**Ishaan Chawla, Student Id: U01748687, Username: Ic57230n**

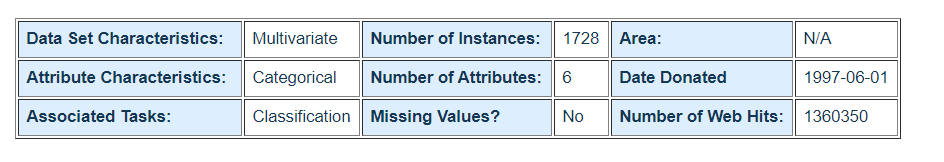
**Email:** [**ic57230n@pace.edu**](mailto:ic57230n@pace.edu)

**Data Mining Final Project Report**

**Abstract:** The aim of the project is to carry out analysis on bank marketing data set to predict that if the client has subscribed a term deposit or not. Weka tool will be used to carry out the data analysis using different data mining algorithms as follows: Classification using Decision Trees, SVM (Support Vector Machines), Nearest Neighbor classification & Clustering. These techniques will help to determine variety of characteristics exist.

**Description of dataset:**

**Link:** [**http://archive.ics.uci.edu/ml/datasets/Car+Evaluation**](http://archive.ics.uci.edu/ml/datasets/Car+Evaluation)

****

**Attribute Information:**

Input variables:  
# bank client data:  
1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular','telephone')  
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# other attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  
# social and economic context attributes  
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17 - cons.price.idx: consumer price index - monthly indicator (numeric)  
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)  
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)  
20 - nr.employed: number of employees - quarterly indicator (numeric)  
  
Output variable (desired target):  
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

**Description of the problem that is intended to solve:** The aim of the project is to carry out analysis on bank marketing data set to predict that if the client has subscribed a term deposit or not.

**Type of data mining intended to perform:** Classification using Decision Trees, SVM(Support Vector Machines), Nearest Neighbor classification & Clustering are the data mining techniques which is intended to perform on this dataset.

**Explanation of this choice:** The reason intended to perform these data mining techniques is because the goal is to predict if the client will subscribe a term deposit or not for which the classification would be the best fit. As, the output goal is to classify the result if the client has subscribed the term deposit or not.

**Description of any transformation in dataset:**  The current format of the data set is not viable for data mining since the values are semicolon (“;”) separated not comma separated. So, a transformation of this dataset must be required in comma separated values. For that I will be using python panda’s library to convert this dataset in proper table format.

**Original Dataset Link:** [**https://raw.githubusercontent.com/IshaanChawla0001/DataMining-final-project/main/bank-additional-full.csv**](https://raw.githubusercontent.com/IshaanChawla0001/DataMining-final-project/main/bank-additional-full.csv)

**Final Converted Table Format Dataset(.CSV) Link:** [**https://raw.githubusercontent.com/IshaanChawla0001/DataMining-final-project/main/final\_bank\_dataset.csv**](https://raw.githubusercontent.com/IshaanChawla0001/DataMining-final-project/main/final_bank_dataset.csv)

**Python program to convert original dataset to table format:** [**https://github.com/IshaanChawla0001/DataMining-final-project/blob/main/convert.py**](https://github.com/IshaanChawla0001/DataMining-final-project/blob/main/convert.py)

Once the data is converted into table format, I need to drop the column “contact” since all the values in “contact” are same and it does not affect the final output. Except “contact” attribute, every other attribute in the dataset has correlation with the final outcome.

**Dataset Preparation:** As Weka uses **.Arff** format, so I have to change the dataset from **.Csv** to **.Arff** format.

* Under Weka landing screen, go to tools -> Arff Viewer.
* Then load **.csv** file into it and save as **.arff** file. Now, the data is ready to mine and modelled.

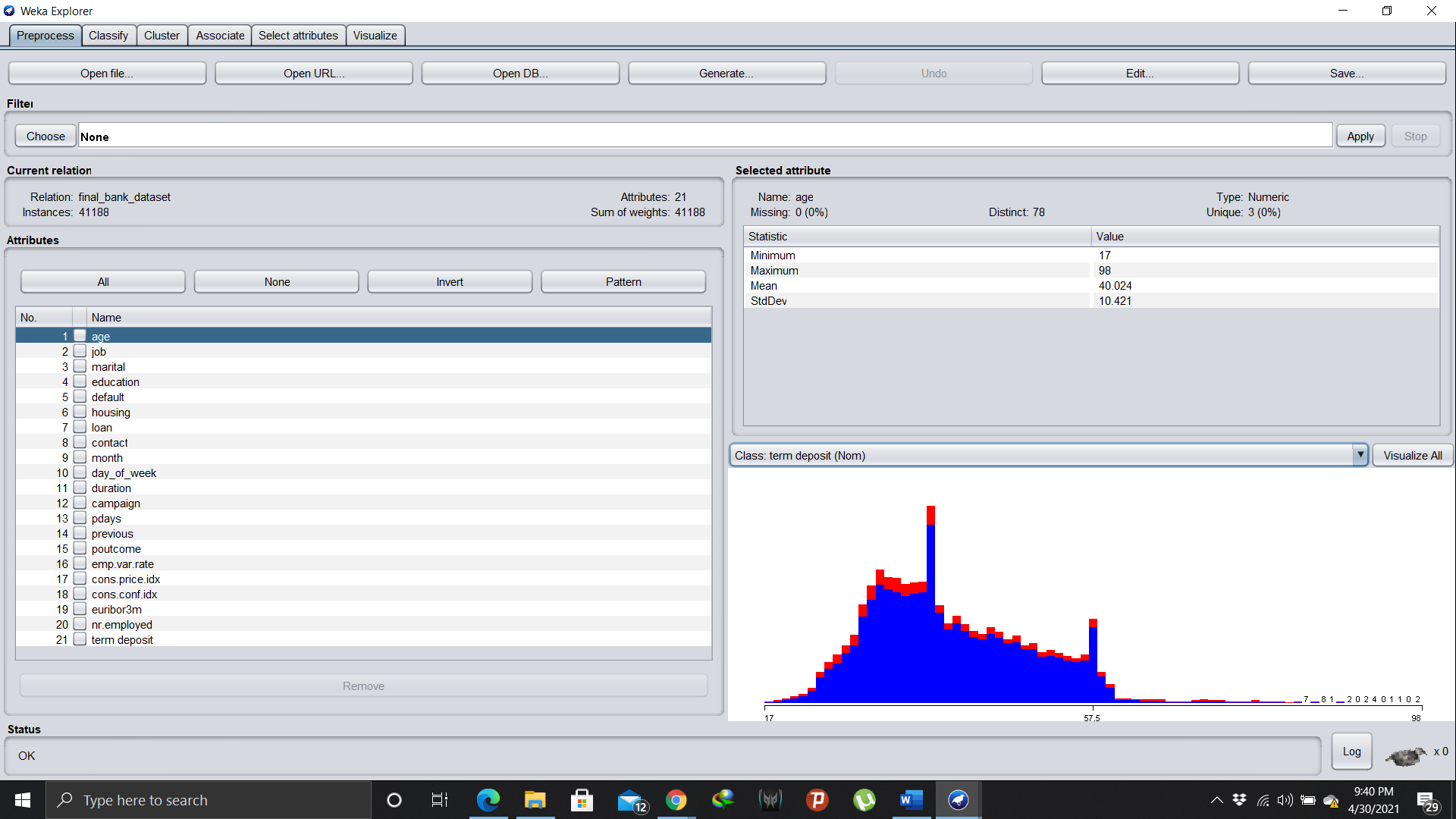
**Dataset Analysis:** For Data Analysis to be performed in **Weka.** I choose 4 Data mining algorithms to build and predict the best model.

1. Classification using Trees **DecisionStump & J48**
2. SVM (Support Vector Machines) with **Function SMO**
3. Nearest Neighbor classification with **IBK**
4. Clustering

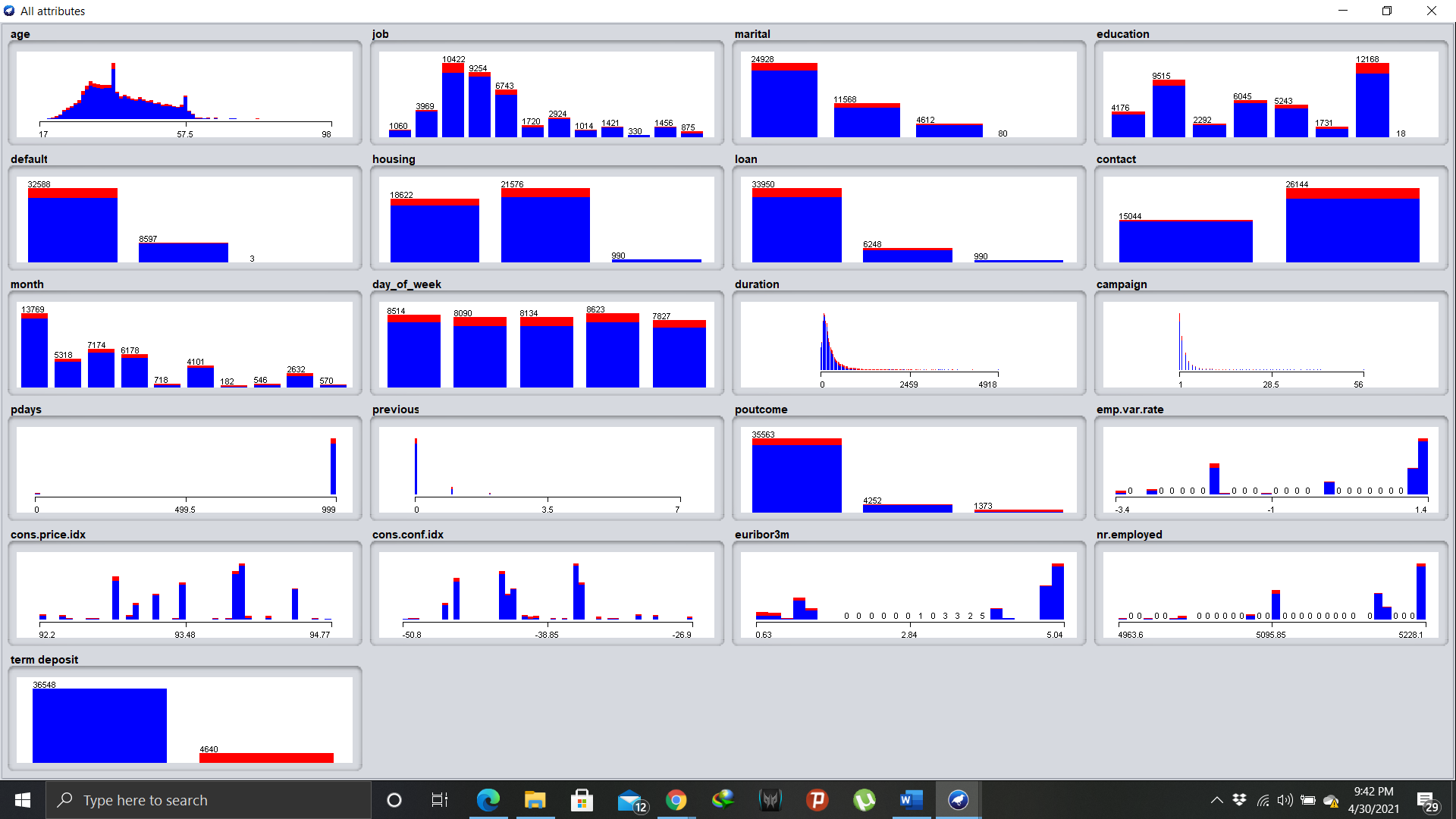
For this the main step is to load the dataset in Weka. The steps are:

1. Open Weka.
2. Click Explorer.
3. Click Open File.
4. Load Dataset as .arff file.

Once loaded in Weka. This is the visual representation.



To visualize Coorelation of attributes Click on **Visualize all.**



1. Classification using Trees **DecisionStump & J48**

**DecisionStump:** It is a one level decision tree with one internal node. The internal nodes are connected to terminal node.

It is mainly called **1-rule algorithm**. It can be used at a very basic level to develop a basic prediction model.

**Decision Stump with 90% Training Dataset:**

Correctly Classified Instances 3636 **88.2739 %**

Incorrectly Classified Instances 483 **11.7261 %**

Mean absolute error **0.1722**

Root mean squared error **0.2925**

**Note:** To check the accuracy of **DecisionStumping** I used another decision tree algorithm. It is one of the best machine learning algorithm for categorical prediction of data.

**J48:**

Correctly Classified Instances 3739 **90.7745 %**

Incorrectly Classified Instances 380 **9.2255 %**

Mean absolute error **0.1171**

Root mean squared error **0.2659**

By using J48 , the success rate has **increased by 2.3%** and mean absolute error & root mean squared error **decreased by 0.621 & 0.0321**.

**Confusion matrix of DecisonStumping & J48:**

|  |  |
| --- | --- |
| a b <-- classified as  3636 0 | a = no  483 0 | b = yes | a b <-- classified as  3487 149 | a = no  231 252 | b = yes |

1. **SVM (Support Vector Machines) with Function SMO:** Sequential minimal optimization also know for quadratic problem solving algorithm is mainly used for training **support vector machines.**

It works by breaking the problem into smaller sub-problems which is solved analytically.

For this dataset SVM would be perfect as that dataset is a combination of numerical and nominal values.

**SMO with 90% Training Dataset:**

Correctly Classified Instances 468 **96.8944 %**

Incorrectly Classified Instances 15 **3.1056 %**

Mean absolute error **0.0311**

Root mean squared error **0.1762**

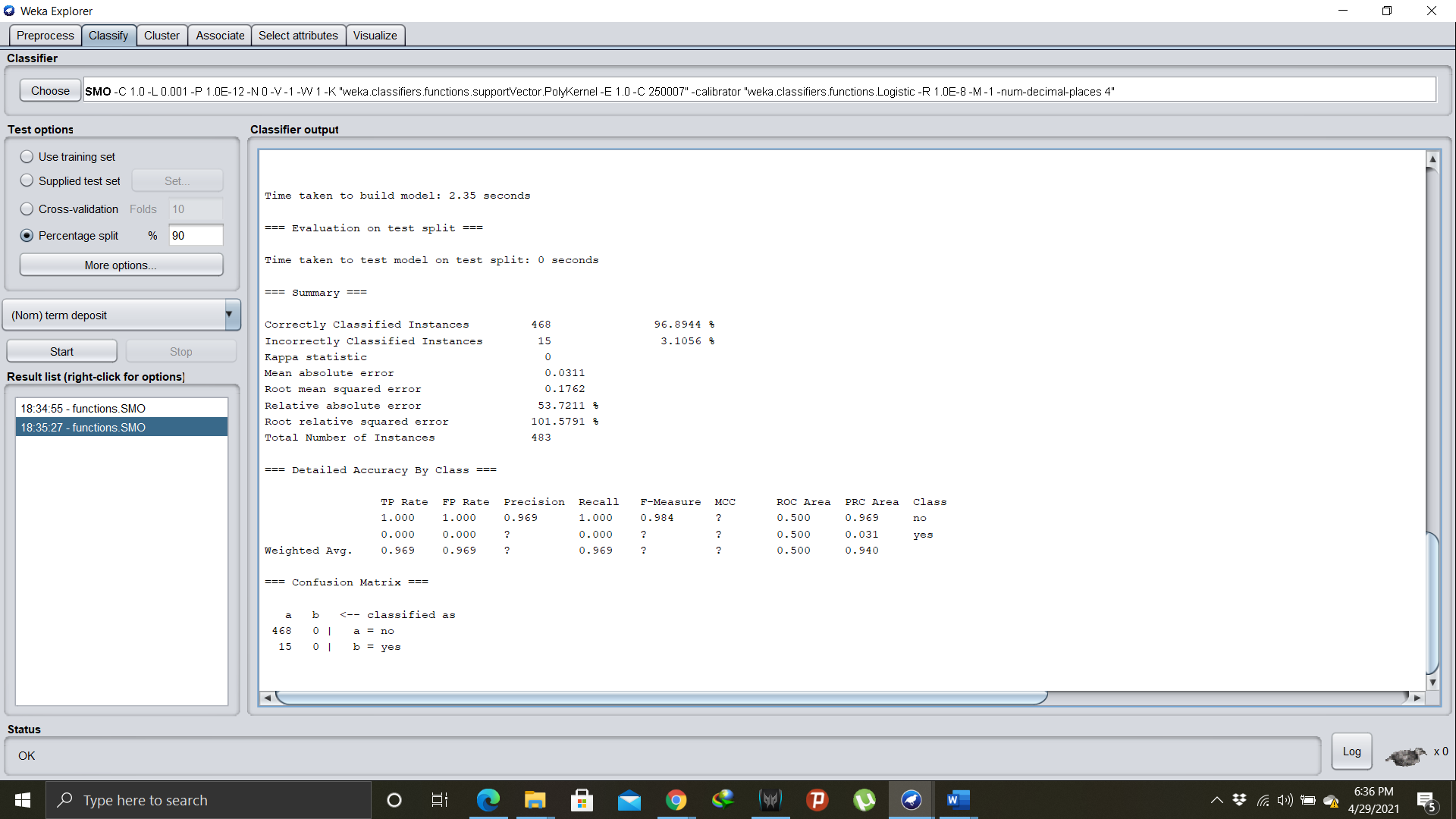
**SVM with SMO has more accuracy with less mean absolute error and root mean squared error.**

**Confusion Matrix:**

a b <-- classified as

12102 259 | a = no

1161 482 | b = yes



1. **Nearest Neighbor classification with IBK:** This **supervised learning** algorithm uses a distance to measure to locate nearest **“K”** instances with trained data for every instance being tested while prediction.

This algorithm does not use a build prediction model for testing.

**IBK with 90% Training Data:**

Correctly Classified Instances 3603 **87.4727 %**

Incorrectly Classified Instances 516 **12.5273 %**

Mean absolute error 0.1253

Root mean squared error 0.3539

**Confusion matrix:**

a b <-- classified as

3428 208 | a = no

308 175 | b = yes

**With IBK the success rate is same as DecisionStumping but the root mean squared error is more.**

1. **Clustering:** This **Unsupervised Learning** algorithm is used to statistical data analysis where it involves a group of datapoints, then each datapoint is classified into a specific group.

**EM Clustering:** It is an iteration algorithm to find maximum estimate of parameters in statistical models numerically.

**EM Clustering:**

Number of clusters selected by cross validation: 6

Number of iterations performed: 0

Cluster

Attribute 0 1 2 3 4 5

(0.1) (0.2) (0.21) (0.15) (0.15) (0.19)

====================================================================================

age

mean 41.8996 40.0639 39.8631 35.4203 40.7981 45.6123

std. dev. 9.0902 8.4165 8.6278 7.7695 8.1393 8.2562

job

housemaid 20 34 16 20 26 39

services 227 162 38 57 73 23

admin. 38 457 44 344 75 28

blue-collar 43 50 651 53 84 555

technician 42 86 83 76 311 54

retired 18 26 22 5 24 52

management 38 71 53 81 64 63

unemployed 10 19 21 19 18 30

self-employed 17 27 27 45 11 23

unknown 12 7 14 6 7 25

entrepreneur 16 39 38 23 34 31

student 9 5 6 30 3 2

[total] 490 983 1013 759 730 925

marital

married 347 706 755 145 569 776

single 72 137 142 515 60 64

divorced 62 128 106 88 91 73

unknown 1 4 2 3 2 4

[total] 482 975 1005 751 722 917

education

basic.4y 18 42 120 23 33 419

high.school 265 479 25 111 237 22

basic.6y 39 37 121 24 48 106

basic.9y 52 65 491 66 63 145

professional.course 48 87 93 64 205 65

unknown 25 46 46 18 34 82

university.degree 38 222 112 448 105 81

[total] 485 978 1008 754 725 920

default

no 87 832 817 656 696 183

unknown 393 141 186 93 24 732

[total] 480 973 1003 749 720 915

housing

no 314 160 237 558 615 614

yes 151 787 739 167 85 258

unknown 16 27 28 25 21 44

[total] 481 974 1004 750 721 916

loan

no 391 780 817 633 616 748

yes 74 167 159 92 84 124

unknown 16 27 28 25 21 44

[total] 481 974 1004 750 721 916

contact

telephone 479 972 1002 748 719 914

[total] 479 972 1002 748 719 914

month

may 479 972 1002 748 719 914

[total] 479 972 1002 748 719 914

day\_of\_week

mon 73 149 196 163 189 178

tue 20 498 48 128 98 430

wed 243 108 196 92 289 87

thu 79 87 423 120 75 105

fri 68 134 143 249 72 118

[total] 483 976 1006 752 723 918

duration

mean 276.1715 262.4614 271.6773 274.9705 264.4889 270.6627

std. dev. 275.6165 248.5258 247.2402 261.3086 241.6786 253.8844

campaign

mean 2.4331 2.1751 2.4825 2.4311 2.2214 2.3209

std. dev. 2.1984 1.9536 2.8157 2.9945 2.2954 2.4817

pdays

mean 999 999 999 999 999 999

std. dev. 0 0 0 0 0 0

previous

mean 0 0 0 0 0 0

std. dev. 0 0 0 0 0 0

poutcome

nonexistent 479 972 1002 748 719 914

[total] 479 972 1002 748 719 914

emp.var.rate

mean 1.1 1.1 1.1 1.1 1.1 1.1

std. dev. 0 0 0 0 0 0

cons.price.idx

mean 93.994 93.994 93.994 93.994 93.994 93.994

std. dev. 0 0 0 0 0 0

cons.conf.idx

mean -36.4 -36.4 -36.4 -36.4 -36.4 -36.4

std. dev. 0 0 0 0 0 0

euribor3m

mean 4.8582 4.8564 4.8578 4.857 4.8572 4.8566

std. dev. 0.0014 0.0011 0.0019 0.0017 0.0014 0.0012

nr.employed

mean 5191 5191 5191 5191 5191 5191

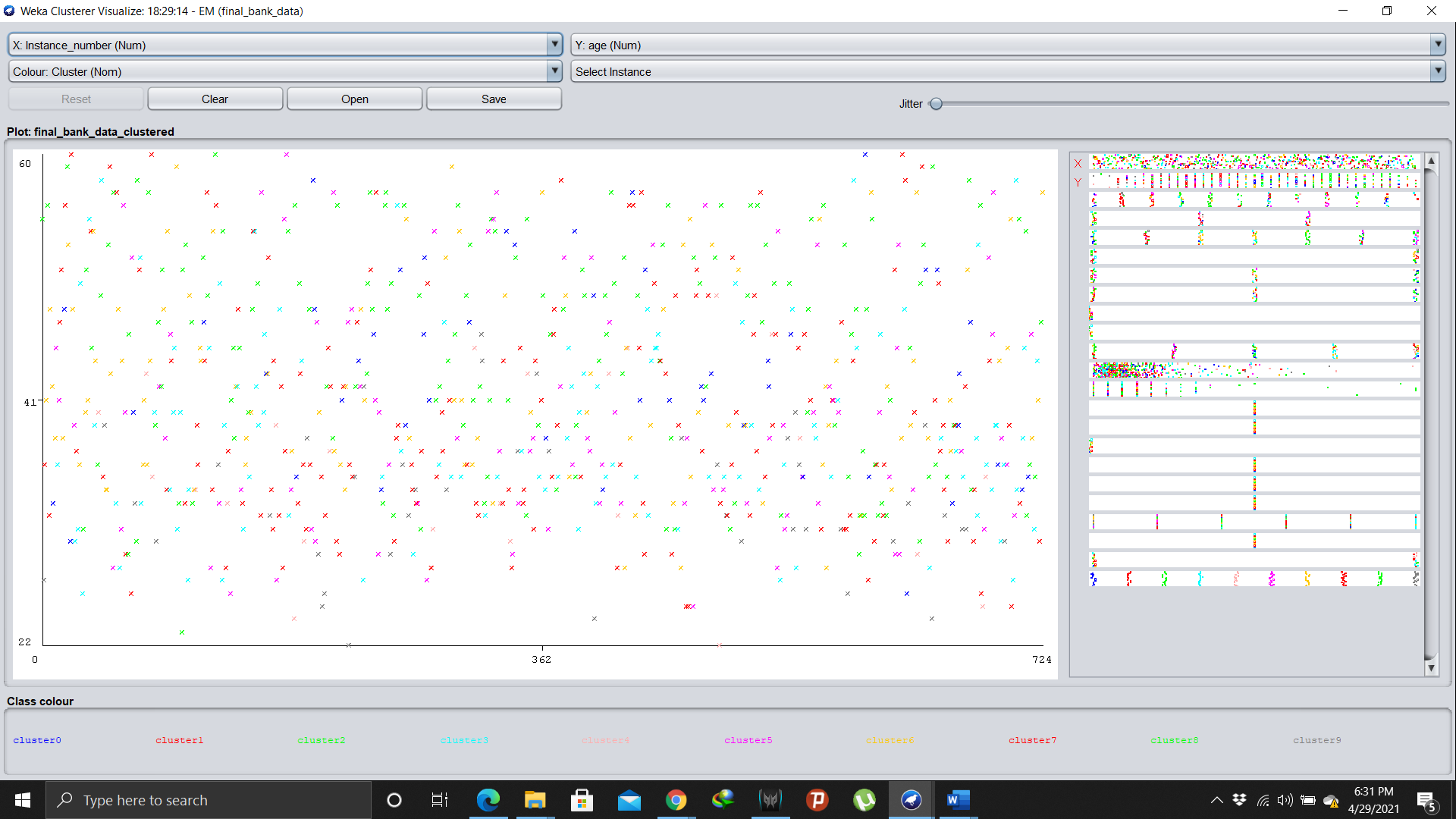
std. dev. 0 0 0 0 0 0

term deposit

no 462 946 970 719 705 894

yes 18 27 33 30 15 21

[total] 480 973 1003 749 720 915



**Clustered Instances:**

0 400 ( 8%)

1 1079 ( 22%)

2 979 ( 20%)

3 730 ( 15%)

4 668 ( 14%)

5 972 ( 20%)

Log likelihood: 61.5488

Class attribute: term deposit

**Classes to Clusters:**

0 1 2 3 4 5 <-- assigned to cluster

378 1053 953 700 658 948 | no

22 26 26 30 10 24 | yes

Correctly Classified Instances 468 **82.8944 %**

Incorrectly Classified Instances 15 **17.2156 %**

**Result:**

**Conclusion:** The different algorithms model developed to classify and predict the outcome on same dataset has different success & error rates. As we can see the **SMO** has the **highest success rate** and the **lowest mean squared error** on this dataset, which simply concludes that **Support vector machines** algorithms works perfectly on finance related datasets to produce optimal outcome. But after SMO, J48 can be a better option as both **decisionstumping** and **j48** performed better in this dataset.

**Appendix:**

[**https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68**](https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68)

[**https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization\_algorithm#:~:text=In%20statistics%2C%20an%20expectation%E2%80%93maximization,depends%20on%20unobserved%20latent%20variables**](https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm#:~:text=In%20statistics%2C%20an%20expectation%E2%80%93maximization,depends%20on%20unobserved%20latent%20variables)**.**

[**https://www.analyticsvidhya.com/blog/2020/03/decision-tree-weka-no-coding/**](https://www.analyticsvidhya.com/blog/2020/03/decision-tree-weka-no-coding/)