Data Mining and Machine Learning-1

Yash Bhargava
Dept. of Computing
National College of Ireland
Dublin, Ireland
x22220861@student.ncirl.ie

***Abstract*—This is a project for data mining and machine learning that includes analysing three datasets from different domains and applying five different models to these datasets. The first dataset is the bank marketing dataset, which is a dataset from the finance domain the second dataset the Seoul Bike dataset, is from the Transportation domain; and finally, the third dataset is from the Agricultural domain. I found these datasets from the UCI Machine Learning Repository, which is a huge library of datasets from every domain. The first dataset is our classification-based problem, in which I must classify, according to the bank marketing campaign held in Portuguese, whether a customer will buy term deposit insurance or not, to which I have applied Logistic Regression and Decision Tree algorithms. The second dataset is our regression-based Problem in which I must predict the rented bike count for the bike-sharing dataset of Seoul, to which I have applied the Linear Regression Model and the support vector for the regression (SVR) algorithm. The third dataset is again of classification based on the features of dry beans and to classify the class of dry beans among 7 different classes, I have used the Random Forest classifier.**

I. Introduction

Bank marketing refers to the methods and actions taken by financial institutions to sell their products and services to consumers. The banking sector is highly efficient for acquiring and retaining consumers in today's world.
The main purpose of bank marketing is to raise brand awareness, promote brands, and increase customer interaction. Understanding the requirements and preferences of target groups, building compelling marketing campaigns, and employing many channels to reach and interact with clients are all part of the process. The project involves various tasks on 3 datasets which are from different domains in which the first dataset is the bank marketing dataset for which a marketing campaign was held by a Portuguese banking institution which consists of 45,211 observations and 17 variables i.e. ("Age", "Job", "Marital Status", "Education", "Default", "Balance", "Housing Loan", "Personal Loan", "contact", "day", "Month", "Duration", "campaign", "pdays", "Previous", "poutcome", "y") of which our target variable (y) is term-deposit ("yes", "no") and it is a classification-based problem in which the research question for this dataset is:

Dataset 1: Bank Marketing Dataset

"Develop a prediction model that can predict whether a customer has purchased a term deposit based on the independent features provided in the dataset?"

Bike-sharing systems help us to provide sustainable urban transportation by reducing traffic congestion on the road and carbon emissions. By assessing and improving bike-sharing demand, we may directly contribute to the development of more ecologically friendly and efficient transportation alternatives. Predicting bike sharing demand with good accuracy is a difficult undertaking that involves machine learning and statistical modelling. Putting this onto a challenge can help me improve my skills in data preprocessing, feature engineering, and model selection, which are useful in different fields.
The second dataset is our Seoul Bike dataset, for which the survey was held in Seoul for analysis of bike sharing demand in the city. This dataset consists of 8,760 observations and 14 variables, i.e., ("Date", "Rented Bike Count", "Hour", "Temperature", "Humidity", "Windspeed", "Visibility", "Dew Point Temperature", "Solar Radiation", "Rainfall", "Snowfall", "Seasons", "Holiday", "Functioning Day") of which our target variable (y) is Rented Bike Count per hour, and it is a regression-based problem in which our research question is:

Dataset 2: Seoul Bike Sharing Dataset

"In this Seoul Bike Sharing demand dataset, predict the Rented Bike Count per hour based upon the independent features?"

Dry beans are a key component in many cuisines worldwide, supplying necessary minerals and protein. We can help improve crop output, quality, and resilience by studying and comprehending trends in the dataset. The dry bean dataset enables us to make significant contributions to agriculture, food production, and community well-being, while also linking my efforts with larger aims of promoting sustainability and solving food security challenges.
The third dataset is our Dry Bean dataset, for which various features are given for the prediction of the class of a dry bean, which consists of 13,611 observations and 17 variables that identify the type of bean it is. The feature's names are ("Area", "Perimeter"," Major Axis Length", "Minor Axis Length", "Aspect Ration", "Eccentricity", "Convex Area", "Equivalent Diameter", "Extent", "Solidity", "Roundness",

"Compactness", "Shape Factor 1", "Shape Factor 2", "Shape Factor 3", "Shape Factor 4", "Class"), in which our target y variable is Class as it is Multi-Class Classification problem for which our research question is as follows:

Dataset 3: Dry Bean Dataset

"Predict the class of the Dry bean based on the features in the dataset using a multiclass classification machine learning algorithm?"

The following is the sequence followed in this report structure. The second section of this report provides a short description of related work done in these datasets and this subject. In the third section of this report, the methodology I followed in my project is mentioned whether it is a KDD or CRISP-DM. In the fourth section of this report, I have mentioned all the analysis done in my project work and what machine learning algorithms I have applied and identified which models were giving the best accuracy measures for that research problem for each dataset. In the final section of this report, I have provided the final results and predictions of this research project and then given the future scope of my research project to get a better accuracy measure in future.

II. Related Work

For the first dataset i.e., the Bank Marketing dataset, there is a research study conducted by Archit V. [1] conducted in 2019 on evaluating the classification algorithms for handling the class imbalance problem on this dataset using WEKA which is a machine learning tool in this study the author applied various classification algorithms like Decision Tree, Naïve Bayes, Multilayer Neural Network, Support Vector Machine, Logistic Regression and Random Forest, the author found that after solving the class imbalance problem using various sampling techniques the precision of the minority class increases and SMOTE gives the highest precision value among all the other sampling techniques. In the case of Recall, the author found out that the recall of the minority class is highest with the SMOTE sampling technique while the two classification methods are Decision Tree and Random Forest which gives greater Recall value with Random Over Sampling of the Minority class which is greater than SMOTE's recall. In the case of the F1 Score, the author concluded that all the classification algorithms applied give the highest F1 score with the SMOTE sampling technique except Random Forest which gives a high F1 Score with Random oversampling of minority class. In the case of the ROC Area of the Minority class, SMOTE gives the highest ROC Area among all other sampling techniques for all classification algorithms applied except Random Forest which gives the highest ROC Area with Random oversampling of minority class. In the case of AUCPR, the author found out that the AUCPR of minority class for all the classification algorithms applied except Random Forest had given the best AUCPR value with SMOTE sampling technique while Random Forest gave the best AUCPR with Random oversampling of minority class technique.

Another research was conducted by Tuba P. and Songul K. A. in 2017 [2] which is about detecting the Important Features of Bank Marketing using Data Mining Techniques the authors used two Feature Selection methods i.e., the Information Gain and Chi-square and applied Naïve Bayes classification algorithm for which the evaluation metrics are Precision, Recall and F measure. The author found the Precision, Recall and F-measure by finding the number of top 5, 8, 10, and 15 features using the IG feature selection method and chi-square feature selection method. They finally concluded the top 10 best features on a ranking basis are "duration", "poutcome", "month", "pdays", "contact", "previous", "age", "job", "housing", and "balance". Another study was conducted by Maulida A.F. and Dany C.F. in May 2021 [3] which was about potential customer segmentation for data mining in the bank marketing dataset for which they first handled class imbalance problem using SMOTE and then the authors applied various classification algorithms such as Naïve Bayes, KNN, Random Forest, SVM, AdaBoost, for which they concluded that Random Forest has given the best accuracy value of 92.71% among all the classification algorithms.

Another research was conducted by Chittem L. K. and Poli V. S. R. in September 2019 [4] in which there were 5 models i.e., Decision Tree, Naïve Bayes, KNN, Support Vector Machine (SVM) and Backpropagation Neural Networks were applied. The authors have done Principal Component Analysis (PCA) for feature reduction and concluded that deep neural network classification algorithms outperformed all the other 4 classifiers in terms of accuracy.

Another study was conducted by Olatunji A. in 2016 [5] in which several models applied to the dataset were Logistic Regression, Decision Tree, Naïve Bayes and Random Forest ensemble and these algorithms were applied on both balanced and original datasets and the methodology used was CRISP-DM. The authors concluded that the best classification accuracy found on the balanced bank dataset was 76.6% which was of Decision Tree followed by Logistic Regression, Naïve Bayes were 75.7% and 75.6% respectively. The Random Forest had a classification accuracy of 74.2% when the number of trees was limited to 200. It was concluded that the Random Forest has a lower AUC value than the Decision Tree, the Random Forest does not improve the performance of the Decision Tree.

Furthermore, there was another study done by Safia A. in 2015 [6] was about deposit subscribe prediction on bank marketing in which the authors applied the Rough Set Theory which is for determining the CORE set of features responsible for informed decision-making i.e. (age, duration, Balance) which have a discriminant behaviour from other features and applied Decision Tree algorithm by which gain ratios have been obtained for each of the features using C4.5 classifier which shows that "Duration" has the maximum gain ratio, "age" feature has the 8th gain ratio and "Balance" has on

10th. Hence, they found that Rough Set Theory (RST) gives better summarization because of the feature reduction process and produces better accuracy in the Decision Tree.

TABLE I
RELATED WORK SUMMARY OF BANK MARKETING DATASET

Authors & Year	Algorithms
Archit Verma (2019)	Decision Tree, Naïve Bayes, Multilayer Neural Network, Support Vector Machine, Logistic Regression and Random Forest
Tuba P. et al (2017)	Naïve Bayes
Maulida A.F. et al (2021)	Naïve Bayes, KNN, Random Forest, SVM, AdaBoost
Chittem L. K. et al (2019)	Decision Tree, Naïve Bayes, KNN, Support Vector Machine (SVM) and Backpropagation Neural Networks
Olatunji A. (2016)	Logistic Regression, Decision Tree, Naïve Bayes and Random Forest
Safia A. (2015)	Decision Tree and Rough Set Theory (RST)

For the second dataset i.e., the Seoul Bike Dataset there was research conducted by Satishkumar V. E., Jangwoo P. and Yongyun C. [7] in 2020 in which the authors applied various models like Linear Regression, Gradient Boosting Machine, Support Vector Machines (SVM), Boosted Trees, XGBoost and the evaluation indices was taken as Root Mean Squared Error(RMSE), Mean Absolute Error(MAE), R squared(R2) and Coefficient of Variation(CV). They concluded that GBM and XGBTree outperformed the performance of the model as their R2 value is higher and RMSE and MAE values are lower which means that R2 described how the model fitted and the lower possibility of errors than the Support Vector Machine (SVM), Linear Regression and Boosted Trees whereas Temperature and Hour were considered as the most important variables in all the other models except Linear Regression model for the prediction of Rented Bike Count.
Another research was conducted in March 2022 by William K. D., Max K., and Pei-yin Y. [8] in which the models were applied to predict the Rented Bike Count were Time series models like ARIMA, FASTER, and non-parametric machine learning algorithms like Random Forest, XGBoost, Long Short Term Memory neural networks (LSTM) in which MAE and RMSE are taken as evaluation matrices. The authors concluded that the LSTM Lag 12 model outperformed with having the least MAE of 78 bikes per hour followed by Lag 24 and XGBoost. The baseline SNAIVE model also outperformed both the Time Series model i.e. ARIMA-based models.
Another research was conducted by Xin Xue L. and Chang L. in 2023 [9] in which they used stack-based ensemble modelling for the prediction of Rented Bike count and applied different Regression Models like Linear Regression, Ridge Regression, Lasso Regression, K- nearest Neighbour, Random Forest, Decision Tree Regression, Support Vector Machine and Gradient Boosting Decision Tree. Evaluation metrics that were considered were R-squared (R2), Explained variance score (EVS), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Median Absolute Error. Finally, the authors concluded that the stacking ensemble models including Decision Tree, Random Forest and GBDT non-linear regressors performed the best and non-tree-based non-linear regressors performed the worst. For the basic predictive model, Random Forest was the best among them.
Furthermore, research was conducted by Thu-Tinh T. N., Hue T.P., Juan A., and Sybil D. [10] in 2022 in which there were two regression models were applied Linear Regression and Random Forest in which the Random Forest outperformed the Linear Regression model as the value of R2 of Linear Regression model was 0.47 which was 50% less than that of Random Forest. The values of consideration are R2, RMSE, and MAE for which the values of RMSE and MAE of the Linear Regression Model were twice those of the Random Forest ones. The values of RMSE, MAE, and R2 of Random Forest are 210, 121, and 0.90 respectively which proves that Random Forest was superior in terms of predicting the Rented Bike Count of the Seoul bike-sharing dataset.

TABLE II
RELATED WORK SUMMARY OF SEOUL BIKE SHARING DATASET

Authors & Year	Algorithms
Satishkumar V. E. et al (2020)	Linear Regression, Gradient Boosting, SVM, Boosted Trees, XGBoost
Williams K.D. et al (2022)	ARIMA, FASTER, Random Forest, XGBoost, Long Short Term Memory Neural Networks (LSTM)
Xin Xue L. et al (2023)	Linear Model, Ridge Regression, Lasso Regression, KNN, Random Forest, Decision Tree Regression, SVM, GBDT
Thu-Tinh T. N. et al (2022)	Linear Model and Random Forest

The third dataset i.e., the Dry Bean dataset for which research was conducted in 2020 by Murat K. and Ilkar A. O. [11] did a study on Dry Beans and the authors applied different models for the multi-class classification of dry beans that whether they have among the 7 classes i.e. Barbunya, Bombay, Cali, Dermason, Horoz, Seker and Sira. The authors have applied multiclass classification algorithms i.e., Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), K-Nearest Neighbour (k-NN), Decision Tree (DT) for which the evaluation metrics considered are accuracy, Rate of Error, Sensitivity, Specificity, Precision, Recall and F1-Measure and found out that SVM performs the best with accuracy of 93.13% among the other models applied and the accuracy rates of prediction of each class i.e., Barbunya, Bombay, Cali, Dermason, Horoz, Seker and Sira were 92.36%, 100.00%,

95.03%, 94.36%, 94.92%, 94.67% and 86.84%, respectively.
Another research was conducted by Md Salauddin Khan et al [12] in 2023 in which the authors applied 8 different models i.e., Logistic Regression, Naïve Bayes, k-nearest Neighbour, Decision Tree, Random Forest, Extreme Gradient Boosting, Support Vector Machine, Multilayer Perceptron. The authors concluded that the Extreme Gradient Boosting Algorithm had performed the best with all the balanced and imbalanced dry beans datasets which was done by the ADASYN algorithm with an accuracy of 95.4% with balanced and 93% with imbalanced dataset. Using ADASYN, the performance of k-NN and Random Forest was also improved in terms of classification accuracy, Sensitivity, Specificity, and Cohen's Kappa coefficient.
Another research was conducted by S. Krishnan et al [13] in 2023 in which the authors applied 22 machine learning classification techniques so that they could identify the best ML models by using cross-validation to classify the class of dry beans and found out that the best classifiers are Random Forest, Extreme Gradient Boosting algorithm and CatBoost algorithm. They concluded that CatBoost performed the best with an accuracy of 94.25% with the balanced dataset and 92.76% with the unbalanced dataset whereas Random Forest had an accuracy of 92.06% with the unbalanced dataset and 93.29% with the balanced dataset while XGBoost got an accuracy of 92.10% with the unbalanced dataset and 93.57 with balanced dataset respectively.

TABLE III
RELATED WORK SUMMARY OF DRY BEANS DATASET

Authors & Year	Algorithms
Murat K. et al (2020)	Multilayer Perceptron (MLP), Support Vector Machine (SVM), k-Nearest Neighbour (KNN), Decision Tree (DT)
Md Salauddin et al (2023)	Logistic Regression (LR), Naïve Bayes (NB), k-nearest Neighbour (KNN), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), Multilayer Perceptron (MLP)
S. Krishnan et al (2023)	Random Forest (RF), Extreme Gradient Boosting (XGB), CatBoost
Jaime C.M. et al (2023)	Random Forest (RF), K-Nearest Neighbour (KNN), Support Vector Machine (SVM)

Furthermore, research was conducted by Jaime C. M. et al [14] in 2023 about data mining approaches to classify dry beans in which the authors have applied 3 machine learning classification algorithms i.e., Random Forest, Support Vector Machine, k-nearest Neighbour. The authors found that Random Forest and k-NN have an accuracy of 95.6% and 96% whereas SVM had an accuracy of 95.1% with the original dataset. After that, the authors did hyperparameter tuning and balancing of the dataset using SMOTE by which the accuracy was increased by around 2.6% for the k-NN classifier model getting the best accuracy of 95.9% respectively.

III. METHODOLOGY

The methodology I used in this project was the KDD methodology. Knowledge Discovery for Databases (KDD) was used to explore the data and do some preprocessing on the data followed by applying data mining algorithms to extract some valuable patterns and insights from the databases. It consists of 5 steps which are data selection, data preprocessing, transformation, data mining and evaluation of results which is shown in Figure 1. [15] It is further described below:

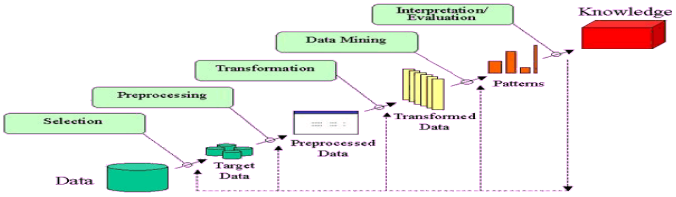

Fig. 1. Flow of KDD Methodolgy

A. Selection of data

This is the first step of the KDD methodology in which I must select all the datasets that I selected all three datasets from the UCI Machine Learning Repository that we use for performing further steps of KDD. In selecting the datasets, there should be a proper problem statement like a regression-based or classification-based dataset in which there should be clear problems for which we can find a solution by performing the further steps of KDD.

1) Dataset 1: Bank Marketing Dataset : This dataset was taken from the UCI Machine Learning Repository in which there were 2 files available in CSV format one was the bank in which there are 10 per cent of total observations i.e. 4521 rows and 17 columns and the other file was bank marketing full data which consists of 45,211 observations and 17 variables in which our dependent variable was y in which we have to predict whether the customer will subscribe to the term deposit or not while the other variables i.e., "Age", "Job", "Marital Status", "Education", "Default", "Balance", "Housing Loan", "Personal Loan", "contact", "day", "Month", "Duration", "campaign", "pdays", "Previous", "outcome" were the independent variables. The dependent variable y would have 2 levels if the customer had subscribed to the term deposit then it was indicated as 1 while if the customer doesn't subscribed it was indicated as 0.

2) Dataset 2: Seoul Bike Sharing Demand Dataset: This dataset was taken from the UCI Machine Learning Repository in which there was a bike sharing demand dataset which has been taken from Seoul which was available in the CSV file it was a regression-based dataset which consisted of 8,760 observations and 14 variables in which we have the dependent variable as Rented Bike Count and independent variables were "Date", "Rented Bike Count", "Hour", "Temperature", "Humidity", "Windspeed", "Visibility", "Dew Point Temperature", "Solar Radiation", "Rainfall", "Snowfall", "Seasons", "Holiday", "Functioning Day" based on which we predict our dependent variable. This dataset was about the bike-sharing demand in Seoul, North Korea which predicts the Rented bike count per hour based on the above-mentioned features.

3) Dataset 3: Dry beans Dataset: This dataset was about the classification of dry beans as it was a multi-class classification problem which I downloaded from UCI Machine Learning Repository which was available in CSV format it includes 13,611 observations and 17 variables in which the features taken from scanning the images of different dry beans and recorded the measurements in the 16 variables which were the independent variables i.e., "Area", "Perimeter"," Major Axis Length", "Minor Axis Length", "Aspect Ration", "Eccentricity", "Convex Area", "Equivalent Diameter", "Extent", "Solidity", "Roundness", "Compactness", "Shape Factor 1", "Shape Factor 2", "Shape Factor 3", "Shape Factor 4" and 1 dependent variable i.e., Class based on the above-mentioned features we will figure out the class of the dry bean i.e., "BARBUNYA", "BOMBAY", "CALI", "DERMASON", "HOROZ", "SEKER", "SIRA" respectively.

B. Data Preprocessing and Transformation of data

After the selection of all the datasets, the next step of KDD methodology was preprocessing and transforming the data which was a very important phase before building the final machine learning model as it will help the machine to better understand the data after performing this step it also increases the efficiency of the machine learning model. This step consists of doing exploratory data analysis (EDA) step i.e., checking for missing values, proper datatypes of variables, outliers in the dataset and cleaning of data i.e., treatment of missing values, outliers, handling of imbalance data, feature engineering on categorical variables and feature scaling of continuous variables and then finally splitting them into training on which the machine learning model was trained and then tested with the test data in ratio of 70:30.

1) Dataset 1: Bank Marketing Dataset: In this dataset, I first checked the structure of the dataset how many observations and variables were in the dataset like datatypes of the variables and out that all the variables had the correct data types then checked for the missing values and outliers and found that there are no missing values in the dataset, but I have found some outliers in the "balance" variables as it can be seen in Figure 2.

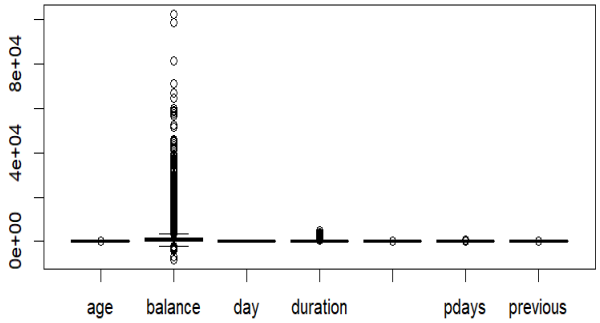

Fig. 2. Detection of Outliers for Dataset 1

Then further I have done the data cleaning of the dataset i.e., treatment of outliers. It was important to treat the outliers so that they would not affect the efficient performance of our model. After the treatment of outliers, I investigated the correlation between all the numeric variables in the dataset and it can be seen in the heatmap as seen in Figure 3 that the two variables "pdays" and "previous" had a very high intermediate positive correlation amongst them.

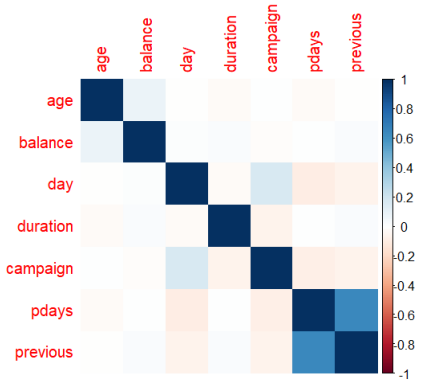

Fig. 3. Correlation Heatmap for Dataset 1

After figuring out the correlation between the variables, I further checked the balance of the dataset and whether it was a balanced or imbalance dataset, it was revealed that the dataset was highly imbalanced as it has 39,922 observations for "no" class and 5,289 observations for the "yes" class as it can be seen in Figure 4.

To resolve this problem, I used the Synthetic Minority Oversampling Technique commonly known as SMOTE and

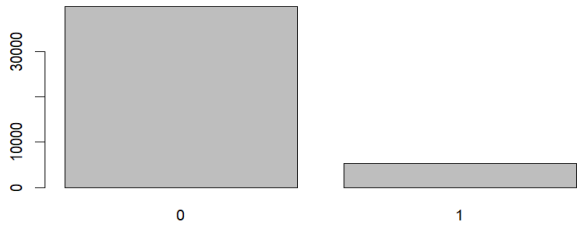

Fig. 4. Before Sampling of Dataset 1

did random oversampling of the minority class i.e., the "yes" class and created a balanced dataset of 55,923 observations and 17 variables now finally our dataset was balanced with 27,945 observations of "no" class and 27,978 observations of "yes" class as it can be seen in Figure 5.

Then I encoded the target variables which is y in the

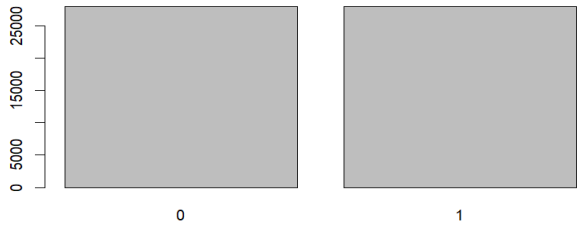

Fig. 5. After Sampling of Dataset 1

dataset as "0" and "1" and encoded the categorical variables i.e., "job", "marital", "education", "education", "default", "housing", "loan", "contact", "month" and "outcome". Further, I have split the dataset into training and testing sets and done the feature scaling of all the numeric variables in the training and testing set as well as transforming it for model building.

2) Dataset 2: Seoul Bike Dataset: This dataset, I first checked the structure of the dataset, and I found that the dataset has 8,760 observations and 15 variables I also checked for missing values and outliers and found that there are outliers in the "Rented Bike Count" variable and then I renamed some variables those have space in between the variable name i.e., "Rented Bike Count", "Wind Speed", "Dew point Temperature", "Solar Radiation", and "Functioning Day". So, in this dataset also there are no missing values and, I did the treatment of outliers as can be seen in Figure 6.

Secondly, I checked the correlations between all the numerical

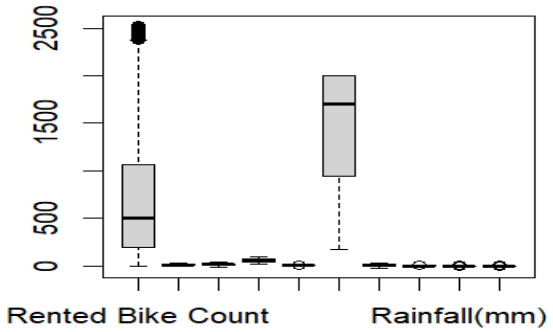

Fig. 6. After Treatment of Outliers for Dataset 2

variables using the heatmap as shown in Figure 7. Then as a

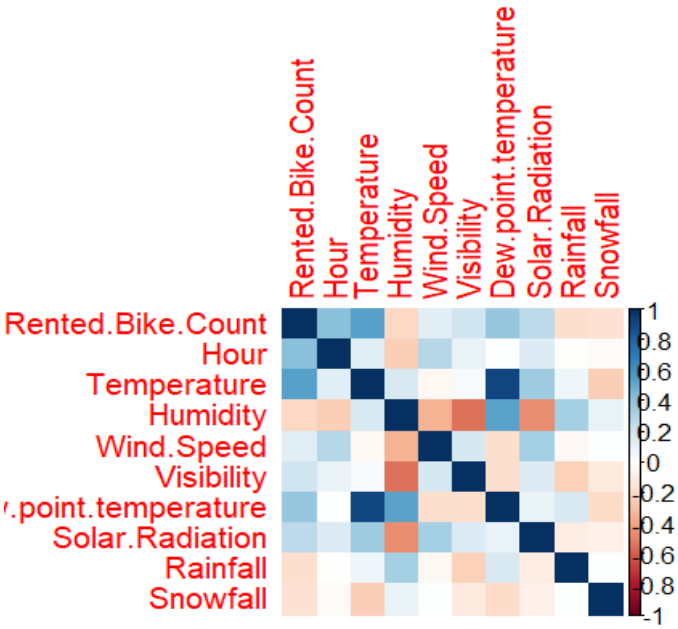

Fig. 7. Correlation heatmap for Dataset 2

most important step, I encoded all the categorical variables i.e., "Seasons", "Holiday", and "Functioning Day". Further, I split the data into training and test sets in the ratio of 70:30 and finally scaled all the numeric variables for both sets and transformed the sets in such a way that it was ready for building the model on the training set and testing the model on the test set.

3) Dataset 3: Dry Beans dataset: In this dataset, I first started checking out the structure of the dry bean dataset and the dataset had 13,611 observations and 17 variables, during

the exploratory data analysis I also checked if the dataset had any missing values and outliers in the variables, it was found out that there are no missing values in the dataset and I have also treated the outliers by defining a function named "rep_out" which did the treatment of outliers by replacing it with 1st and 99th percentile of values with outliers. After the treatment of outliers, I checked the correlation between the variables by plotting the heatmap as all the variables except the dependent variable "Class" are continuous and found out that the "Area" variable had a strong positive relation with "Perimeter", "Convex_Area", "Equiv_Diameter". The "Shapefactor3" had a strong negative correlation with "AspectRation" and "Eccentricity" as can be seen in Figure 8.

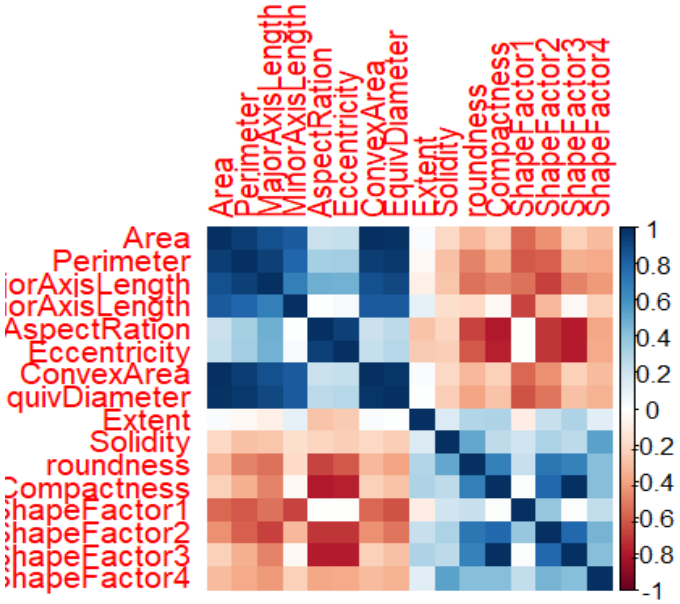

Fig. 8. Correlation heatmap for Dataset 3

Further, I have encoded the target variable i.e., "Class" as factors. Finally, after performing all the preprocessing steps and transformation I have split the dataset into training and test sets for model building which we further discussed in data mining and interpretation/evaluation steps.

C. Data Mining

This is the fourth step of KDD methodology which was done after preprocessing and transformation of data. In this step, I have selected the appropriate data mining algorithms based on the research question of the datasets. The two datasets were classification-based models i.e., the bank marketing dataset and dry beans dataset while the third dataset was a regression-based model i.e., Seoul bike data. It is an important step to select data mining algorithms as it is important in the extraction of knowledge from databases. I have applied 5 machine learning algorithms on the three datasets and did a comparison of their performance on each dataset by evaluating some key metrics discussed in further steps of KDD methodology.

1) Dataset 1: Bank Marketing Dataset: For this dataset, I have applied two classification machine learning algorithms i.e., Logistic Regression which was a very good classification model and Decision Tree classification algorithm. In this dataset, the main reasons for applying the Logistic Regression model were it was the best model as it performs well in case of binary classification problems, logistic regression perfectly handles both categorical and continuous variables of the dataset, and it was the computationally efficient machine learning algorithm as it can easily handle large datasets as our dataset has 45,211 observations. It also provides probabilistic prediction, allowing for checking the likelihood of an event occurring. Understanding the likelihood of a client subscribing to a term deposit might be useful for marketing decision-making. Secondly, I have applied the Decision Tree Classifier and my main intention for choosing the Decision Tree Classifier were it can easily handle complex datasets with non-linear relationships between the independent and dependent variables where the decision tree can easily outperform the linear classification models, it can easily handle imbalanced dataset on its own and could provide meaningful and valuable insights as it is a good practice to handle imbalanced dataset for the efficient performance of the model.

2) Dataset 2: Seoul Bike Dataset: In this dataset, I have applied two regression-based machine learning models i.e., Linear Regression and Support Vector Regression as it was a regression-based problem as I must predict the "Rented Bike Count" in this dataset. The main reason for choosing the Linear Regression model was the simplicity of this model as it was the easiest to apply among all the other regression models, it assumes that all the variables in the dataset have a linear relationship amongst each other, and the summary of Linear Regression model provides all the coefficients of all the independent variables with the importance of all the independent variables that how much each of the features were responsible for predicting the dependent variable. The reasons for choosing the Support Vector Regressor for this dataset were that it was a non-linear model and manages the non-linear relationship between the independent variables, as I have said this dataset had many variables and SVR performs well with handling the high dimensional spaces and works efficiently with the high number of variables and finally it can easily handle both regression and classification problems efficiently.

3) Dataset 3: Dry Beans Dataset: For this dataset, I have applied only one classifier i.e., the Random Forest Classifier as in this dataset I have to predict the class of dry beans based upon their features and it was the multi-class classification problem and the main reason for choosing the Random Forest

Classifier was that it performs extremely well with multi-class classification problems as compared to other classification algorithms and Random forest consists of multiple decision trees as it was ensemble learning algorithm which combines the prediction outputs of all the decision trees and provides the prediction based upon that of combined decision trees. Another rationale for choosing this model was it also takes care of feature importance for predicting the dependent variable based on the significance of features.

IV. Evaluation

This was the final step of the KDD methodology. In this step, all 5 models that are applied to the 3 datasets based upon their research questions in the previous steps of KDD were discussed with the results and the performance was evaluated based on some evaluation metrics i.e., for the regression-based problems i.e., Seoul bike dataset I have evaluated the performances of the regression models were listed below:

- **R-squared R2**- It is the evaluation measure for regression-based models which represents the proportion of variance of the predicted variable and the goodness of the fit of the model. The value of R-squared ranges from 0 to 1 and a higher value of R2 represents how well the model fits the dataset.
- **Mean Absolute Error (MAE)**- It represents the magnitude of the average difference between actual and predicted values of the dataset. The lower the value of MAE, the lesser the average error, better the model.
- **Mean Absolute Percentage Error (MAPE)**- This evaluates the average percentage difference between the actual and predicted values in the dataset. The lower the percentage of absolute errors the better the model.
- **Mean Squared Error (MSE)**- It represents the average of the squared difference between actual and predicted values. The lower value of MSE indicates better performance, on average, the squared differences between actual and predicted values are smaller.
- **Root Mean Squared Error (RMSE)**- It is a measure which is evaluated by taking the square root of mean squared error (MSE). The lower the value of RMSE better the performance of the model.

For the evaluation of classification-based modelling, we have considered important measures which are as follows:

- **Accuracy**- It is defined as the ratio of correctly predicted observations to the total number of observations in the dataset.
- **Confusion Matrix**- It consists of 4 metrics namely:
  - True Positives (TP)- The values that are positive and are correctly predicted as positive by the model.
  - False Positives (FP)- The values that are negative but are incorrectly predicted as positive by the model.
  - True Negatives (TN)- The values that are negative but are correctly predicted as negative by the model.
  - False Negatives (FN)- The values that are positive but are incorrectly predicted as negative by the model.
- **Precision**- It is defined as the ratio of true positive predictions to the total number of observations predicted as positive by the model.
- **Recall**- It is defined as the ratio of true positive predictions to the total number of actual positive observations.
- **F1 Score**- The F1 Score is the harmonic mean of precision and recall. It ranges from 0 to 1.

1) Dataset 1: Bank Marketing Dataset: In this dataset as I have discussed in the previous section, I have applied two classification models i.e., Logistic Regression and Decision Tree Classifier and I did the comparison of performance between the two models and found that Logistic Regression performs better than the Decision Tree Classifier on our balanced dataset which I have done using SMOTE. Logistic Regression has an accuracy of 83%, Precision of 0.96 and Recall value of 0.83 as can be seen in the confusion matrix of Logistic Regression in Figure 9

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10029   311
         1  1948  1276

               Accuracy : 0.8335
                 95% CI : (0.8271, 0.8397)
    No Information Rate : 0.883
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4431

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8374
            Specificity : 0.8040
         Pos Pred Value : 0.9699
         Neg Pred Value : 0.3958
             Prevalence : 0.8830
         Detection Rate : 0.7394
   Detection Prevalence : 0.7623
      Balanced Accuracy : 0.8207

       'Positive' Class : 0
```

Fig. 9. Confusion Matrix of Logistic Regression for Dataset 1

while the Decision Tree has an accuracy of 74%, Precision of 0.97 and Recall value of 0.72 and it can be seen in the confusion matrix of the Decision Tree in Figure 10.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 8671   201
         1 3306 1386

               Accuracy : 0.7414
                 95% CI : (0.734, 0.7488)
    No Information Rate : 0.883
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3231

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7240
            Specificity : 0.8733
         Pos Pred Value : 0.9773
         Neg Pred Value : 0.2954
             Prevalence : 0.8830
         Detection Rate : 0.6393
   Detection Prevalence : 0.6541
      Balanced Accuracy : 0.7987

       'Positive' Class : 0
```

Fig. 10. Confusion Matrix od Decision Tree for Dataset 1

Logistic Regression outperformed the Decision Tree classifier

and it better predicted whether a customer would buy a term deposit or not as compared to the Decision Tree.

2) Dataset 2: Seoul Bike Dataset: In this dataset, I have two regression-based models i.e., Linear Regression and Support Vector Regression in which we have found that Linear Regression performs better than Support Vector Regression as the Linear Regression model has lesser values of MAE, MSE, and RMSE i.e., 317.86, 173737.04 and 416.81 respectively as compared to Support Vector Regressor has got these values as 626.72, 674649.48 and 821.37 which was much greater than the final linear regression. In this dataset, I have created 4 linear regressions by selecting the significant features from each of the models by just checking the p-values and variance inflation factor (VIF) and the final evaluation metric values I got from the final Linear Regression Model have been compared with Support Vector Regressor and found out that Linear Regression model has lower values of evaluation metrics as that of SVR hence we can say that Linear Regression model outperformed the SVR and predicted better values of the dependent variable in this dataset. The summary of both the Regression models can be seen in Figures 11 and 12.

```
> summary(lr_regressor)

Call:
lm(formula = Rented.Bike.Count ~ Date + Hour + Temperature +
    Humidity + Wind.Speed + Solar.Radiation + Rainfall + Holiday +
    Functioning.Day, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-1272.2  -273.2   -46.9   206.2  1780.9

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.583e+04  1.148e+03 -13.793  < 2e-16 ***
Date               1.028e-05  7.475e-07  13.755  < 2e-16 ***
Hour               1.873e+02  5.825e+00  32.163  < 2e-16 ***
Temperature        3.311e+02  7.569e+00  43.743  < 2e-16 ***
Humidity          -1.364e+02  7.195e+00 -18.964  < 2e-16 ***
Wind.Speed         2.605e+01  6.090e+00   4.277 1.92e-05 ***
Solar.Radiation   -6.830e+01  7.153e+00  -9.548  < 2e-16 ***
Rainfall          -9.547e+01  5.716e+00 -16.703  < 2e-16 ***
Holiday1          -1.534e+02  2.626e+01  -5.844 5.35e-09 ***
Functioning.DayYes  8.660e+02  3.052e+01  28.375  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 422.5 on 6207 degrees of freedom
Multiple R-squared:   0.56,     Adjusted R-squared:  0.5594
F-statistic: 877.8 on 9 and 6207 DF,  p-value: < 2.2e-16
```

Fig. 11. Summary of Linear Regression Model for Dataset 2

```
> summary(svm_regressor)

Call:
svm(formula = Rented.Bike.Count ~ Date + Hour + Temperature + Humidity + Wind.Speed +
    Solar.Radiation + Rainfall + Holiday + Functioning.Day, data = training_set,
    type = "eps-regression")


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1
    epsilon:  0.1


Number of Support Vectors:  4077
```

Fig. 12. Summary of Support Vector Regression for Dataset 2

3) Dataset 3: Dry beans Dataset: In this dataset, I have applied only one machine learning algorithm i.e., Random Forest classifier which has an accuracy of 92%, Precision for Class 1,2,3,4,5,6 and 7 as 0.90, 1.00, 0.92, 0.90, 0.96, 0.93 and 0.88 respectively. I have also got Recall value for Classes 1, 2, 3, 4, 5, 6, and 7 as 0.90, 1, 0.91, 0.92, 0.95, 0.93 and 0.86, respectively. The F1 score I got for the Random Forest classifier was 0.91. In this classifier, I have chosen the number of decision trees as 10, but if we can increase the number of trees the accuracy and all the other evaluation metrics can also be increased. The confusion matrix as well as the classification report can be seen in Figure 13.

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5   6   7
         1 361   0  27   0   0   4   8
         2   0 157   0   0   0   0   0
         3  25   0 449   0  12   0   2
         4   0   0   0 984   7  20  74
         5   2   0   9   0 550   0   7
         6   3   0   1  18   0 571  15
         7   6   0   3  62   9  13 685

Overall Statistics

               Accuracy : 0.9199
                 95% CI : (0.9112, 0.9281)
    No Information Rate : 0.2605
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9032

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7
Sensitivity           0.90932  1.00000   0.9182   0.9248   0.9516   0.9391   0.8660
Specificity           0.98942  1.00000   0.9892   0.9666   0.9949   0.9894   0.9718
Pos Pred Value        0.90250  1.00000   0.9201   0.9069   0.9683   0.9391   0.8805
Neg Pred Value        0.99023  1.00000   0.9889   0.9733   0.9920   0.9894   0.9679
Prevalence            0.09721  0.03844   0.1197   0.2605   0.1415   0.1489   0.1937
Detection Rate        0.08839  0.03844   0.1099   0.2409   0.1347   0.1398   0.1677
Detection Prevalence  0.09794  0.03844   0.1195   0.2657   0.1391   0.1489   0.1905
Balanced Accuracy     0.94937  1.00000   0.9537   0.9457   0.9732   0.9643   0.9189
```

Fig. 13. Confusion Matrix of Random Forest for Dataset 3

V. Conclusion and Future Scope

The methodology used in this research was KDD in which we have done all the steps and after the completion of the analysis on the three datasets and the successful application of 5 different machine learning models, I can conclude that the results of the applied models were really good and here I can conclude from the results of the first datasets that Logistic Regression and Decision Tree classifier were applied in which Logistic Regression performed the best in terms of all the evaluation metrics that were considered and we can say that in future this model can helps in the bank marketing domain to predict the chances of buying a product or policies of the bank based on their features. In the second dataset, which was our Seoul bike dataset I applied Linear Regression and Support Vector Regression (SVR) in which the Linear Regression model performs the best in comparison to SVR as it has done better predictions of Rented Bike count per hour. In future, this model will help the transportation authorities and companies to investigate the use of sharing bike demand and take necessary actions to increase the sharing bike demand of customers. In our third dataset, which is our dry bean dataset I have applied a Random forest classifier with 10 trees which has given us a good accuracy of 92% in the future this model

can be used by the agriculturists to check the features of the dry bean and predict its class by using this model. Moreover, in the future, we can apply more advanced machine learning algorithms and check if other models can perform better and use other techniques for feature engineering and selection of crucial features like PCA, hyperparameter tuning and many more to achieve better outcomes.

REFERENCES

[1] A. Verma, "Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA," International Research Journal of Engineering and Technology (IRJET), vol. 06, pp. 54–61, Mar. 2019.

[2] T. PARLAR and S. K. ACARAVCI, "Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data", IJEFI, vol. 7, no. 2, pp. 692–696, 2017.

[3] M. A. Fitriani and D. C. Febrianto, "Data Mining for Potential Customer Segmentation in the Marketing Bank Dataset," JUITA: Jurnal Informatika, vol. 9, no. 1, p. 25, May 2021, doi: 10.30595/juita.v9i1.7983.

[4] C. L. Krishna and P. V. S. Reddy, "Deep Neural Networks for the Classification of Bank Marketing Data using Data Reduction Techniques," International Journal of Recent Technology and Engineering, Sep. 30, 2019. https://doi.org/10.35940/ijrte.c5522.098319

[5] O. Apampa, "Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction," CSUSB ScholarWorks. https://doi.org/10.58729/1941-6679.1296

[6] S. Abbas, "Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset," International Journal of Computer Applications, Jan. 16, 2015. https://doi.org/10.5120/19293-0725

[7] V. E. Sathishkumar, J.-W. Park, and Y. Cho, "Using data mining techniques for bike sharing demand prediction in a metropolitan city," Computer Communications, Mar. 01, 2020. https://doi.org/10.1016/j.comcom.2020.02.007

[8] Davis III, William K., Max Kutschinski, and Pei-Yin Yang. "Seoul Bike Demand Analysis." (2022).

[9] X. X. Lin and C. Lu, "A Stacking-Based Ensemble Model for Prediction of Metropolitan Bike Sharing Demand," American Journal of Information Science and Technology 7, pp. 62–69, 2023, [Online]. Available: http://article.ajinfoscitech.org/pdf/10.11648.j.ajist.20230702.13.pdf

[10] T.-T. T. Ngo, H. T. Pham, J. G. Acosta, and S. Derrible, "Predicting Bike-Sharing Demand Using Random Forest," Journal of Science and Transport Technology, pp. 13–21, May 2022, doi: 10.58845/jstt.utt.2022.en.2.2.13-21.

[11] M. Köklü and İ. A. Özkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," Computers and Electronics in Agriculture, Jul. 01, 2020. https://doi.org/10.1016/j.compag.2020.105507

[12] M. Salauddin Khan et al., "Comparison of multiclass classification techniques using dry bean dataset," International Journal of Cognitive Computing in Engineering, vol. 4, pp. 6–20, Jun. 2023, doi: 10.1016/j.ijcce.2023.01.002.

[13] S. Krishnan, S. K. Aruna, K. Kanagarathinam, and E. Venugopal, "Identification of Dry Bean Varieties Based on Multiple Attributes Using CatBoost Machine Learning Algorithm," Scientific Programming, vol. 2023, pp. 1–21, Apr. 2023, doi: 10.1155/2023/2556066.

[14] J. C. Macuácua, J. A. S. Centeno, and C. Amisse, "Data mining approach for dry bean seeds classification," Smart Agricultural Technology, vol. 5, p. 100240, Oct. 2023, doi: 10.1016/j.atech.2023.100240.

[15] S. Chumbar, "KDD Process in Data Science: A Beginner's Guide - Shawn Chumbar - Medium," Medium, Sep. 22, 2023. https://medium.com/@shawn.chumbar/kdd-process-in-data-science-a-beginners-guide-426d1f0fc062