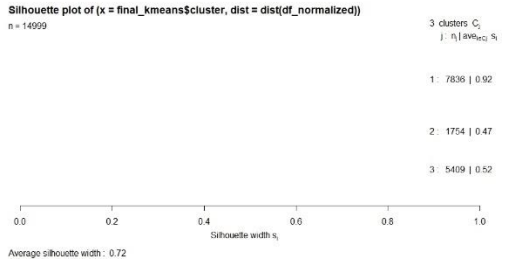


Fig. 18: Silhouette Plot

Step 4: In the data mining process, the optimal number of clusters is determined using Elbow Method. Based on the Elbow Method shown in Fig. 17, the number of clusters chosen is 3. K-means algorithm is run, and silhoutte scores are calculated, which tells about the clustering quality as shown in Fig. 18. It can be seen that the average silhouette width of Cluster 1 is quite high. Clusters 2 and 3 have a moderately average silhouette width. A high average silhouette width for all clusters is 0.72, which indicates that clusters are well separated from each other and grouped closely.

```
> print(cluster_summary)
# A tibble: 3 × 5
  cluster HelpfulnessNumerator_mean
  <fct>                       <dbl>
1 1                               0
2 2                            8.80
3 3                            1.43
# i 3 more variables:
#   HelpfulnessNumerator_median <dbl>,
#   HelpfulnessDenominator_mean <dbl>,
#   HelpfulnessDenominator_median <dbl>
```

Fig. 19: Cluster Summary

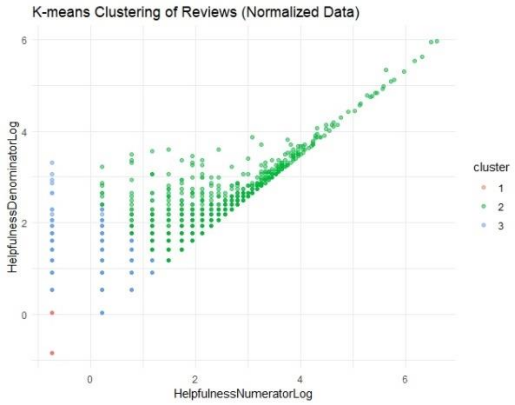


Fig. 20: Clustering of Reviews

Step 5: Evaluation is done using the silhoutte score, and a scatter plot of the normalized data with the clusters is displayed to visualize cluster assignments of reviews, as seen in Fig. 20. It shows that the cluster forms a pattern like a diagonal line, which indicates that there is a correlation between HelpfulnessNumeratorLog and HelpfulnessDenominatorLog. This means that the more people vote on a helpfulness review, the higher will be the number of helpful votes. From the cluster summary shown in Fig. 19, it can be inferred that Cluster 1 has reviews that are not marked as helpful, Cluster 2 contains reviews where opinions seem divided and Cluster 3 has popular reviews that are marked as helpful. This can help the business understand customer feedback and improve products.

C) Dataset 3: California Housing

This dataset [16] was sourced from Kaggle and contains information about housing data in California. It contains 10 columns and 20640 rows. This data contains information about 1990 census data from California.

TABLE III. VARIABLES AND DATA TYPES FOR DATASET 3`

Variable	Type
Longitude	Numerical, Discrete
Latitude	Numerical, Discrete
Housing Median Age	Numerical, Discrete
Total Rooms	Numerical, Discrete
Total Bedrooms	Numerical, Discrete
Population	Numerical, Discrete
Households	Numerical, Discrete
Median Income	Numerical, Continuous
Median House Value	Numerical, Continuous
Ocean Proximity	Categorical

a) Multiple Linear Regression

Step 1: The dataset [16] is selected from Kaggle as mentioned above. All the necessary libraries for preprocessing the data and for modeling are loaded. The dataset is in .csv format. The read.csv function is used to load the housing.csv dataset in R Studio.

```
> summary(df)
   longitude         latitude     housing_median_age  total_rooms     total_bedrooms
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00       Min.   :    2   Min.   :   1.0
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00       1st Qu.: 1448   1st Qu.: 296.0
 Median :-118.5   Median :34.26   Median :29.00       Median : 2127   Median : 435.0
 Mean   :-119.6   Mean   :35.63   Mean   :28.64       Mean   : 2636   Mean   : 537.9
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00       3rd Qu.: 3148   3rd Qu.: 647.0
 Max.   :-114.3   Max.   :41.95   Max.   :52.00       Max.   :39320   Max.   :6445.0
                                                                      NA's   :207
   population      households     median_income     median_house_value
 Min.   :    3   Min.   :   1.0   Min.   : 0.4999   Min.   : 14999
 1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634   1st Qu.:119600
 Median : 1166   Median : 409.0   Median : 3.5348   Median :179700
 Mean   : 1425   Mean   : 499.5   Mean   : 3.8707   Mean   :206856
 3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432   3rd Qu.:264725
 Max.   :35682   Max.   :6082.0   Max.   :15.0001   Max.   :500001

ocean_proximity
Length:20640
Class :character
Mode  :character
```

Fig. 21: Summary statistics for Dataset 3

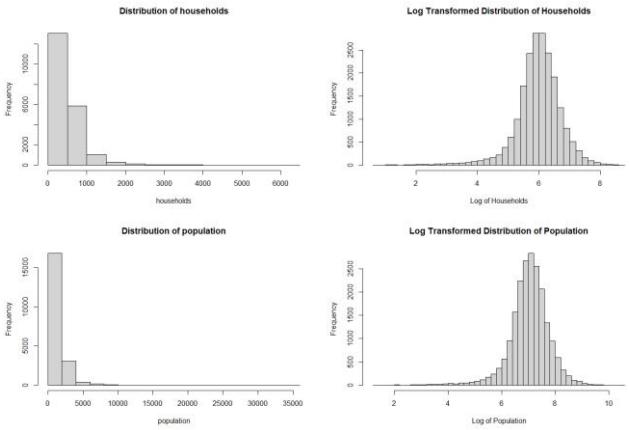


Fig. 22: Histograms for households and population and its transformations

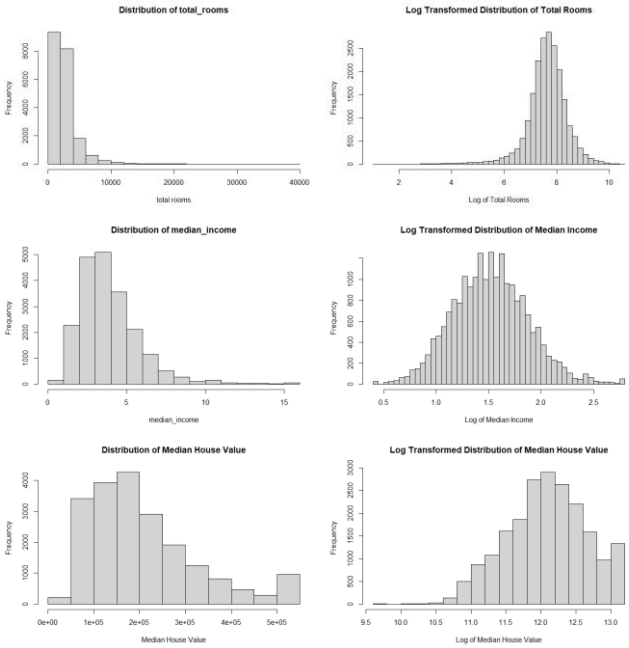


Fig. 23: Histograms for total_rooms, median_income, and median_house_value and population and its transformations

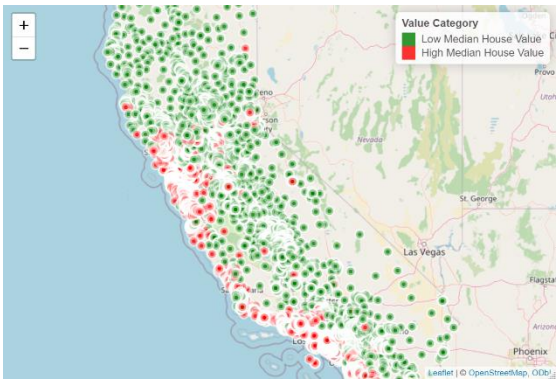


Fig. 24: Map for house price distribution

Step 2: The basic structure, summary, and data dimensions are checked as shown in Table III and Fig. 21. Variables are transformed to normalize it. As seen in Figs. 22 and 23. Categorical variables are converted into factor type. Dummy variables are created for the categorical variables. The media house prices are categorized into high and low. A map is generated which plots these high and low median house values using leaflet library. Most of the high median house vales are distributed along the coast.

Step 3: Outliers in log transformed median_house_value are identified and removed. The outliers and missing values in the dataset were removed to reduce the computational workload on the machine, as the dataset has a large number of rows. Including them increased the computational time. Hence, for this analysis, the outliers were omitted. However, domain expertise is required when handling outliers. After removing outliers, the number of rows is now 20407.

```
> print(cor_matrix)
                     longitude    latitude housing_median_age total_rooms_log population_log
longitude          1.00000000 -0.92502291         -0.10907612      0.03050225     0.11014280
latitude          -0.92502291  1.00000000          0.01199778     -0.03185926    -0.13609155
housing_median_age -0.10907612  0.01199778          1.00000000     -0.31508428    -0.24438642
total_rooms_log    0.03050225 -0.03185926         -0.31508428      1.00000000     0.86403877
population_log     0.11014280 -0.13609155         -0.24438642      0.86403877     1.00000000
households_log     0.05635466 -0.08748925         -0.24283497      0.93152767     0.93178404
median_income_log -0.01550764 -0.08503239         -0.13758506      0.24996516     0.02876465
median_house_value_log -0.02208819 -0.19316219     0.07778832      0.18116528     0.01987572
total_bedrooms     0.06970179 -0.06653960         -0.32112776      0.76069928     0.72361683
                   households_log median_income_log median_house_value_log total_bedrooms
longitude              0.05635466       -0.01550764              -0.02208819     0.06970179
latitude              -0.08748925       -0.08503239              -0.19316219    -0.06653960
housing_median_age    -0.24283497       -0.13758506               0.07778832    -0.32112776
total_rooms_log        0.93152767        0.24996516               0.18116528     0.76069928
population_log         0.93178404        0.02876465               0.01987572     0.72361683
households_log         1.00000000        0.04412660               0.10797348     0.78813317
median_income_log      0.04412660        1.00000000               0.68210101     0.01703556
median_house_value_log 0.10797348        0.68210101               1.00000000     0.07829596
total_bedrooms         0.78813317        0.01703556               0.07829596     1.00000000
```

Fig. 25: Correlation Matrix

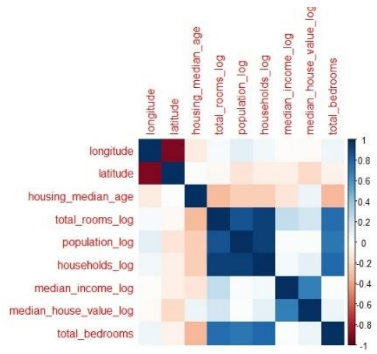


Fig. 26: Correlation heatmap

```
> summary(mlr_model)

Call:
lm(formula = mlr_formula, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71504 -0.21918 -0.02161  0.19331  2.26794

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.9782768  0.4680476  -2.090 0.036622 *
longitude               -0.1381770  0.0054237 -25.477  < 2e-16 ***
latitude                -0.1344406  0.0053423 -25.165  < 2e-16 ***
housing_median_age       0.0021072  0.0002251   9.363  < 2e-16 ***
median_income_log        0.9043805  0.0077909 116.081  < 2e-16 ***
ocean_proximity_INLAND  -0.3177353  0.0093788 -33.878  < 2e-16 ***
ocean_proximity_ISLAND   0.6193178  0.1667092   3.715 0.000204 ***
`ocean_proximity_NEAR BAY`   0.0098274  0.0103549   0.949 0.342602
`ocean_proximity_NEAR OCEAN` 0.0091855  0.0085028   1.080 0.280030
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3332 on 16317 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6531
F-statistic:  3843 on 8 and 16317 DF,  p-value: < 2.2e-16
```

Fig. 27: Summary of model

Step 4: The correlation matrix shown in Fig. 25 along with the heatmap shown in Fig. 26 are displayed to understand the relationship between the variables in the dataset. The dataset is split into training and test sets, and the Multiple Linear Regression (MLR) model is applied to training data using the 'lm' function to make predictions on both training and test data. The model summary is shown in Fig. 27. It can be inferred that the chosen predictors are the most significant. Initially, all predictors were used and the model did not perform well. When only the significant predictors were chosen, the model performed comparatively better. Forward selection was applied initially, however, due to the large size of the dataset, it was time consuming. Hence, only significant predictors were chosen for this modeling.

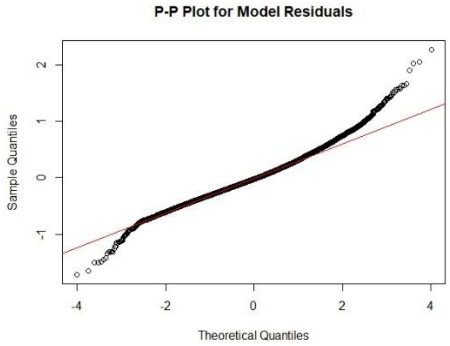


Fig. 28: P-P plot

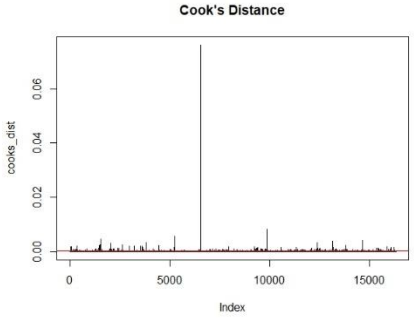


Fig. 29: Cook's Distance

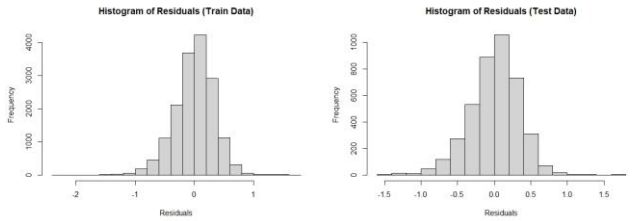


Fig. 30: Histogram of Residuals

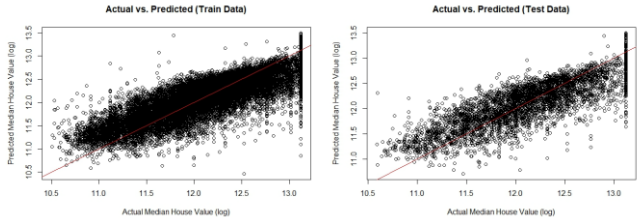


Fig. 31: Actual vs Fitted

Step 5: Model performance on both the training and test data is evaluated using Root Mean Square Error (RMSE) and R-squared values. The RMSE value is 33% , and the R-squared value is 65.33% for the training data. The RMSE value is 33.3% and the R-squared value is 65% on test data. P-P plots for normality shown in Fig. 28 and plots of histograms of residuals as shown in Fig. 30 is used for residual analysis. The P-P Plot shows that most of the residuals follow the diagonal line closely. This indicates the residuals are normally distributed. There are slight deviations at the ends and it could indicate residuals have heavier tails i.e. a slight right skewness as seen in Fig. 30. Influential points are identified using Cook's Distance, as shown in Fig. 29. Plots comparing actual values vs. predicted values for both training and test datasets are visualized as shown in Fig. 31.

b) Random Forest

Step 1: Data selection is done as mentioned above. The dataset [16], which is in .csv format is loaded . The basic structure, summary, and data dimensions are checked as shown in Table III and Fig. 21. Missing values are omitted to reduce computational workload.

Step 2:. Categorical variables are converted into factor type and dummy variables are created for the same. The outliers are removed to reduce computational workload. However, domain expertise is required in this case of removing outliers and missing values. In this scenario, both have been omitted to reduce computational time as the dataset is large. The map is plotted to show the distribution of median house prices, as shown is Fig. 24.

Step 3: Log transformations are done to normalize the distributions of those variables, as shown in Figs. 22, and 23. Histograms are plotted to visualize both before and after log transformation.

```
Call:
 randomForest(x = train_data[, predictor_vars], y = train_data[,      outcome_var], ntree =
100, mtry = sqrt(length(predictor_vars)),      nodesize = 5, importance = TRUE, na.action =
na.omit)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 3

          Mean of squared residuals: 0.0548372
                    % Var explained: 82.86
```

Fig. 32: Model Summary

Step 4: The dataset is split into training and testing data. The Random Forest model is trained using the variables. The number of trees and variables to consider at each split is mentioned. The summary of the model is shown in Fig. 32.

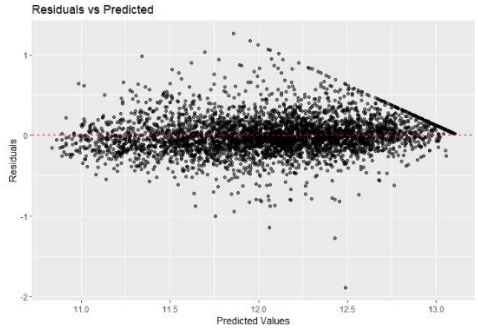


Fig. 33: Residual vs Predicted

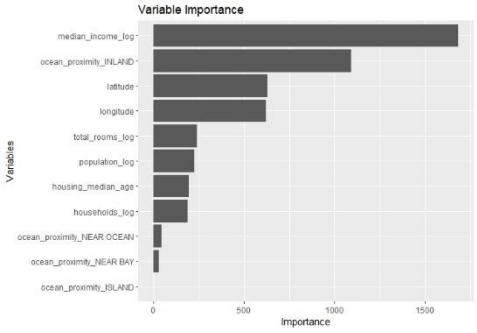


Fig. 34: Variable Importance

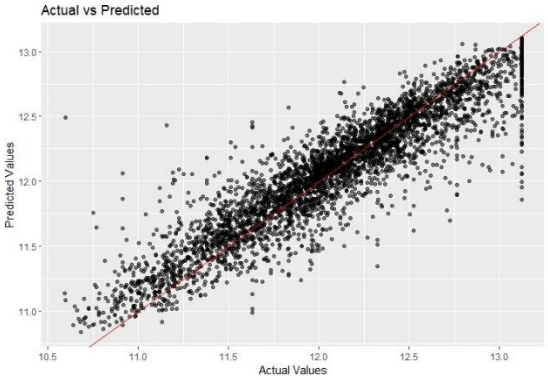


Fig. 35: Actual vs Predicted

Step 5: The performance of the model on test data is calculated using the RMSE which is 22.5%, and the R-squared value which is 84.3%. The residuals plot in Fig. 33 shows that most residuals are clustered around zero, which indicates good predictions, especially for lower predicted values and variable importance plot was created to analyse the model's accuracy and important variables. The plot showing variable importance in Fig. 34 shows that median_income_log is the most important predictor for the model, followed by the other variables shown in the figure. Actual vs. predicted values are plotted as shown in Fig. 35. The points cluster around the red line, which implies that the model predictions agree with the actual values. .

On comparing both Linear Regression and Random Forest modeling outcomes, the Random Forest

outperforms Multiple Linear Regression. The RSME value of Linear Regression on test data is 33.3% and R-squared value of 65% whereas, the RSME value of Random Forest is 22.5% and R-squared value is 84.3%.

V. Conclusion

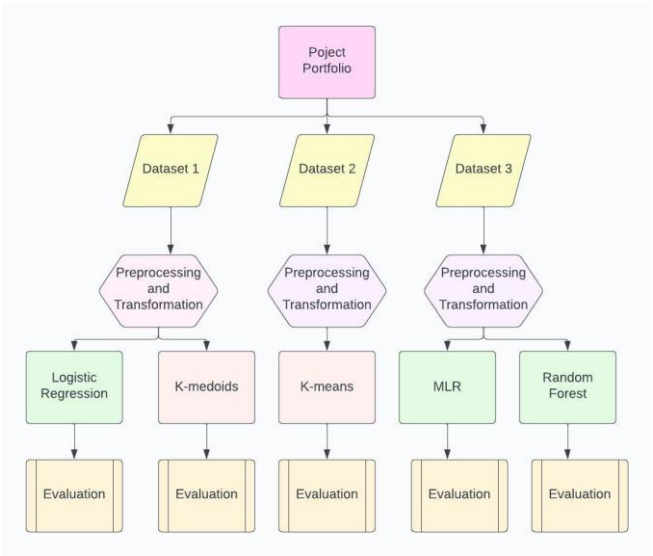


Fig. 36: Overview

This report has analysed three distinct datasets that represent three different sectors. Three supervised learning algorithms and two unsupervised learning algorithms were applied to these datasets. The overview of the entire project portfolio is shown in Fig. 36. The Logistic Regression model applied to Dataset 1 shows moderate predictive capability with an AUC value of 0.77 for training and 0.765 for test data. The overall accuracy is 82.4%, and Kappa value of 0.3 indicates that the model is reasonably good.
The K-medoids clustering algorithm applied to Dataset 1 helped in identifying distinct customer segments. Cluster 3 represents newer customers. K-means clustering algorithm was applied on Dataset 2. Which helped to cluster reviews based on helpfulness. MLR and Random Forest applied to Dataset 3 showed that Random Forest performed better at predicting housing sale prices. KDD methodology was followed throughout the entire analysis.

References

[1] I. Huseyinov and O. Okocha, "A Machine Learning Approach To The Prediction Of Bank Customer Churn Problem," 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2022, pp. 1-5, doi: 10.1109/IISEC56263.2022.9998299

[2] X. Li and Z. Chen, "Customer Churn Prediction in Bank Based on Different Machine Learning Models," 2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM), Montreal, ON, Canada, 2022, pp. 274-279, doi: 10.1109/ISPCEM57418.2022.00061.

[3] Arno De Caigny, Kristof Coussement, Koen W. De Bock, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees,European Journal of Operational Research,Volume 269, Issue 2,2018,Pages 760-772,ISSN 0377-2217

[4] I. Huseyinov and O. Okocha, "A Machine Learning Approach To The Prediction Of Bank Customer Churn Problem," 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2022, pp. 1-5, doi: 10.1109/IISEC56263.2022.9998299..

[5] S. AlZu'bi, A. Alsmadiv, S. AlQatawneh, M. Al-Ayyoub, B. Hawashin and Y. Jararweh, "A Brief Analysis of Amazon Online Reviews," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 555-560, doi: 10.1109/SNAMS.2019.8931816..

[6] C. Fry and S. Manna, "Can We Group Similar Amazon Reviews: A Case Study with Different Clustering Algorithms," 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2016, pp. 374-377, doi: 10.1109/ICSC.2016.71..

[7] C. Kumaresan and P. Thangaraju, "Ranking the Customer Reviews from Mobile Commerce Big data: K means Clustering," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 2022, pp. 1-6, doi: 10.1109/ICDCECE53908.2022.9792800.

[8] S. Kone, S. M. Farheen, B. Lokesh and T. S. Pavani, "A Novel Approach to Recommend Products in E-Commerce," 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, 2021, pp. 17-21, doi: 10.1109/ICISSGT52025.2021.00015.

[9] Salim Lahmiri, Stelios Bekiros, Christos Avdoulas,A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization,Decision Analytics Journal,Volume 6,2023,100166,ISSN 2772-6622

[10] D. G. I. Simanungkalit, B. Meylia, J. Salim, I. S. Edbert and D. Suhartono, "House Base-Price Prediction with Machine Learning Methods," 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), Jakarta Selatan, Indonesia, 2023, pp. 318-323, doi: 10.1109/ICIMCIS60089.2023.10349077.

[11] T. Patel, J. Sahu and M. R. K, "Comparative Analysis of Regression Techniques for Accurate House Price Prediction Using Machine Learning: A Statistical Analysis Method," 2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), New Raipur, India, 2023, pp. 1-6, doi: 10.1109/ICBDS58040.2023.10346417.

[12] U. Gupta, H. Vaidya, R. Chauhan and C. Bhatt, "House Price Prediction Using Gradient Boosting and Linear Regression," 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), Ghaziabad, India, 2023, pp. 829-832, doi: 10.1109/ICSEIET58677.2023.10303034.

[13] M. Waseem and S. Abidin, "Issues and Challenges of KDD Model for Distributed Data Mining Techniques and Architecture," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 1612-1616.

[14] bank-customer-churn-dataset.[Online] . Available : https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset. Last accessed : Dec 31, 2023.

[15] amazon-product-reviews dataset.[Online] . Available : https://www.kaggle.com/datasets/jillanisofttech/amazon-product-reviews . Last accessed : Dec 31, 2023.

[16] california-housing-prices dataset.[Online] . Available : https://www.kaggle.com/datasets/camnugent/california-housing-prices/data . Last accessed : Dec 31, 2023

[17] R Studio Available : https://posit.co/download/rstudio-desktop/