

Data Mining and Visualisation Report

by Gowri Pamarthi

Submission date: 04-May-2023 07:51PM (UTC+0100)

Submission ID: 204773308

File name: 42643_Gowri_Pamarthi_Data_Mining_and_Visualisation_Report_493224_1312409871.docx
(473.41K)

Word count: 4532

Character count: 23348



Student Name: GOWRI PAMARTHI

Student Number: 2206771

Module Name:

Data Mining and Visualisation (ACCA7015)

Assignment Title:

Data Mining and Visualisation Report

Table of Contents

Table of Contents	2
List of Figure	3
1 Introduction	4
1.1 Primary and Secondary Data Pre-Processing	5
1.2 Dataset Description	5
1.3 Data Exploration	6
1.4 Data Analysis	8
2 Results and Analysis	10
3 Conclusion and Recommendation	21
3.1 Conclusion	21
3.2 Recommendation and Future Scope	22
4 References	23

List of Figure

Figure 1.1 Weka working procedure	4
Figure 2.1 Dataset Used	10
Figure 2.2 Storage of the dataset in the internal storage of the PC	11
Figure 2.3 Old and New Datasets stored in the internal memory of the pc	11
Figure 2.4 Initially observable Initialisation	11
Figure 2.5 Main Screen After importing The Dataset	12
Figure 2.6 non-Null values in Column	12
Figure 2.7 Visualise All The Graphs	13
Figure 2.8 Attribute Selection	13
Figure 2.9 After Applying Normalisation	14
Figure 2.10 Apply Class Balancing	15
Figure 2.11 Data Splitting	15
Figure 2.12 Target column	15
Figure 2.13 Output Summary of Implementation of random forest classifier	16
Figure 2.14 Output Summary of Implementation of Decision Tree Classifier	16
Figure 2.15 Test mode of Dataset	17
Figure 2.16 Visualisation Screen Initial Output	17
Figure 2.17 Visual output of job vs age column	18
Figure 2.18 Plot of Duration vs Campaign	18
Figure 2.19 Visual output of loan vs housing column	19
Figure 2.20 Visual output of campaign vs duration	19

1 Introduction

Weka, one of the best tools for data visualization and modeling, enables people and businesses to move towards a data-driven culture [1]. It is from the group of machine learning (ML) algorithms that is used mostly for data mining. Weka is a tool that creates models and visualization graphs devoid of any coding. This tool is capable enough of providing the data analysis activities such as classification, grouping, visualization, regression, etc. [2]. This application is easily accessible to everyone because it is open-source software. Using this software, the visualizations are considered at the initial stages and then considered for the obtaining the outputs based on the consistency of the data in the dataset. For satiating system demand, it is helpful. The dataset is often in the form of Excel, CSV, or another type of file; following pre-processing, it will be ready for output generation. Weka has been selected as the algorithm that might be utilized to determine which data mining technique would produce the ideal results for the system's graph visualization [3].

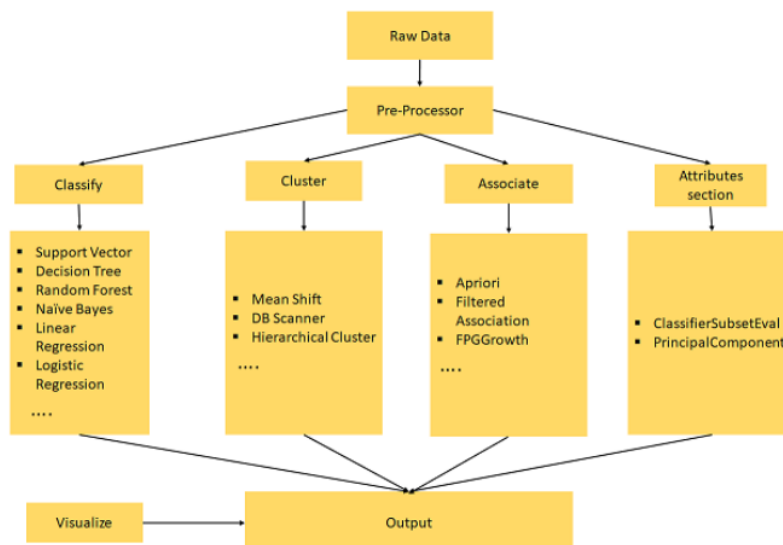


Figure 1.1 Weka working procedure

Source: Own Diagram.

To develop the requisite Machine Learning models, big data may be used. Here are the excellent field data that were collected. This data might contain several null values and extra fields. if the author uses the WEKA data pre-processing tools to clean the data. The author should save the results of these pre-processing tools in local storage to use ML algorithms. Based on these algorithms the models are built after pre-processing [4]. Features can be

automatically picked to produce a reduced dataset with the aid of attribute selection. Recognize that WEKA provides the implementation of a wide range of algorithms for each category. The desired parameters would be specified, an algorithm would be selected, and the algorithm would be applied to the dataset. It provides a tool for data exploration and visualization [5]. Different models can be tested using the same dataset. After comparing the outcomes of many models, the author can select the one that best meets their requirements. Therefore, using WEKA accelerates the creation of machine learning models as a whole. In the chapter that follows, we'll learn how to install WEKA on a local computer after examining what it is and what it can do.

1.1 Primary and Secondary Data Pre-Processing

A dataset may consist large amount of unnecessary information which may produce lot of incorrect data. Inconsistency in data, null values and irrelevant columns tend to create massive impact on the final output. Data processing facilitates output visualization for data inspection in many formats. The processing data means after cleaning the data removing the null, duplicate, and data type conversion values are helpful for visualization. Different models can be pre-processed for the next steps that include visualization and model building activities. Different outputs from all ML modes are compared to see the best-performing model with good accuracy to proceed further. The obtained results will be retrieved for the SQL formats [6], and the dataset in this work is related to market the campaigns of a banking institution in Portuguese [7]. This proposed project based on market campaigns was based on the phone class. In this dataset, deciding whether a product would be accepted or rejected required several interactions with the same client [8]. The reason for collecting data is to develop the product for plotting the various graphs.

1.2 Dataset Description

If the dataset consists of quality data using Weka to create the best algorithms for visualization is possible. the processing of the algorithm initially uses the researcher should give proper data for using the best algorithm as per the view of data [9]. The processes of the data are used to check the quality of the dataset and it is also used for transforming values towards different activities of statistical analysis. There is a need to check the dataset quality by systems to ensure that they are ready to apply for the processing of the system to get system requirements.

In this work, the dataset consisting of parameters based on these parameters is used to identify a stable method to generate suitable results. The marketing on banking dataset consists of 20

different parameters and 45211 instances. These 20 attributes are decided target column results. Here in this dataset, everyone can observe age, job house loan, employee variation rate, etc these types of parameters, and how they impacting on the target variable. This dataset consists of fully required parameters based on this kind of dataset everyone can easily create the best algorithm for visualization.

Here this dataset contains various attributes like age, job, housing, loan, contact, month, day, etc. This also explains different forms used to identify best method to analyse marketing trends of the banks [10]. There are different types of methods for having a consistency of maintaining the dataset, and it is an important task to analyse the content in the dataset.

1.3 Data Exploration

Selected Attribute

Name: Age

Missing: 0 (0%)

The percentage split 75 %

Evaluation of Random Forest implementation dataset

Correlation Coefficient 0.91140

Mean absolute error 0.1166

Root mean squared error 0.2429

Relative absolute error 0.5833 %

Root relative squared error 0.7682 %

Total number of Instances 41188

Evaluation of Decision Tree implementation dataset

Correlation Coefficient 0.8873

Mean absolute error 0.1697

Root mean squared error 0.2913

Relative absolute error 0.8485 %

Root relative squared error 0.9211 %

Total number of Instances 41188

Age:

The age of the customer is presented here it is a numerical column. the data range between 17 to 98

Job:

The type of the job is the categorical column here 12 classes such as admin blue collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, and unknown.

Marital:

marital status of the customer is also a categorical column such as divorced, married, single, and unknown.

Education,

The customer education status is a categorical column such as basic.4y, basic.6y, basic.9y, high school, illiterate, professional course, university degree, and unknown.

Default:

has credit in default or not it is also categorical such as yes and no.

Housing:

Whether a customer has a housing loan or not is also categorical such as yes and no.

Loan:

Whether a customer has a personal loan or not is also categorical such as yes and no.

Contact:

Here in this column, the customer type of contact communication is also categorical the classes such as cellular and telephone.

Month:

last contact month of the year here the month is classification January, February, up to December.

Day of the week:

last contact day of the week here is the week's classification such as Sunday, to Saturday.

Duration:

last contact duration, in seconds the data ranges between 0 and 4918.

Campaign:

number of contacts performed during this campaign and for this client the data ranges between 1 and 56

previous days:

number of days that passed by after the client was last contacted from a previous campaign the data ranges between 0 and 999

previous:

number of contacts performed before this campaign and for this client the data range between 0 and 7

previous outcome:

the outcome of the previous marketing campaign is categorical such as failure, non-existent, and success.

employee variation rate:

employment variation rates the data range between -3.4 and 1.4

the consumer price of the index:

consumer price index the data range between 92.021 and 94.767

the consumer confidence index:

consumer price index a monthly indicator the data range between -50.8 and -26.9.

Euribor 3-month rate:

Euribor 3-month rate a daily indicator of consumer confidence indexes the data range between 0.634 and 5.045.

number of employees:

number of employees a quarterly indicator of the data range between 4963.6 and 5228.1.

y:

number of employees has two classes such as yes and no.

1.4 Data analysis

This process uses a decision tree and a random forest classifier to provide the output for the system. The technique makes use of the ML algorithm, which might be used to keep visualizing the required amount of data. Here, decision trees or random forest files are utilized to create the files which are used for categorization based on the system's requirements. To offer the

necessary format while studying the system's visualization process to ascertain the various values linked with it, the decision tree's bagged form is employed. The system also makes use of a decision tree to provide a procedure for submitting a method for validating information. Weka used its own decisions to generate the visualization based on the column values that is used to identify values that are required by system. This helps to decide values that are created based on their values that are used in data implementation by using decision tree capabilities [11]. This might be used to generate values for the application by entering them there and executing the visualization process, allowing it to be recognized for the data visualization.

2 Results and Analysis

The given dataset is about the direct marketing campaigns of a Portuguese banking institution refers to the targeted promotional activities that the bank engages in to encourage customers to open new accounts, apply for loans, and invest in financial products. These campaigns are typically conducted through direct mail, phone calls, email, or other forms of direct communication. This application can be developed by using the tool named Weka. Weka is one of the most widely used tools this tool is mainly used for data pre-processing and data mining. Weka provides a user-friendly graphical user interface for exploring and analyzing data, as well as a suite of powerful algorithms and tools for building predictive models. Weka supports a wide range of data formats, including CSV, ARFF, and JSON. It also allows users to create their own custom plugins and extensions, enabling them to extend the functionality of the tool and integrate it with other software systems.

age	job	marital	education	default	housing	loan	contact	month	day_of_w	duration	campaign	pdays	previous	poutcome	emp.var.r	cons.price	cons.conf	euribor3m	nr.employ	y
56	housemai	married	basic.4y	no	no	no	telephoni	may	mon	261	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
57	services	married	high.schoi	unknown	no	no	telephoni	may	mon	149	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
37	services	married	high.schoi	no	yes	no	telephoni	may	mon	226	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
40	admin.	married	basic.6y	no	no	no	telephoni	may	mon	151	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
56	services	married	high.schoi	no	no	yes	telephoni	may	mon	307	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
45	services	married	basic.9y	unknown	no	no	telephoni	may	mon	198	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
59	admin.	married	profession	no	no	no	telephoni	may	mon	139	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
41	blue-colla	married	unknown	unknown	no	no	telephoni	may	mon	217	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
24	techniciar	single	profession	no	yes	no	telephoni	may	mon	380	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
25	services	single	high.schoi	no	yes	no	telephoni	may	mon	50	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
41	blue-colla	married	unknown	unknown	no	no	telephoni	may	mon	55	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
25	services	single	high.schoi	no	yes	no	telephoni	may	mon	222	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
29	blue-colla	single	high.schoi	no	no	yes	telephoni	may	mon	137	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
57	housemai	divorced	basic.4y	no	yes	no	telephoni	may	mon	293	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
35	blue-colla	married	basic.6y	no	yes	no	telephoni	may	mon	146	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
54	retired	married	basic.9y	unknown	yes	yes	telephoni	may	mon	174	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
35	blue-colla	married	basic.6y	no	yes	no	telephoni	may	mon	312	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
46	blue-colla	married	basic.6y	unknown	yes	yes	telephoni	may	mon	440	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
50	blue-colla	married	basic.9y	no	yes	yes	telephoni	may	mon	353	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
39	managem	single	basic.9y	unknown	no	no	telephoni	may	mon	195	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
30	unemploy	married	high.schoi	no	no	no	telephoni	may	mon	38	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
55	blue-colla	married	basic.4y	unknown	yes	no	telephoni	may	mon	262	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
55	retired	single	high.schoi	no	yes	no	telephoni	may	mon	342	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no
41	techniciar	single	high.schoi	no	yes	no	telephoni	may	mon	181	1	999	0	nonexiste	1.1	93.994	-36.4	4.857	5191	no

Figure 2.1 Dataset Used

Figure 2.1 describes the dataset used in the Weka tool. Weka tool is one of the most widely used tools this tool is mainly used for data pre-processing and the visualization purpose. Users can easily build the machine learning model based on the requirements of the dataset without any coding user plot the different graphs and build the best model within a short time. The given dataset is used in the Weka tool. The dataset is related to the digital marketing campaigns with the Portuguese bank. There are several columns are displayed in this dataset. The problem

statement of this dataset is to predict whether the client will subscribe to a term deposit or not. In the output variable having the values yes or no it is a classification problem.

C:\Users\GOWRI\Downloads\bank-additional-full.csv"

Figure 2.2 Storage of the dataset in the internal storage of the PC

Figure 2.2 describes the path of the dataset used to store it in the computer. After storing the dataset, the next step is the load the dataset into the Weka tool for loading the dataset into required tool first open the tool and click on the open file option automatically open the window box displayed in that window select the required path and select the required file and click on open button. Automatically the required file will be loaded into the Weka tool.



Figure 2.3 Old and New Datasets stored in the internal memory of the pc

Figure 2.3 represents the data that will be stored in different files one file is in CSV format and the other one is in Arff format. While using the Weka application the files which are loaded in the tool will be in the .arff extension for that first download the CSV file and convert that CSV file into an arff file by clicking and saving the CSV file again. After clicking and saving, the CSV file will be converted into an arff file and stored in the system's internal memory.

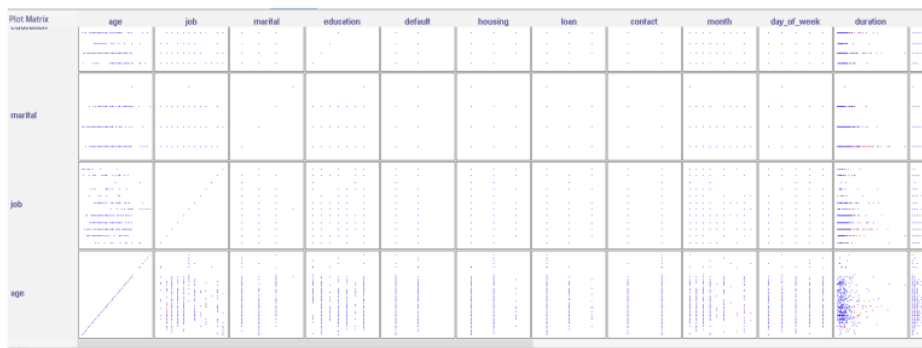


Figure 2.4 Initially observable Initialization

Figure 2.4 describes After loading the dataset into the Weka tool go to the visualize option and click on it automatically visualization of the required dataset will be loaded. These visualizations are not created manually these are updated or loaded automatically after loading the dataset. A scattered plot will be displayed based on the features of the dataset.

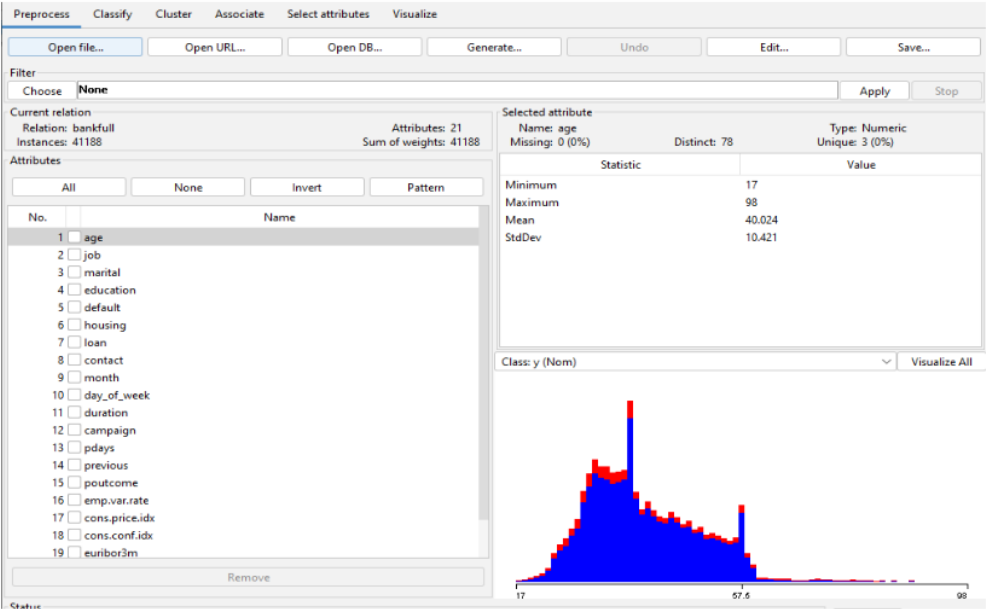


Figure 2.5 Main Screen After importing The Dataset

Figure 2.5 Describes the main screen after importing the dataset. Once import the dataset on the Weka tool it will display all the available columns in the dataset and also displays the minimum value and maximum value in each column and also displays the non-null values percentage and variance and std values etc. All this information can be useful for the analysis of the dataset by using this information user get the minimum knowledge about the dataset. For better understanding, all these are very useful.

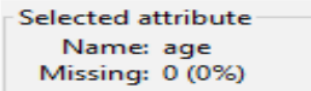


Figure 2.6 non-Null values in the Column

Figure 2.6 Describes the non-null values in the columns. This information can be useful for finding the null values and non-null values in the given dataset. Non-null values could be useful to provide better and more efficient results to the user.

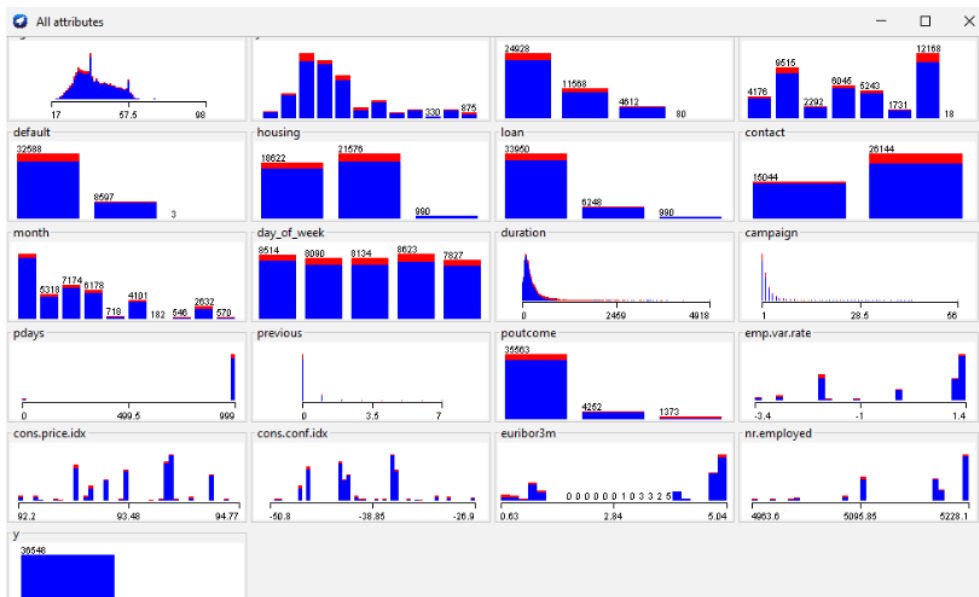


Figure 2.7 Visualise All the Graphs

After loading the source file into the Weka tool, the right side of the window can visualize the graphs of the columns to visualize all the graphs of the columns then click on the visualize all option button then automatically display all the available graphs of the dataset. The above figure displays all the available graphs of the columns in the specific dataset. The main advantage of the visualize all option is the user can able to see all the available columns in one window.

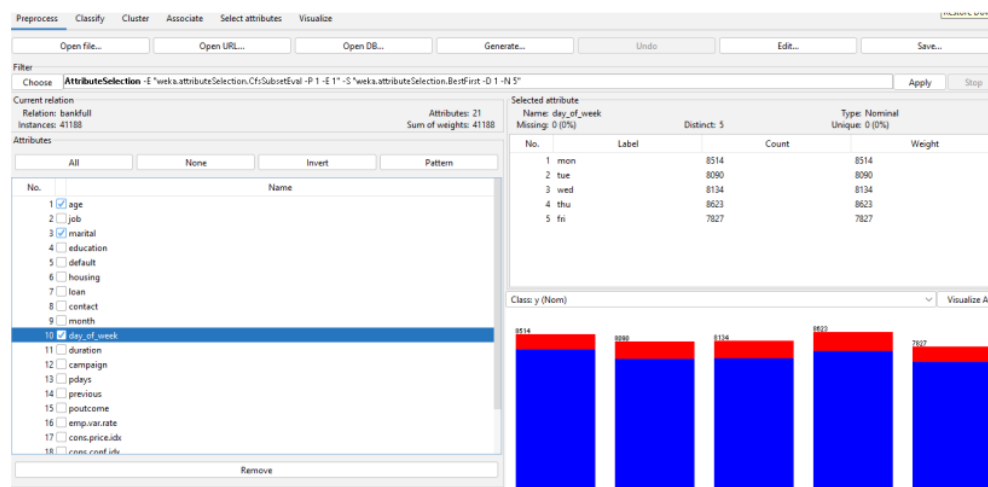


Figure 2.8 Attribute Selection

Figure 2.8 shows how to select the required attributes from the given dataset. For building any machine learning algorithm attribute selection is the most important feature. In the given dataset there are many columns available out of all the columns there is a need to select the specific attributes based on the problem statement. Attribute selection can be useful for giving the best output. For attribute selection users first, select the choose option in the Weka home window after that click on the filters option, and then next go to the attributes folder where many options are available out of all the options select the attribute selection option. After selecting the required option then select the columns that you required and click on the apply button. Automatically the selected features will remain and the remaining attributes are removed. After applying the changes click on the save button.

Selected attribute	
Name: age	Type: Numeric
Missing: 0 (0%)	Unique: 3 (0%)
Distinct: 78	
Statistic	Value
Minimum	0
Maximum	1
Mean	0.284
StdDev	0.129

Figure 2.9 After Applying Normalization

Normalization is the most important step for building any machine learning model. Normalization can be applied to the dataset to convert all the columns into one scale. If don't apply the normalization step then the user got a scaling problem so to escape from this problem needs to normalize the required columns in the dataset. The steps included in this process are just to select the choose option on the left side of the Weka window. And click on the filters option one sub-window will be displayed in that window select the unsupervised option and click on the normalize option. After that select the features and click on the apply button automatically all the column values in the dataset will be converted into one scale of 0 and 1. The above figure display that after applying the normalization age column the values will be converted into minimum value into 0 and maximum value into 1.

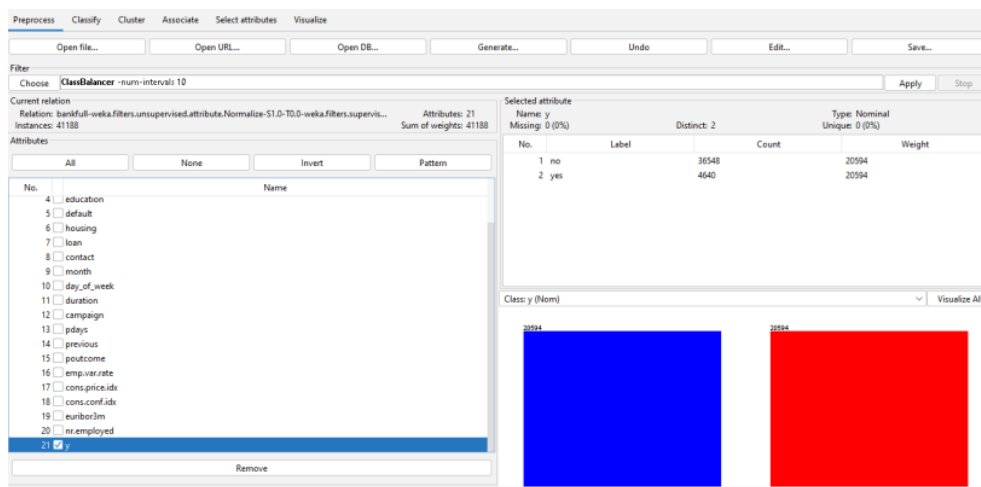


Figure 2.10 Apply Class Balancing

Above figure 2.10 displays the figure after the class balancing in the target column. The imbalance dataset causes the main problem most of the time imbalance dataset may not give accurate or correct results. So, to resolve all the problems need to balance the target column for that click on the choose option and select the supervised option and one pop-up will be displayed Select the required option like class balancer after that click on the specific column in which column needs to apply and select on the apply button and click on the save button. An automatically imbalanced dataset is converted into balance.

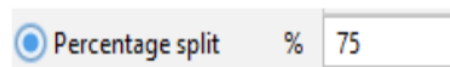


Figure 2.11 Data Splitting

Figure 2.11 Describes the Splitting of the dataset. Splitting is used to predict the results efficiently. Splitting the dataset can be used for identifying the process and performing the different methods based on the performance of the dataset.

No.	Label	Count	Weight
1	no	36548	36548
2	yes	4640	4640

Figure 2.12 Target column

The above figure 2.12 Describes the target column in the given target column there are two categories yes or no it is a classification process this information gives the idea to the required users which algorithms can be used in the dataset. Based on the performance models user can change the algorithms or apply the different classification algorithms.

```

=== Summary ===

Correctly Classified Instances      37539      91.1406 %
Incorrectly Classified Instances    3649       8.8594 %
Kappa statistic                    0.5004
Mean absolute error                 0.1166
Root mean squared error             0.2429
Relative absolute error             58.3319 %
Root relative squared error         76.8227 %
Total Number of Instances          41188

```

Figure 2.13 Output Summary of Implementation of random forest classifier

Figure 2.13 Displays the results for the random forest classifier algorithm. This algorithm generates a 91% accuracy based on the values of the columns in the dataset. For better accuracy, the user needs to apply different machine-learning algorithms to improve the performance of the model.

```

=== Summary ===

Correctly Classified Instances      36548      88.7346 %
Incorrectly Classified Instances    4640      11.2654 %
Kappa statistic                    0
Mean absolute error                 0.1697
Root mean squared error             0.2913
Relative absolute error             84.8517 %
Root relative squared error         92.1183 %
Total Number of Instances          41188

```

Figure 2.14 Output Summary of Implementation of Decision Tree Classifier

Figure 2.14 Displays the results of the decision tree classifier this model is used to increase the performance of the model. The decision tree classifier algorithm has an accuracy of 88%. Based on the given dataset and columns in the dataset decision tree classifier will give very accurate and efficient results.

```

Test mode:      10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

```

Figure 2.15 Test mode of Dataset

Figure 2.15 Displays the information regarding the training data what model applied and how many cross-validations are used and how much time will it take to run the model. In the given dataset user used the 10-fold cross-validation in the training dataset for better accuracy and to increase the performance of the model. And it takes 100 iterations along with bagging.

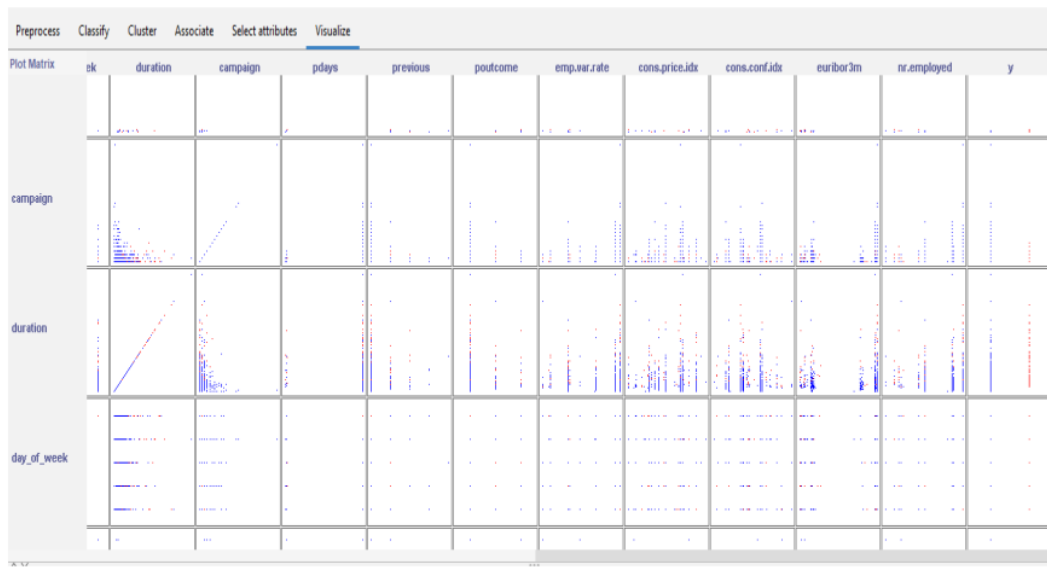


Figure 2.16 Visualization Screen Initial output

Figure 2.16 Displays the visualizations of the required dataset. For every machine learning algorithm visualisation plays a very important role in building the model. In the visualization part, the user has a better understanding of the dataset and the user got the knowledge about which are affecting the target column. Plotting graphs gives a better understanding and idea about the dataset.

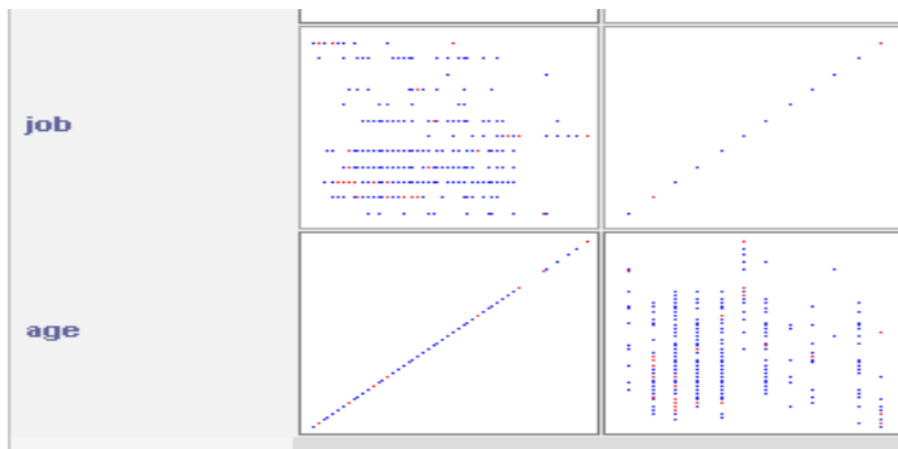


Figure 2.17 Visual output of job vs age column

Figure 2.17 Describes the plots generated based on the columns of job and as well as age column. This information is mainly used for identifying the user requirements in the system and is also useful for the internal process or working of the verification of the system.

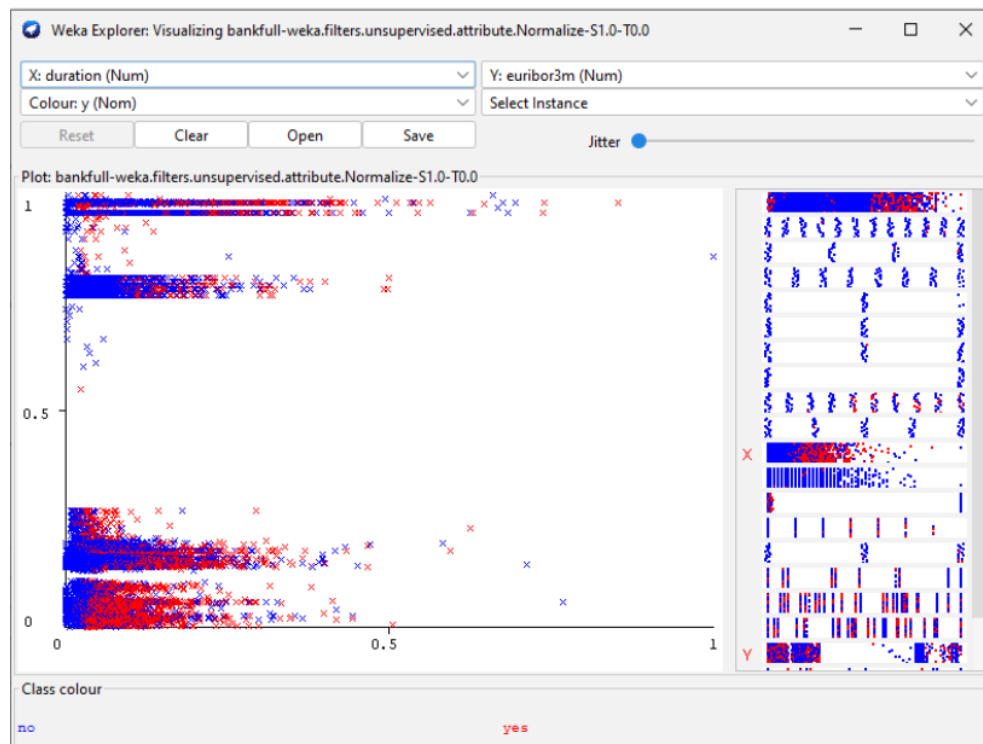


Figure 2.18 Plot of Duration vs Campaign

The above plot 2.18 Describes the information regarding the duration of the camping in the different places. If the user clicks on any data point it displays the information regarding the particular column. By using this information users get the idea of which place will it take the most time for camping. This information will be useful for future predictions.

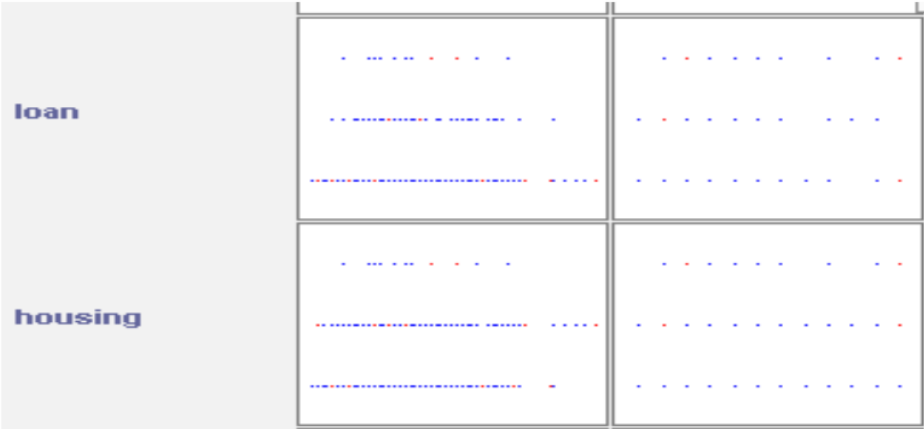


Figure 2.19 Visual output of loan vs housing column

Figure 2.19 Describes the generated plots between the columns in the given dataset. Using the basic visualization plots user got the knowledge about the values in the columns how it's varied based on the target columns and the effect on the target column all this information can be useful for building a better machine learning model.

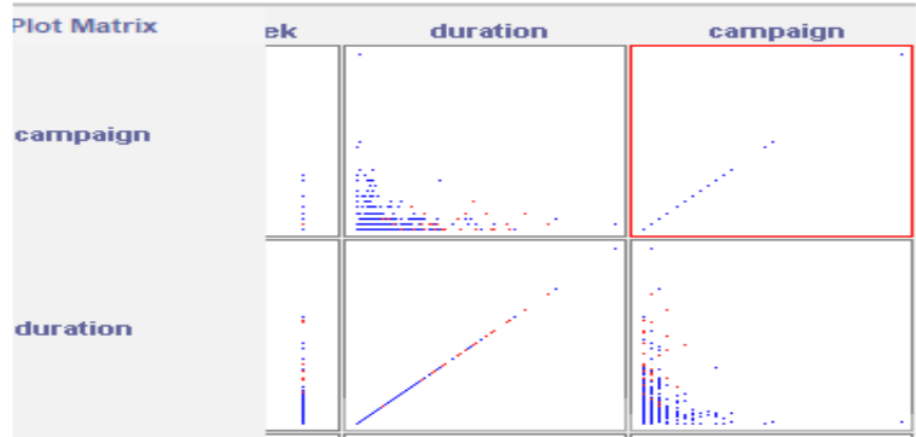


Figure 2.20 Visual output of campaign vs duration

Figure 2.20 Describes the visualizations of the columns like the campaign and duration of the column. Based on these plots get information like how the data can be varied from place to

place and also know how much time will it take for the campaign in different places. How the time will vary based on location can be easily understandable by the user. This information can cause future predictions of better decision-making about the campaign.

3 Conclusion and Recommendation

3.1 Conclusion

Here the author used WEKA tools to visualize all the data information because it is an ML platform where without too much coding knowledge a person can complete any ML project. That's why the author chooses this platform to complete this work study. While trying to implement some of the visualization and model parts the author faces some major issues or difficulties such as dataset upload, finding the proper visualization part as well as model implementation part. After some time, the author applies proper visualization to extract some of the information to represent the dataset attribute or any hidden information. Because with the help of visualization, it is possible to represent the data more easily so that everyone can easily understand the data information. Here in this dataset, there is no null value as well as no duplicate values also so no need to apply any data cleaning part. In real-time the entire time user may not get a good dataset the dataset may contain null values and duplicate values. For pre-processing that type of dataset no need to write the code just go to choose an option and select the filter and click on the supervised in that click on the instance folder and select the required options like clean null and remove duplicate values automatically the dataset will be converted into a clean one. Apart from that there all the attributes are neat and clean, that's why there is no required data transformation and derived columns because all the attributes are representing a good quality of information. Here the author used a scattered plot to show or find the relationship between each attribute. Apart from the scatter plot user wants to see the histograms in the given dataset for that after loading the dataset into a tool on the left side of the window displays the plots if they need to see all the histograms available in this dataset, then click on the visualize all option in the left side of the window, then automatically it will display all the available plots. The author chooses this plot because it is easy to implement and understand to find the relationship between all the attributes. Apart from that here the author finds some of the other basic information of the dataset attribute to gain or better understand the data. After applying proper visualization techniques, the author implemented different techniques of classification algorithms. But there are multiple classification algorithms available in ML, out of them the author chooses Random Forest (RF) and Decision Tree (DT) because they are considered the highest predicting model among all algorithms. After applying the RF model, the author got 91 percent of accuracy and by applying the DT model the author got 89 percent of accuracy. As it is a classification problem that's why the author considered

accuracy as the evaluation matrix and finalize the RF model as the best suitable model for this problem statement.

3.2 Recommendation and Future Scope

As the author applied only a scatter plot to extract the relationship between each column and here applied only 2 algorithms for model selection, so, there is a chance that maybe the author got higher accuracy by applying different algorithms also. So, it is recommended to apply more visualization techniques and more algorithms for better data information and accurate result. Apart from that there is a chance in the future, that the bank organization may use these techniques to analyze their data because it is very easy to implement and understand. So, with less effort, they can analyze their data to increase their business as well as generate more revenue.

4 References

- [1] Concha. B, and P. Larrañaga. Data-driven computational neuroscience: machine learning and statistical models. Cambridge University Press, 2020.
- [2] Suraj. N, D. Debnath, P. Sarkar, and A. Biswas. "An Efficient Implementation of Data Mining Techniques in Agriculture." *Journal of Computational and Theoretical Nanoscience* 17, no. 1 (2020): 154-161.
- [3] Mohammed. S. "State of the art of data mining and learning analytics tools in higher education." *International Journal of Emerging Technologies in Learning (iJET)* 15, no. 21 (2020): 58-76.
- [4] Malathi. M, K. Kaliannan, S. Balraj, K. Singaram, T. Kaliannan, and J. Renoald Albert. "A hybrid deep learning model for effective segmentation and classification of lung nodules from CT images." *Journal of Intelligent & Fuzzy Systems* 42, no. 3 (2022): 2667-2679.
- [5] Ei. C., E, H. H. Hein, Z. Lynn, S. J. Hla, and T. Tran. "Investigation on the influence of infill pattern and layer thickness on the mechanical strength of PLA material in 3D printing technology." *J. Eng. Sci. Res* 3, no. 2 (2019): 27-37.
- [6] Michael. A, A. Ghodsi, R. Xin, and M. Zaharia. "Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics." In *Proceedings of CIDR*, vol. 8. 2021.
- [7] Jennifer. A, and K. P. Sinaga. "Machine learning approaches for a marketing campaign in portuguese banks." In *2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS)*, pp. 1-6. IEEE, 2021.
- [8] Liang-Ting. T, H.L. Chi, T.H. Wu, and S.C. Kang. "AR-based automatic pipeline planning coordination for on-site mechanical, electrical, and plumbing system conflict resolution." *Automation in Construction* 141 (2022): 104400.
- [9] Xichen. C, A. Y. Chang-Richards, A. Pelosi, Y. Jia, X. Shen, M. K. Siddiqui, and N. Yang. "Implementation of technologies in the construction industry: a systematic review." *Engineering, Construction and Architectural Management* 29, no. 8 (2022): 3181-3209.
- [10] Florian. K, and S. Thalmann. "Applications of Artificial Intelligence in commercial banks—A research agenda for behavioral finance." *Journal of Behavioral and experimental finance* 27 (2020): 100352.

- [11] Mojtaba. N, P. Nayyeri, H. Jabani, S. Shahab, and A. Mosavi. "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis." IEEE Access 8 (2020): 150199-150212.

Data Mining and Visualisation Report

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

Instructor

70/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

GRADING FORM: DATA MINING / VIS A2_2022/3

GOWRI PAMARTHI

70

KNOWLEDGE AND UNDERSTANDING

Dataset description (15 marks)



Marketing on banking dataset taken fro UCI repository used for data analysis in submission. Reasons given for undertaking analysis needed to be more explicit. Explanation of listed attributes are clear and coherent.

12

COGNITIVE/ INTELLECTUAL SKILLS

Primary and Secondary data analysis, reflection (35 marks)



Stages of analysis and relevance of each included algorithm is shown in your submission. Fair element of self reflection is also shown concerning your evaluation of included dataset.

30

PRACTICAL APPLICATION SKILLS

Primary and Secondary data processing (20 marks)



Pre-processing undertaken described sufficiently well.

15

TRANSFERABLE SKILLS

Data reporting and report writing (30 marks)



General report structure and content meets normal report writing requirements. However, there are grammatical errors evident in your submission. This does affect the conveyance of report's ideas.
In terms of data reporting this is consistently depicted and information is easily ascertained. Diagrams are connected to literary content.

13