American International University-Bangladesh (AIUB)

Mid Term
Assignment

NAME : DOHAN, DIN MOHAMMAD

ID : 17-34465-2

COURSE NAME : DATA WAREHOUSING AND DATA MINING

SECTION : D

COURSE TEACHER : DR. MD MAHBUB CHOWDHURY

SEMESTER : SPRING 2020-2021

Dataset: COVID-19 Country Vaccination Progress.

Abstract: The purpose of this experiment is to know vaccination progress for COVID-19 diseases and which vaccine is better how much we get accurate results from specific vaccine as well as.

Tools:

I. Microsoft Excel
II. WEKA 3.9.5

Theory and Methodology: In this experiment, WEKA is being used for data analysis purposes. WEKA is an open source software that provides tools for data processing, implementation of several machine learning algorithms and visualization tools.

Dataset Selection: To select the desired dataset , many website has been searched but finally collected from *kaggle* which is the world's largest data science community with powerful tools.So finally prepared dataset is " COVID-19 Country Vaccination Progress".

The reason behind choosing this dataset is toward a pandemic situation. This is the great challenging situation for the world. Many people in the world died from COVID-19 and affected highly day by day but didn't get well so easily because many people still don't know how to recover themself. Though the world got COVID-19 vaccine but still don't know which vaccine is more secure and more useful.That's why this dataset is used for experiment and know the final feedback.

Dataset Attribute: There are 12 attributes and 5689 instances in dataset.

I. country
II. date
III. total_vaccinations
IV. people_vaccinated
V. people_fully_vaccinated
VI. daily_vaccinations_raw
VII. daily_vaccinaions
VIII. total_vaccinations_per_hundred
IX. people_vaccinated_per_hundred
X. people_fully_vaccinated_per_hundred
XI. daily_vaccinations_per_million
XII. vaccines

Here "*vaccines*" is a *class attribute* in the dataset.

Dataset Cleaning: After inserting the dataset there were lots of missing values data so had to clean first with ReplaceMissingValues and got some values those are considered as mean value for all attribute.

Relation: country_vaccinations

No.	1: country	2: date	3: total_vaccinations	4: people_vaccinated	5: people_fully_vaccinated	6: daily_vaccinations_raw	7: daily_vaccinations	8: total_vaccinations_per_hundre
	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	Albania	1/10...						0.0
2	Albania	1/11...					64.0	
3	Albania	1/12...	128.0	128.0			64.0	0.0
4	Albania	1/13...	188.0	188.0		60.0	63.0	0.01
5	Albania	1/14...	266.0	266.0		78.0	66.0	0.01
6	Albania	1/15...	308.0	308.0		42.0	62.0	0.01
7	Albania	1/16...	369.0	369.0		61.0	62.0	0.01
8	Albania	1/17...	405.0	405.0		36.0	58.0	0.01
9	Albania	1/18...	447.0	447.0		42.0	55.0	0.02
10	Albania	1/19...	483.0	483.0		36.0	51.0	0.02
11	Albania	1/20...	519.0	519.0		36.0	47.0	0.02
12	Albania	1/21...	549.0	549.0		30.0	40.0	0.02
13	Albania	1/22...					34.0	
14	Albania	1/23...					26.0	
15	Albania	1/24...					21.0	
16	Albania	1/25...					15.0	
17	Albania	1/26...					9.0	

Figure-1: Uncleaned Dataset

Relation: country_vaccinations-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.ReplaceWithMissingValue-Rfirst-last-S1-P0.1-weka.filt..

No.	1: country Nominal	2: date Nominal	3: total_vaccinations Numeric	4: people_vaccinated Numeric	5: people_fully_vaccinated Numeric	6: daily_vaccinations_raw Numeric	7: daily_vaccinations Numeric	8: total_vaccinations_per_hundred Numeric
1	Albania	1/10...	0.022036092817...	0.0279757195169...	0.019923969938355122	0.0281405374161456...	0.026771644079...	0.0
2	Albania	1/11...	0.022036092817...	0.0279757195169...	0.019923969938355122	0.02799221665731293	2.903252564539...	0.06294604972784525
3	Canada	1/12...	1.366169822721...	2.0953197971200...	0.019923969938355122	0.02799221665731293	2.903252564539...	0.0
4	Albania	1/13...	2.006561927122...	3.0775009520200...	0.019923969938355122	2.0659527881582342...	2.857169190499...	7.61556621734826E-5
5	Albania	1/14...	2.839071662843...	4.3543364533900...	0.019923969938355122	2.6857386246057043...	2.995419312620...	7.61556621734826E-5
6	Albania	1/15...	3.287346135924...	5.0418632618200...	0.019923969938355122	1.4461669517107639...	2.811085816459...	7.61556621734826E-5
7	Albania	1/16...	3.938411442065...	6.0404141026350...	0.019923969938355122	2.100385334627538E-5	2.811085816459...	7.61556621734826E-5
8	Albania	1/17...	4.322646704705...	0.0282970097733...	0.019923969938355122	1.2395716728949405...	2.626752320297...	7.61556621734826E-5
9	Albania	1/18...	4.770921177786...	7.3172496040050...	0.019923969938355122	1.4461669517107639...	0.027277637312...	1.523113243469652E-4
10	Canada	1/19...	5.155156440426...	7.9065582969450...	0.019923969938355122	1.2395716728949405...	2.304168702015...	1.523113243469652E-4
11	Albania	1/20...	5.539391703067...	8.4958669898850...	0.02026793336242598	1.2395716728949405...	2.119835205854...	1.523113243469652E-4
12	Albania	2/28...	5.859587755267...	8.9869575673350...	0.019923969938355122	1.0329763940791171...	1.797251587572...	1.523113243469652E-4
13	Albania	2/28...	0.022036092817...	0.0279757195169...	0.019923969938355122	0.02799221665731293	1.520751343330...	0.06294604972784525
14	Albania	1/23...	0.022036092817...	0.0279757195169...	0.019923969938355122	0.02799221665731293	1.152084351007...	0.06294604972784525
15	Albania	1/24...	0.022036092817...	0.0279757195169...	0.019923969938355122	0.0281405374161456...	9.216674808062...	0.06294604972784525
16	Albania	1/25...	0.022036092817...	0.0279757195169...	0.019923969938355122	0.02799221665731293	0.027277637312...	0.06294604972784525
17	Albania	1/26...	0.022036092817...	0.0279757195169...	0.019923969938355122	0.02799221665731293	3.686669923225...	0.06294604972784525
18	Albania	1/27...	0.022036092817...	0.0279757195169...	0.019923969938355122	0.02799221665731293	1.382501221209...	0.06294604972784525

Figure-2: Cleaned Dataset

Interface of Initial Values: After cleaning the dataset and converting it to csv file to arff file, dataset has been prepared to analyze and got initial information of the dataset.

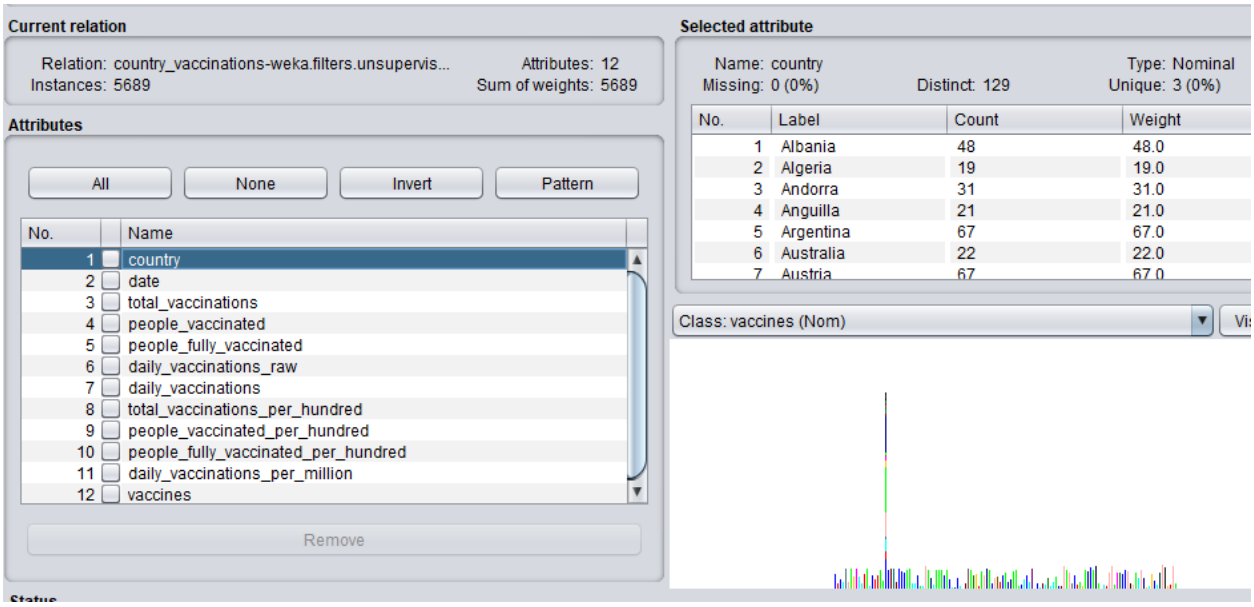

Figure-3: Basic Information of Dataset

Figure-3 shows that the total instances are there 5689, total attributes 12, sum of weights 5689 on the left section and the right section has shown the visual graph and summary of the selected attribute.

Vaccines with Graph: In this dataset , there are 24 vaccines that are used in the whole world. We can see there are no missing values because we have already applied ReplaceMissingValues on Processor window. This dataset contains all nominal values.This database shows that class attributes all weight and we can see Moderna/Oxford/AstraZeneca, Pfizer/BioNTech are top most weight than other vaccine

weight which is 1486.The computation of the weight of an identity attribute is fundamentals to a data machine technique often offered to as “probabilistic” machine. If the classification is correct the associated weight is increased.Since Moderna/Oxford/AstraZeneca, Pfizer/BioNTech produces highest weights that means these classification is more correct than other weights and most of the people used these vaccines.

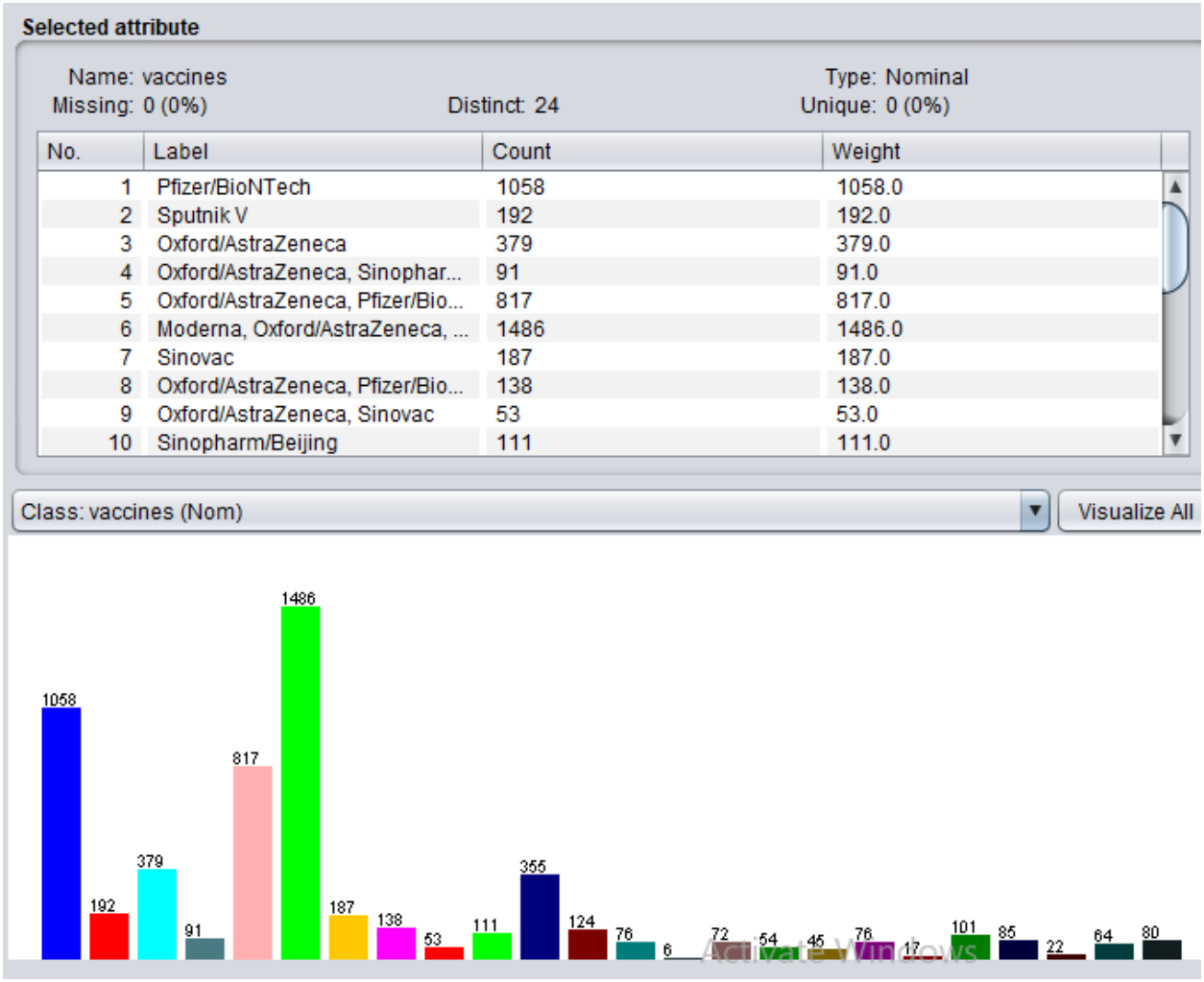

Figure-4: Vaccines with Graph

Data Normalization: Also referred to as data pre-processing which is a basic element of data mining. The main purpose of data normalization is to minimize or even exclude duplicated data.This is the process of recalling one or more attributes to the range of 0 to 1.This means that the largest value for each attribute is 1 and the smallest value is 0.

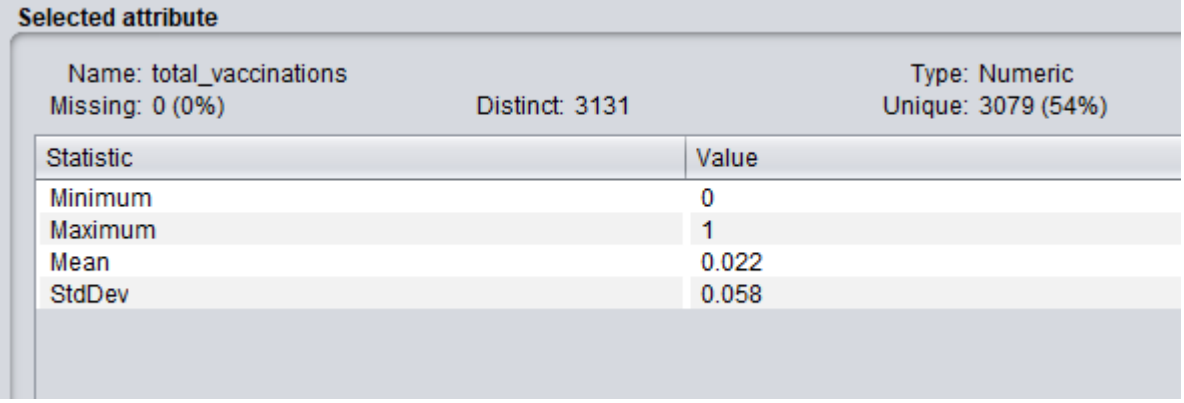

Figure-5: After normalization Min, Max, Mean, StdDev values

Algorithm:

K Nearest Neighbour (KNN): KNN algorithm is one of the simplest classification algorithms and is one of the most used learning algorithms. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

Naive Bayes: The Naive Bayes classifier is a simple classifier that classifies based on probabilities of events based on the prior knowledge of the conditions that might be related to the event. If we know the conditional probability, we can use the bayes rule to find out the reverse probabilities.

Decision Tree: Decision Tree is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.There are two types in decision tree classification technique CART and ID3 (Iterative Dichotomiser 3). CART uses the gini index as a classification matrix where ID3 uses information gain.

Algorithm Selection: To select algorithm on the dataset, classify window has been used in WEKA then choose Naive Bayes and start it, put it with cross validation fold 10.Then the result shown based on vaccines class attribute.

```
=== Summary ===

Correctly Classified Instances        2085               36.6497 %
Incorrectly Classified Instances      3604               63.3503 %
Kappa statistic                          0.3118
Mean absolute error                      0.0541
Root mean squared error                  0.2142
Relative absolute error                 75.2892 %
Root relative squared error            112.9919 %
Total Number of Instances             5689
```

Figure-6: Naive Bayes Accuracy Summary on cross validation

Then we would like to perform K Nearest Neighbors classification on the same dataset IBK classifier from lazy folder. Then the output is -

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4923               86.5354 %
Incorrectly Classified Instances       766               13.4646 %
Kappa statistic                          0.8413
Mean absolute error                      0.0268
Root mean squared error                  0.0988
Relative absolute error                 37.2399 %
Root relative squared error             52.1029 %
Total Number of Instances             5689
```

Figure-7: KNN Accuracy Summary on cross validation

Finally we apply ID3 algorithm but ID3 can only handle nominal values that's why we have applied J48 that handles nominal and numeric values in WEKA.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5248               92.2482 %
Incorrectly Classified Instances       441                7.7518 %
Kappa statistic                          0.9097
Mean absolute error                      0.0079
Root mean squared error                  0.0653
Relative absolute error                 10.9267 %
Root relative squared error             34.4543 %
Total Number of Instances             5689
```

Figure-8: ID3 produce High Accuracy on cross validation

Cross Validation: Cross validation is a standard evaluation technique. Divide a dataset into 10 folds then hold out each folds in turn on testing and train on the remaining 9 together. It gives 10 evaluation results on average, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.

Correctly Classified Instances: This is the sum of True Positive and True Negative. The total number of correctly instances divided by total number of instances gives the accuracy.From figure-8, we can see that coefficient is 92.25% accuracy which is excellent.

Kappa statistic:

$$\text{Kappa static} = \frac{observed\ accuracy - expected\ accuracy}{1 - expected\ accuracy}$$

Kappa statistic is a metric that compares an observed accuracy with an expected accuracy. It is used not only as a single classifier but also as an evaluate classifier. From figure-8 Kappa statistic is 0.9097 which is good.

Mean Absolute Error: Mean absolute error is a quantity used to measure how close forecasts or prediction are to eventual outcomes. The mean absolute error is as the name suggests, the mean absolute error is an average of the absolute errors. Here the MAE is 0.0079 which is very low and good.

Root Mean Squared Error: The root mean squared error is a frequently used measure of the differences between values predicted by a model. The RMSE represents the square root of the second sample moment of the difference between predicted values and observed values. RMSE is always no-negative, and a value 0 which is almost never achieved in practice would indicate a perfect fit to the data. But in general, a lower RMSE is better than a higher one. From the figure 8, it's shown that the RMSE value is 0.0653, which is comparatively lower.

Relative Absolute Error: Relative absolute error is a way to measure the performance of a predictive model. It's primarily used in machine learning, data mining. The RAE is expressed as a ratio, comparing a mean error to errors produced by a trivial or naïve model. A reasonable model will result in a ratio of less than one. From *figure 8,* we can see the ratio of RAE is 10.92% which indicates the created model is efficient and reasonable.

Root Relative Squared Error: Root relative squared error (RRSE) is relative to what it would have been if a simple predictor has been used. The predictor is just the average of the actual values. The RRSE takes the total squared error and normalizes it by dividing the total squared error of the simple predictor.From figure 8, we can see the value of RRSE is 34%.

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances        1052               92.4429 %
Incorrectly Classified Instances        86                7.5571 %
Kappa statistic                          0.9113
Mean absolute error                      0.008
Root mean squared error                  0.064
Relative absolute error                 11.1195 %
Root relative squared error             33.7466 %
Total Number of Instances             1138
```

Figure-9: ID3 Accuracy Summary on 80% test split

For the training and testing ratio,80/20 has been used to get better result.

From the above information it has decided that ID3 algorithm produces the best accuracy 92.44% and root relative squared error is 33.74% where Naive Bayes produces 36.65% with root relative squared error approximately 113% and KNN produces 86.54% and root relative squared error is 52.10%.

That’s why we use ID3 algorithm on this database to get accurate result.

Detailed Accuracy By Class:

Total Number of Instances 5689

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.954	0.021	0.914	0.954	0.933	0.918	0.993	0.980	Pfizer/BioNTech
	0.953	0.004	0.893	0.953	0.922	0.920	0.993	0.964	Sputnik V
	0.892	0.001	0.991	0.892	0.939	0.936	0.986	0.954	Oxford/AstraZeneca
	0.923	0.000	1.000	0.923	0.960	0.960	0.976	0.935	Oxford/AstraZeneca, Sinopharm/Beijing, Sputnik V
	0.923	0.010	0.940	0.923	0.931	0.920	0.992	0.974	Oxford/AstraZeneca, Pfizer/BioNTech
	0.951	0.037	0.902	0.951	0.926	0.899	0.991	0.981	Moderna, Oxford/AstraZeneca, Pfizer/BioNTech
	0.872	0.001	0.982	0.872	0.924	0.923	0.989	0.932	Sinovac
	0.906	0.000	0.992	0.906	0.947	0.947	0.987	0.947	Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing
	0.887	0.000	1.000	0.887	0.940	0.941	0.988	0.917	Oxford/AstraZeneca, Sinovac
	0.901	0.001	0.971	0.901	0.935	0.934	0.984	0.944	Sinopharm/Beijing
	0.825	0.019	0.746	0.825	0.783	0.769	0.970	0.872	Moderna, Pfizer/BioNTech
	0.927	0.001	0.935	0.927	0.931	0.930	0.990	0.960	Pfizer/BioNTech, Sinovac
	0.882	0.000	1.000	0.882	0.937	0.938	0.999	0.956	Sinopharm/Beijing, Sinopharm/Wuhan, Sinovac
	0.667	0.000	1.000	0.667	0.800	0.816	0.913	0.678	Moderna
	0.931	0.000	1.000	0.931	0.964	0.964	0.998	0.954	Moderna, Oxford/AstraZeneca, Pfizer/BioNTech, Sinophar
	0.870	0.000	0.979	0.870	0.922	0.922	0.980	0.914	Covaxin, Oxford/AstraZeneca
	0.889	0.000	1.000	0.889	0.941	0.942	0.988	0.926	Pfizer/BioNTech, Sinopharm/Beijing
	0.882	0.001	0.957	0.882	0.918	0.918	0.965	0.919	Oxford/AstraZeneca, Pfizer/BioNTech, Sputnik V
	0.941	0.000	1.000	0.941	0.970	0.970	0.970	0.941	Sinopharm/Beijing, Sputnik V
	0.921	0.000	1.000	0.921	0.959	0.959	0.997	0.949	Oxford/AstraZeneca, Sinopharm/Beijing
	0.929	0.000	1.000	0.929	0.963	0.964	0.987	0.957	EpiVacCorona, Sputnik V
	0.864	0.000	1.000	0.864	0.927	0.929	0.977	0.927	Johnson&Johnson
	0.844	0.000	1.000	0.844	0.915	0.918	0.975	0.917	Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing
	0.888	0.000	1.000	0.888	0.940	0.941	0.997	0.934	Johnson&Johnson, Moderna, Pfizer/BioNTech
Weighted Avg.	0.922	0.016	0.926	0.922	0.923	0.911	0.989	0.960	

Figure-10: Detailed Accuracy By Class

TP Rate: True Positive rate. Also known as “Sensitivity” or “Recall”.

$$\text{TP Rate} = \frac{TP}{TP+FN}$$

Out of all the positive classes, how much we predicted correctly. It should be as high as possible.Here Pfizer/BioNtech have a high TP rate which is 0.94

FP Rate: False Positive case. We predicted yes but they don't have the actual vaccinated. Also known as a “Type I Error”.Higher value indicate more False information.Here Moderna/Oxford/AstraZeneca,/BioNTech don't have actual vaccinated which FP Rate is 0.037

Precision:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Out of all the positive classes we have predicted correctly, how many are actually positive. Here

Pfizer/Moderna/Oxford/AstraZeneca,/BioNTech/Wuhan/Sinovac/Sinopharm/Johnson have high precision which is 1.00

Recall: Same as TP Rate.

Recall = $\frac{TP}{TP+FN}$

Out of all the positive classes, how much we predicted correctly. It should be as high as possible.Here Pfizer/BioNtech have a high Recall which is 0.94.

F-Measure:

F-Measure = $\frac{2*Recall*Precision}{Recall+Precision}$

It is difficult to compare to model with low precision and high recall.So, to make them comparable, we use F-Score. F-Score helps to measure recall and precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.Here Sinopharm/Beijing/Sputnik V has high F-Measure value which is 0.970

MCC: MCC is used in machine learning as a measure of the quality of binary classifications also measure which can be used even if the classes are of very different sizes.Here Sinopharm/Beijing/Sputnik V has high MCC value which is 0.970

ROC Area: One of the most important values output by WEKA. They give an idea of how the classifiers are performing in general. Here, Moderna/Oxford/AstraZeneca, Pfizer/BioNTech have better performance which value is 0.999

PRC Area: The Precision Recall plot is more informative than the ROC plot when evaluating Binary classifiers on imbalanced datasets. Here Moderna/Oxford/AstraZeneca, Pfizer/BioNTech have high PEC values which is 0.981

Confusion Matrix: Also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. Each row of the matrix represents the instances in a predicted class where each column represents the instances in an actual class.It is extremely useful for measuring Precision, Recall, Specificity, Accuracy and most importantly AUC-ROC Curve.
From the following matrix we can see that Moderna/Oxford/AstraZeneca has True Positive result high (1413) which much more give accurate results than others.

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j    k    l    m    n    o    p    q    r    s    t    u    v    w    x   <-- classified as
 1009    5    0    0    8   25    1    1    0    1    7    1    0    0    0    0    0    0    0    0    0    0    0    0 |    a = Pfizer/BioNTech
    6  183    0    0    0    2    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0    0 |    b = Sputnik V
   22    1  338    0    1   11    1    0    0    0    5    0    0    0    0    0    0    0    0    0    0    0    0    0 |    c = Oxford/AstraZeneca
    1    1    0   84    0    4    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0 |    d = Oxford/AstraZeneca,
   15    2    0    0  754   31    0    0    0    1   14    0    0    0    0    0    0    0    0    0    0    0    0    0 |    e = Oxford/AstraZeneca,
   25    1    0    0    9 1413    1    0    0    1   35    0    0    0    0    0    0    1    0    0    0    0    0    0 |    f = Moderna, Oxford/Ast
    4    1    0    0    1   10  163    0    0    0    8    0    0    0    0    0    0    0    0    0    0    0    0    0 |    g = Sinovac
    1    0    0    0    2    8    0  125    0    0    2    0    0    0    0    0    0    0    0    0    0    0    0    0 |    h = Oxford/AstraZeneca,
    0    0    0    0    0    1    0    0   47    0    5    0    0    0    0    0    0    0    0    0    0    0    0    0 |    i = Oxford/AstraZeneca,
    3    0    0    0    0    6    0    0    0  100    1    1    0    0    0    0    0    0    0    0    0    0    0    0 |    j = Sinopharm/Beijing
    8    9    3    0   15   25    0    0    0    0  293    2    0    0    0    0    0    0    0    0    0    0    0    0 |    k = Moderna, Pfizer/Bio
    0    1    0    0    2    3    0    0    0    0    1  115    0    0    0    0    0    2    0    0    0    0    0    0 |    l = Pfizer/BioNTech, Si
    0    0    0    0    0    8    0    0    0    0    1    0   67    0    0    0    0    0    0    0    0    0    0    0 |    m = Sinopharm/Beijing,
    0    0    0    0    0    2    0    0    0    0    0    0    0    4    0    0    0    0    0    0    0    0    0    0 |    n = Moderna
    0    0    0    0    1    1    0    0    0    0    3    0    0    0   67    0    0    0    0    0    0    0    0    0 |    o = Moderna, Oxford/Ast
    0    0    0    0    0    2    0    0    0    0    4    1    0    0    0   47    0    0    0    0    0    0    0    0 |    p = Covaxin, Oxford/Ast
    4    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0   40    0    0    0    0    0    0    0 |    q = Pfizer/BioNTech, Si
    0    0    0    0    1    4    0    0    0    0    3    1    0    0    0    0    0   67    0    0    0    0    0    0 |    r = Oxford/AstraZeneca,
    0    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0   16    0    0    0    0    0 |    s = Sinopharm/Beijing,
    2    0    0    0    1    2    0    0    0    0    3    0    0    0    0    0    0    0    0   93    0    0    0    0 |    t = Oxford/AstraZeneca,
    2    0    0    0    0    2    0    0    0    0    2    0    0    0    0    0    0    0    0    0   79    0    0    0 |    u = EpiVacCorona, Sputn
    0    1    0    0    0    0    0    0    0    0    0    2    0    0    0    0    0    0    0    0    0   19    0    0 |    v = Johnson&Johnson
    2    0    0    0    5    2    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0   54    0 |    w = Oxford/AstraZeneca,
    0    0    0    0    1    4    0    0    0    0    4    0    0    0    0    0    0    0    0    0    0    0    0   71 |    x = Johnson&Johnson, Mo
```

Figure-11: Confusion Matrix

Visualize Threshold Curve:

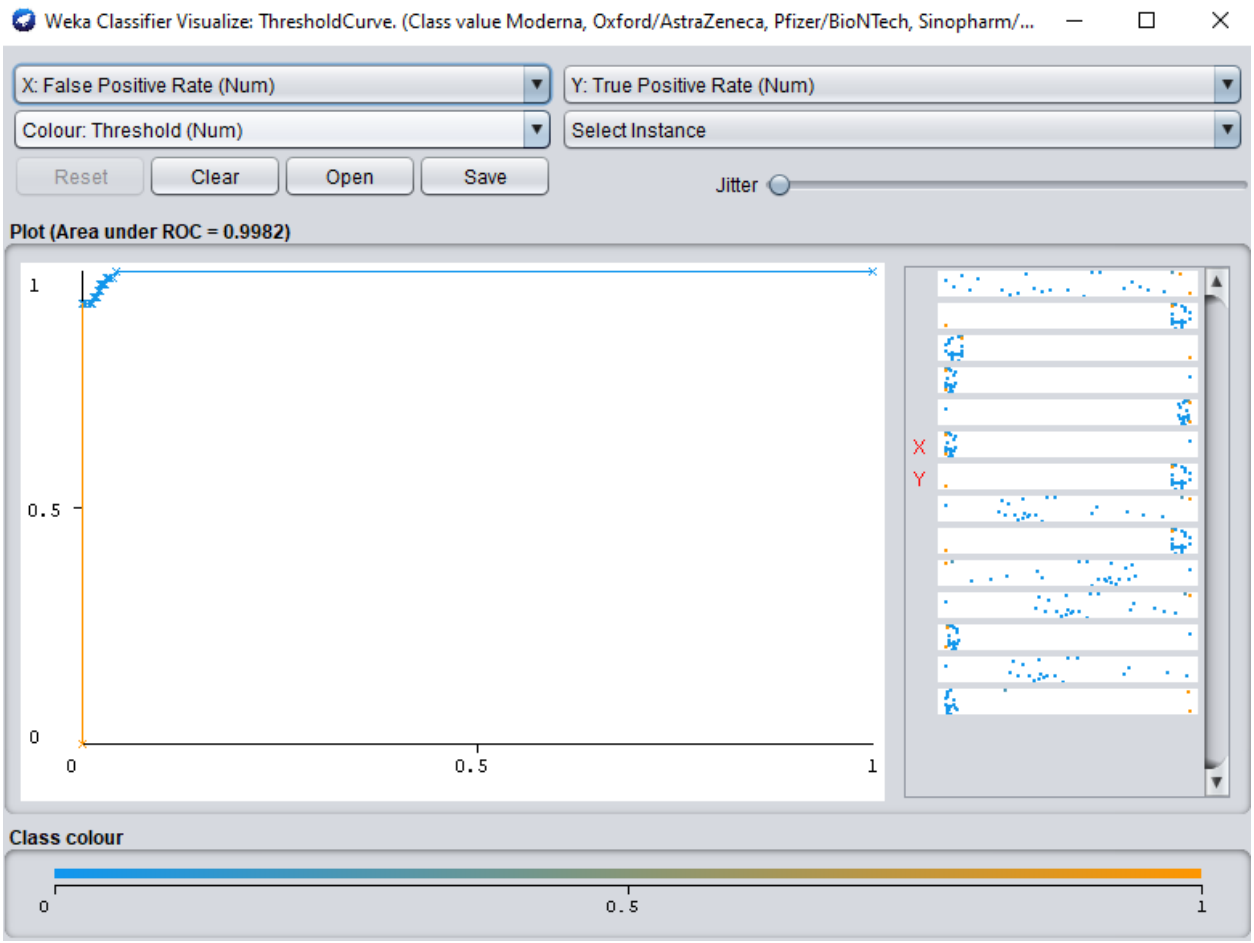

Figure-12: Threshold Curve

From the figure-12, we can see X axis represents false positive rate and Y axis represents true positive rate. ROC is 0.9982 after 0.9982 the threshold curve is increased to 1.18 and then the curve is stable to X axis.

Performance Test Validation: After running the dataset we can see the dataset has successfully tested performance and there were no errors found.

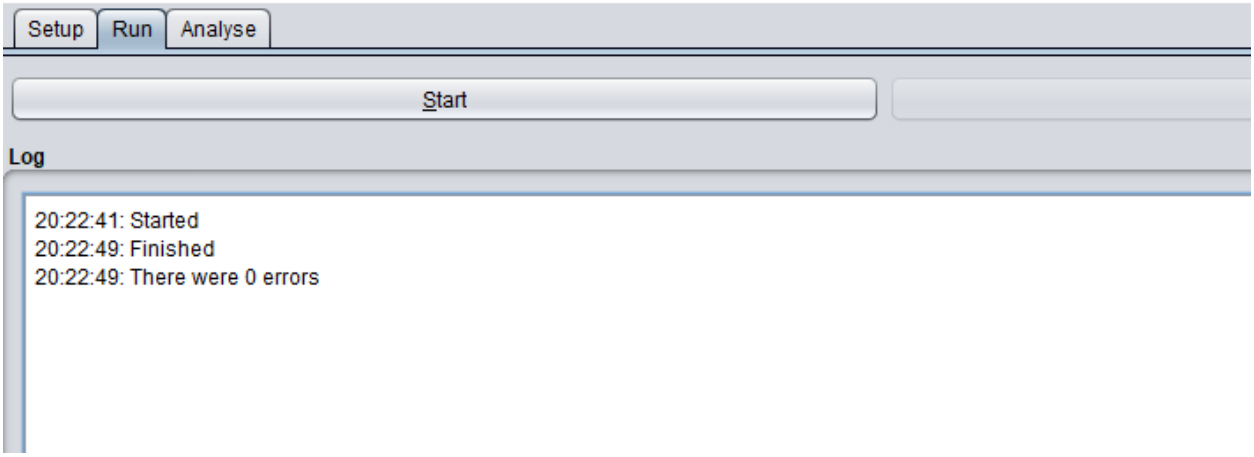

Figure-13: Performance Test Validation

Visualize Classifier Errors: This plot X axis shows people vaccinated and Y axis shows predicted vaccines. From this plot we visualize that Moderna/Oxford/AstraZeneca, Pfizer/BioNTech vaccines are highly used and give better accurate results than others.

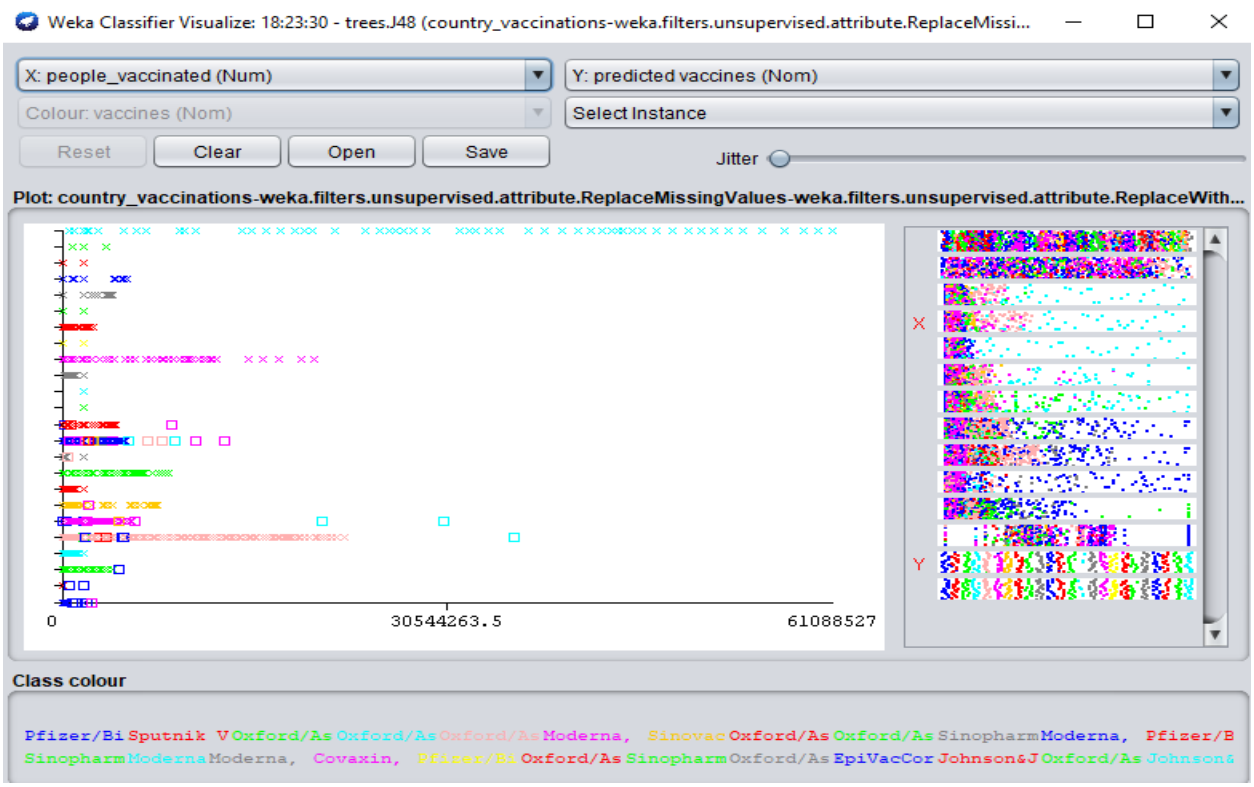

Figure-14: Visualize Classifier Errors

Conclusion: In this dataset I have used ID3 algorithm because this algorithm performs better and gives an accurate result which is 92.44%, compared to other algorithms gives better results and training test is also good for this dataset.More useful for real time prediction. It can overfit the training data and understandable prediction rules are created from the training data also it builds a short tree in relatively small time. This algorithm is test enough to attributes until data is classified.

References:

[I] https://www.kaggle.com/gpreda/covid-world-vaccination-progress