DMAT – Assignment 2

Name 1: Subhasree Vadukoot Student Number: 3014289
Name 2: Minna George Kaiprambadan Student Number: 3008351

Course	MSCBD-DMAT
Stage / Year	1
Module	Data Mining Algorithms & Techniques
Semester	2
Assignment	Assignment 2
Date of Title Issue	16th April
Assignment Deadline	6th May at 23:55
Assignment Submission	Upload to Moodle
Assignment Weighting	25% of module

Group Assignment

You will be working in groups of two to complete this assignment. Email abubakr.siddig@griffith.ie with the names of the people in your group and cc the other group member (Only one email by group). I suggest that you work through Google Drive with this document being stored as a Google Doc so that you can both work together.

Objective

1. To successfully apply a set of data mining skills imparted in this module to a previously unseen datasets to achieve knowledge discovery.
2. Evaluate a well-regarded peer reviewed paper or journal article which concerns the application of one of the techniques covered in this module and comment on its relevance to your dataset.

Deliverables

A single zip called firstName1_lastName1_studentNumber1_firstName2_lastName2_studentNumber2_assignment2.zip to be uploaded to Moodle containing the following files:

This file edited to contain the results of your investigation. Each of the **NUMBERED HEADINGS IN RED** should be expanded to satisfy the requirements of the section.

- A set of supporting files including but not limited to the following, which should be clearly referenced from your documentation. You only need to submit the files relevant the techniques you have explored. oThe original dataset file

    - dataset.arff
    - trainigSet.arff
    - testingSet.arff
    - j48tree.arff
    - associationrules.arff
    - kmeans.arff
    - dbscan.arffmlp.arff
    - The research paper.

Choosing Your Dataset

1. Your dataset should concern a real-world problem that lends itself to easy understanding by your classmates.
2. It should not be identical to the dataset you used in assignment1.
3. It should have >1000 tuples/rows/instances.
4. It should have >=10 attributes
5. It should have attributes which can serve as labels so that the accuracy of your data analysis can be determined.
6. If you cannot find one dataset which is suitable for use with all techniques then you may choose 2. Please clearly indicate which dataset was used in which case.

The list below should help you on your search, student please share additional sources on Moodle discussion form.

- **UCI Machine Learning Repository**- A repository of more than 200 data sets for machine learning and data mining
- **Movie Ratings Data**- Real movie ratings data from**www.movielens.org**Web site. Contains ratings on 1600+ movies by 1000 users
- **Kaggle.com Competition Data Sets**- Data sets from a variety of competitions.
- **Stanford Large Network Dataset Collection**- A variety of network data sets, including data from social networks, product reviews, online communities, etc.
- **Yelp Data Set Challenge**- Reviews and check-in data on thousands of businesses.
- **Million Song Dataset**- Freely-available collection of audio features and metadata for a million contemporary popular music tracks.
- **Public Data sets on Amazon Web Services**- Large public data sets (including data sets for US Census, Wikipedia, Freebase, human genome project), ready for big data analytics on the cloud.
- **Data.gov**-Publically available data sets from Federal, State, and local government, including economic, geological, demographic and many other types of data sources. This site also includes a list of other**Open Data Sites**with similar publicly available data sources from various cities, states, and countries.
- **KDnugget's list of data sets for data mining**
- **Infochimps Data Market**- Thousands of data sets, including data from various social networks and collaborative tagging sites such as Twitter, Delicious, Last.fm, MusicBrainz, as well as data sets from many other domains.

Initial Tasks

1. Description of your dataset(s) and findings – 20%

Two different datasets are used for the purpose of this assignment, as the first dataset- *Adult Censusdataset*, is not suitable for Time Series Forecasting. Hence the second dataset- *Madrid Weather dataset* is used for Time Series.

Dataset for Classification, Clustering – Adult Census Dataset

- **Title**: Data Mining using Classification and Clustering techniques on the Adult Census Dataset
- **Data description:**

1. The problem domain

The world we live in is controlled by the economies, which is extremely dependent on an individual's income. Cambridge defines income as "money that is earned from doing work or received from investments"(INCOME | meaning in the Cambridge English Dictionary, 2020). An individual's income is very much affected by his age, occupation and unfortunately factors like gender. The dataset used here is originally the US Census data collected in 1994. However, all of the attributes which are factors that directly or indirectly affect the income of people, is valid even today. A study conducted on 'Annual Survey of Hours and Earnings'of the year 2016 by the UK Government gives insights about the factors that can affect earnings, which is also useful for income. Age, gender, sector, skill group etc. were found to be relevant factors (UK Government, 2020).

Understanding what affects an individual's income is important now more than ever when the world is trying hard to mitigate the effects of the Novel Coronavirus on the economy. The dataset is chosen after understanding how important income of individuals and a stable economy is for any country, especially during the time of a crisis. Another interesting article on the wall street journal suggests that the pandemic might balance wealth distribution by decreasing income equality. (Hannon, 2020) Thus the dataset that provides vital information about income and characteristics is always a relevant problem in any society.

2. The source of the data

The dataset originally comes from the US Census Bureau (Bureau, 2020). The United States Census Bureau is a principal agency of the U.S. Federal Statistical System, which is responsible for producing data about the people and economy of America. (United States Census Bureau, 2020)The organization releases the Census data for the public to use. Their censuses and surveys help in informed decision making and strategy building in the United States. The dataset is available in UCI Machine Learning Repository. It is the data that was extracted by Barry Becker from the 1994 Census database. The donors of this dataset are Ronny Kohavi and Barry Becker of Silicon Graphics. The data used here is downloaded from Kaggle which is named as Adult income dataset and this data is made available in Kaggle from UCI repository.

Extraction was performed by Barry Becker using some condition such as((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)). This ensures that the dataset doesn't have lots of unclean or noisy records. (UCI Machine Learning Repository: Adult Data Set, 2020)

3. The agencies working with the data

The Adult Census Dataset is cited widely and is used in various researches worldwide. The dataset was originally used by the US Census Bureau. This dataset was made publicly available and after the extraction performed by Barry Becker, the dataset continues to be used by various researchers at several academic institutions or organizations worldwide in order to study classification algorithms or to gain valuable insights. Some of the notable agencies or researchers working with the data are:

3.1. "Privacy Preserving OLAP" – Paper presented on SIGMOD Conference. 2005.

This research paper was presented on the Proceedings of the 2005 ACM SIGMOD International conference on Management of data. The authors of the paper are members of IBM Almaden and Stanford University. The paper presents the concept of data perturbation to protect from data breaches. This paper cited the Adult Census Dataset from the UCI Machine Learning Repository. The dataset was used for conducting experiments for the empirical evaluation of their algorithms.

3.2. "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid"- Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996

The dataset was utilised in this very famous research paper that was presented in the second International Conference on Knowledge Discovery and Data Mining by Ron Kohavi. This paper introduced the Naïve Bayes Hybrid Algorithm, the NBTree which then became very popular amongst data scientists. (PDF) A Statistical Approach to Adult Census Income Level Prediction, 2020)The algorithm was performed on many many datasets available on the UCI repository, mainly the Adult Census Dataset.

4. The intended use of the data

The data was collected and prepared to understand the factors/attributes that are associated with the income of an individual and several other characteristics that relate to and individual's socio-economic life.The dataset is mainly used for Classification and Clustering Tasks. The classification task is to determine whether or not a person makes over 50K US Dollars a year.This information can be obtained by selecting the useful features and applying classification algorithms on them. The dataset is widely used and cited, hence the maximum achieved accuracy of the classification task is also available for reference. It is also useful to identify clusters among the population which helps us better understand the clusters of people and their characteristics that affect their income, even today.

Gender discrimination in wages is still an issue faced by countries worldwide. The dataset can also be used to understand if gender affected income in the US in the 1990s. According to the data released by the U.S Department of Labor and Fortune, women make up 63 percent of workers earning the federal minimum wage. Women are in the lowest rung of the income ladder even after their significant presence in health care and food industry. (Gender Economic Inequality - Inequality.org, 2020)This clearly shows a disparity which has to be studied. Thus, the dataset has much greater significance in helping us understand the social, political and economic context that narrates the story of income in any society. New interventions with more data mining and analytics can help improve policy making and to reduce the concentration of wealth in a very small

percentage of the society. Identifying patterns from this dataset is very beneficial to develop an understanding of the most crucial factors that can increase or decrease income level of an individual.

5. Attribute types of the data

The dataset is multivariate and consists of 14 attributes

Attribute Name	Description	Type	Distinct
age	Age of the individual	Numeric	67
workclass	Work class of an individual	Nominal	8
fnlwght	The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for use by Population Division here at the Census Bureau	Numeric	4508
education	Education of an individual	Nominal	16
education.num	Education number of an individual	Numeric	16
marital.status	Marital status of an individual. It can be Widowed,Divorced, Separated,Never-married, Married-civ-spouse, Married-spouse-absent, Married-AF-spouse	Nominal	7
occupation	Occupation of an individual	Nominal	14
relationship	Relationship of an individual. It can be Husband, Wife, Not-in-family, Unmarried, Own-child and other relative	Nominal	6
race	Race of an individual. It can be White, Black, Asian-Pac-Islander, Other and Amer-Indian-Eskimo	Nominal	5
sex	Gender of the individual. It is categorised as Female and Male	Nominal	2
capital.gain	Capital gain of an	Numeric	79

	individual		
capital.loss	Capital loss of an individual	Numeric	56
hours.per.week	Hours per week of an individual	Numeric	78
native.country	Native country of an individual	Nominal	39
income	The class attribute. Income can be either greater than 50,000 or less than or equal to 50,000.	Nominal	2

Resampling of data

The dataset as obtained from Kaggle contained 32561 instances.

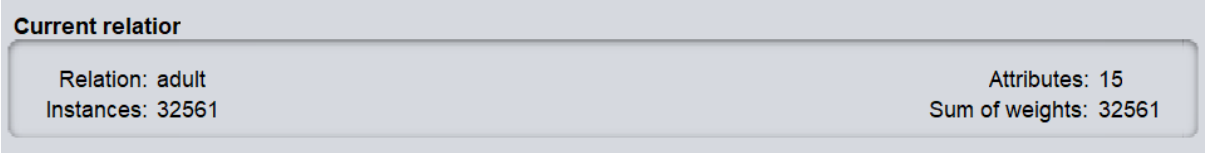


Figure 1 Original data

Processing of large data with 32561 instances and 15 attributes, particularly Feature Selection, Association rules causes delay in MLP and other advanced techniques. Hence it is ideal to resample the dataset to a smaller size without affecting performance.

Thus, the data is resampled to reduce the size of the data, for the purpose of this assignment. Resampling is done randomly and since this is a de-identified dataset, resampling doesn't affect the quality of the data. The details of resampling are:

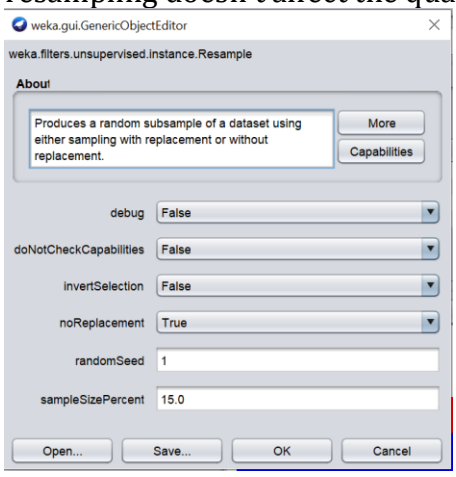


Figure 2 Resample Filter

The data is resampled to 15% sample without replacement. So now there are 4884 instances of data

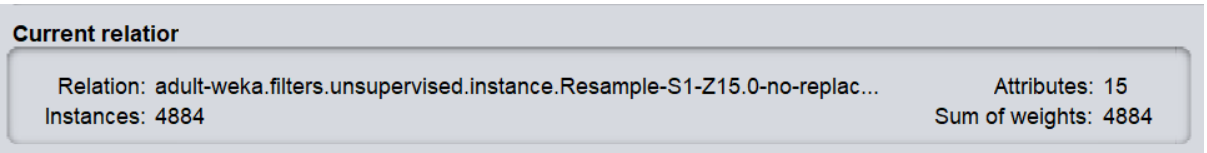


Figure 3 After resampling

The dataset is stored as “adult_resampled.arff” .

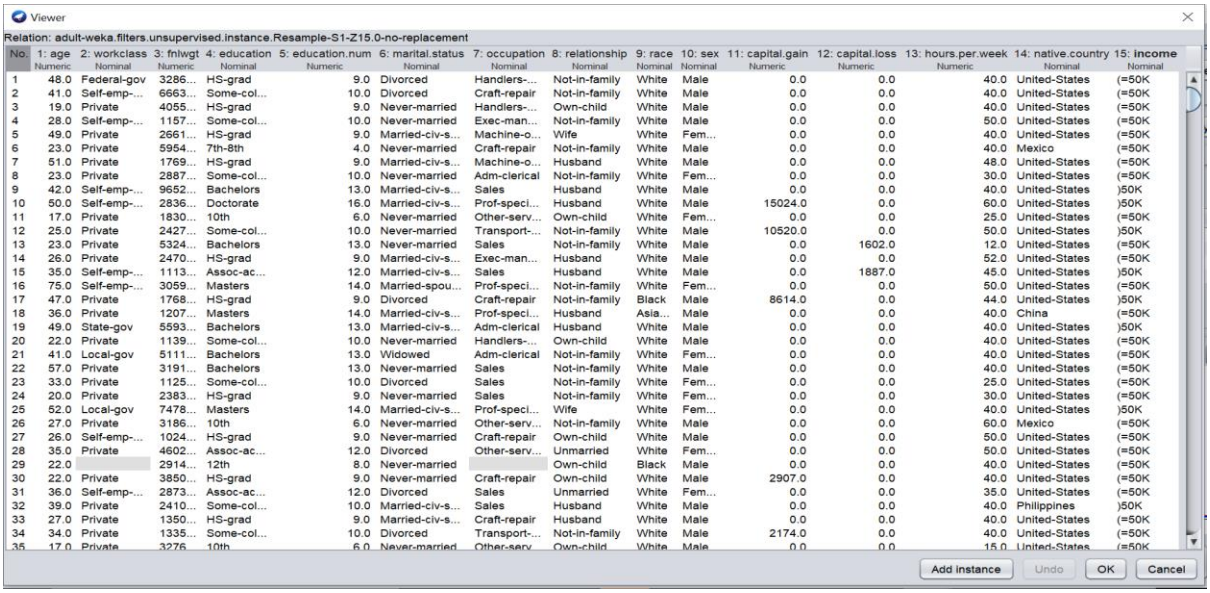


Figure 4 View of the dataset

Screenshots of summary and graphs of attributes

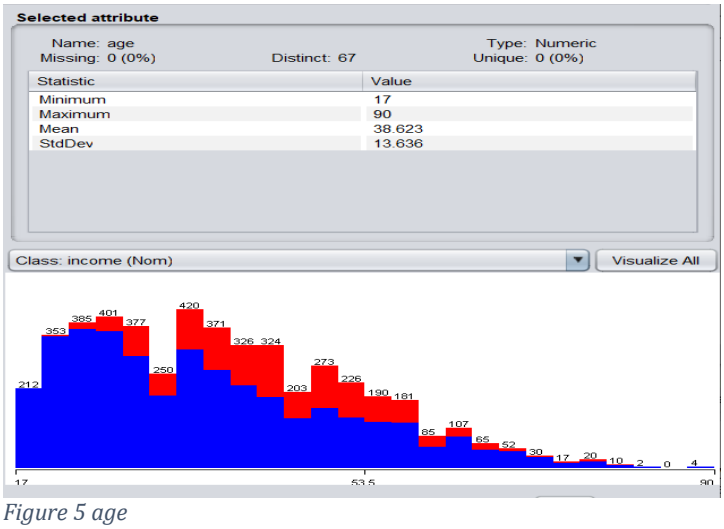


Figure 5 age

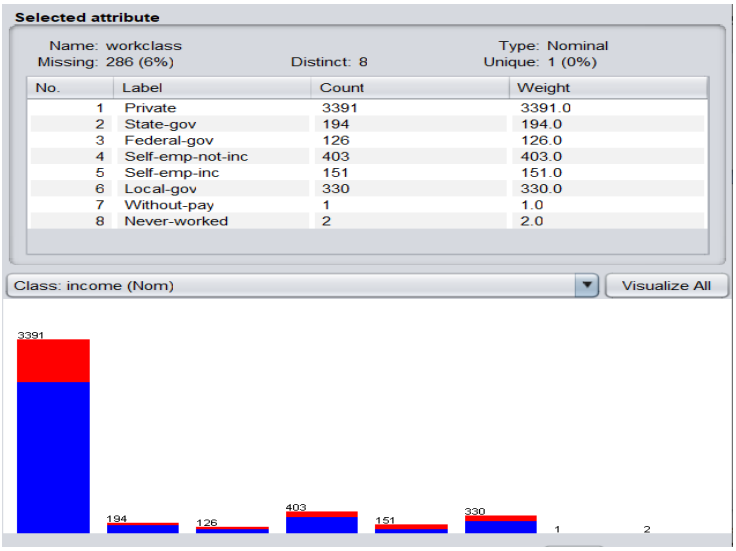


Figure 6workclass

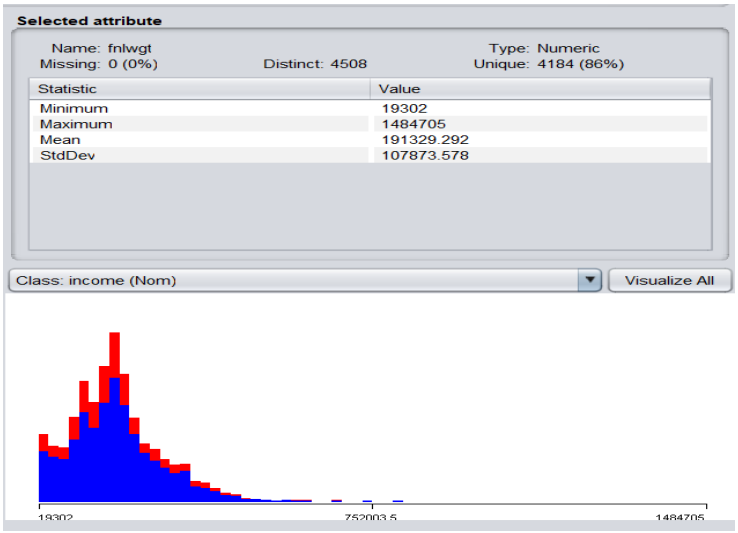


Figure 7fnlwgt

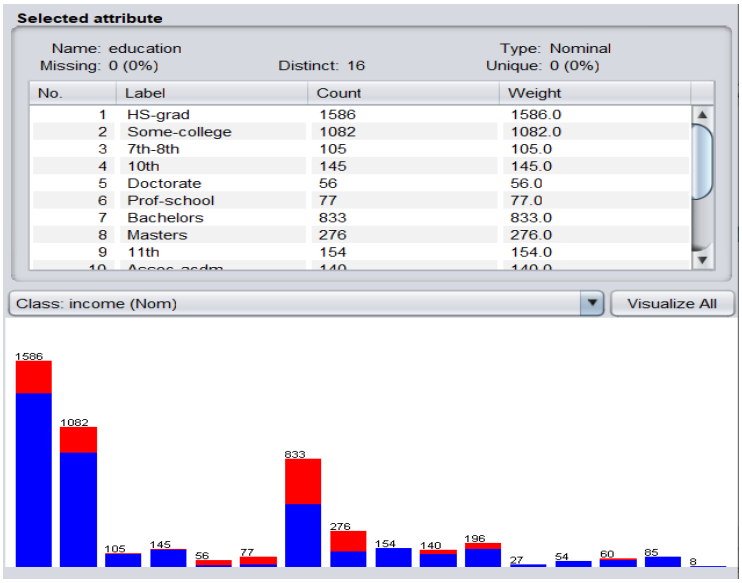


Figure 8 education

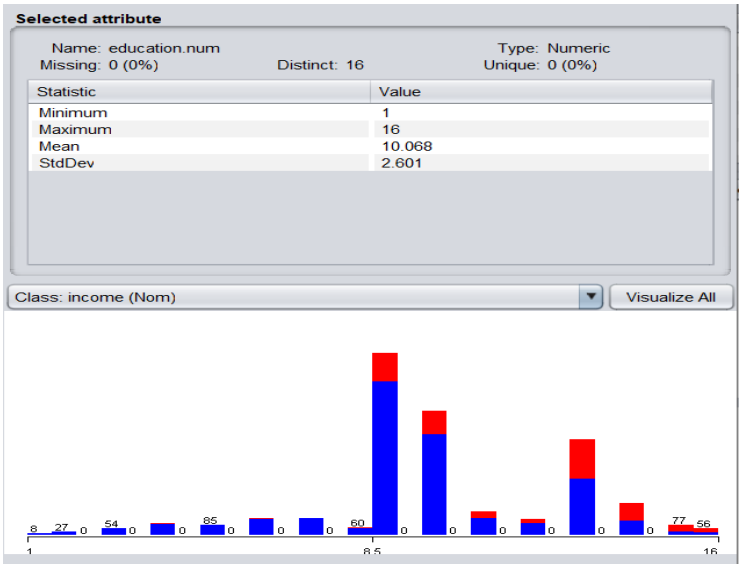


Figure 9education.num

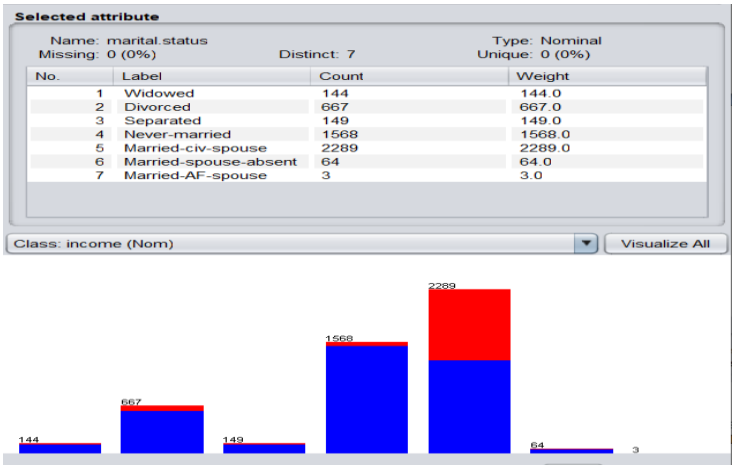


Figure 10marital.status

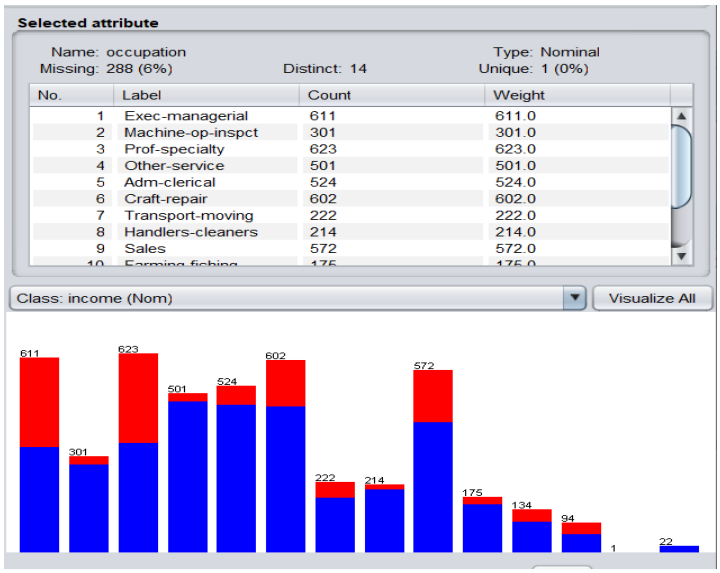


Figure 11 occupation

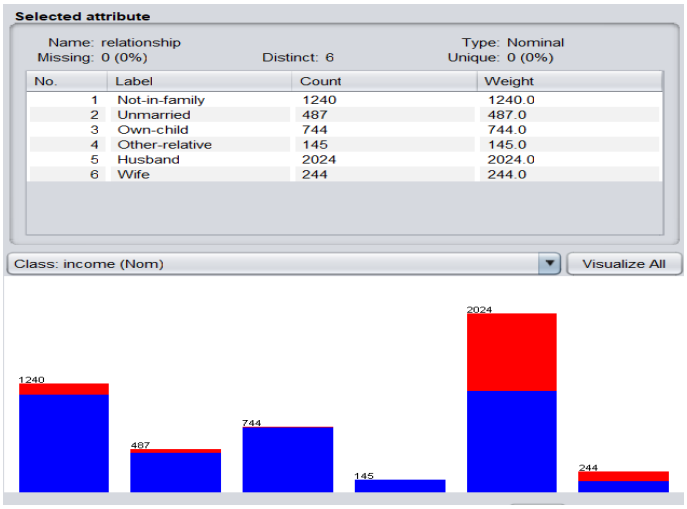


Figure 12 relationship

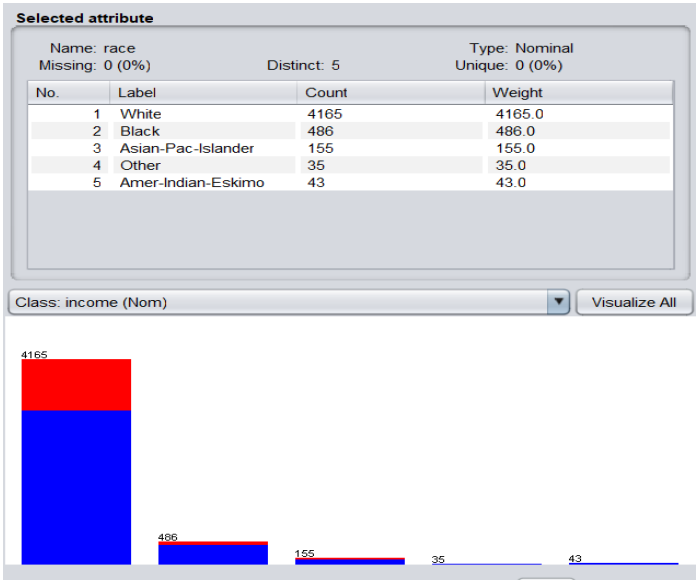


Figure 13 race

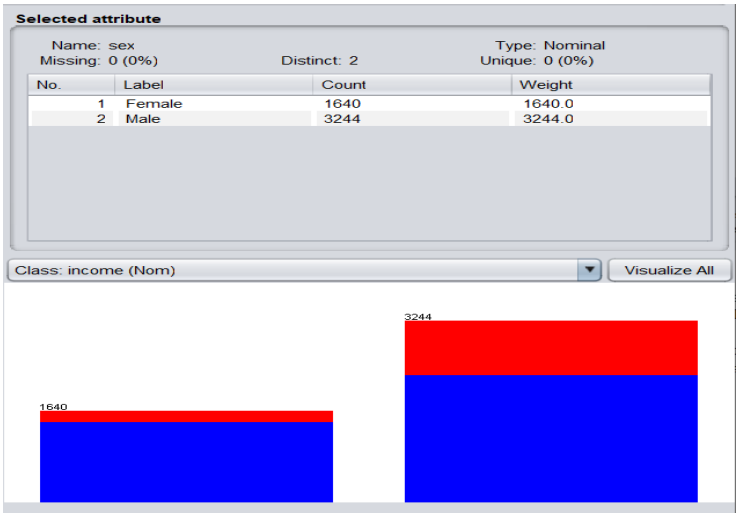


Figure 14 sex

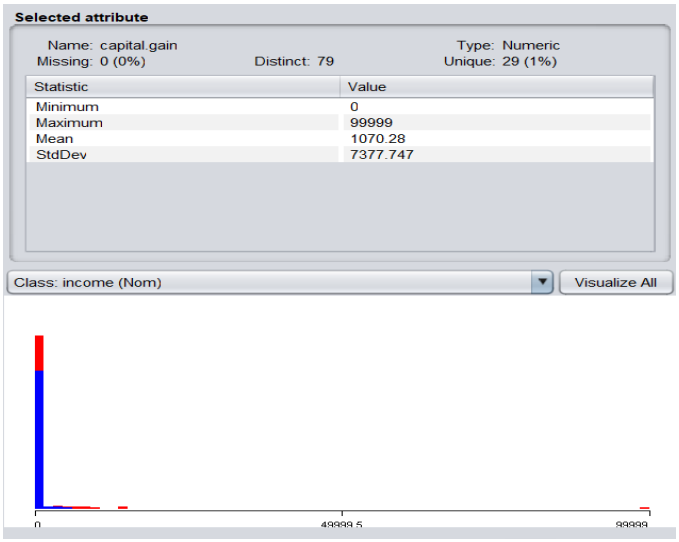


Figure 15capital.gain

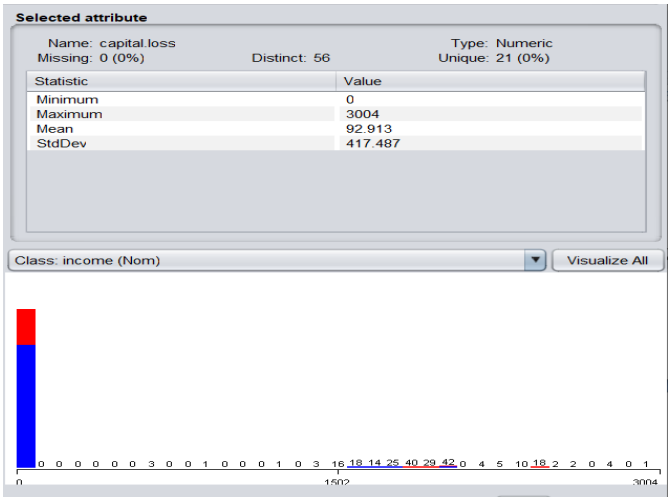


Figure 16capital.loss

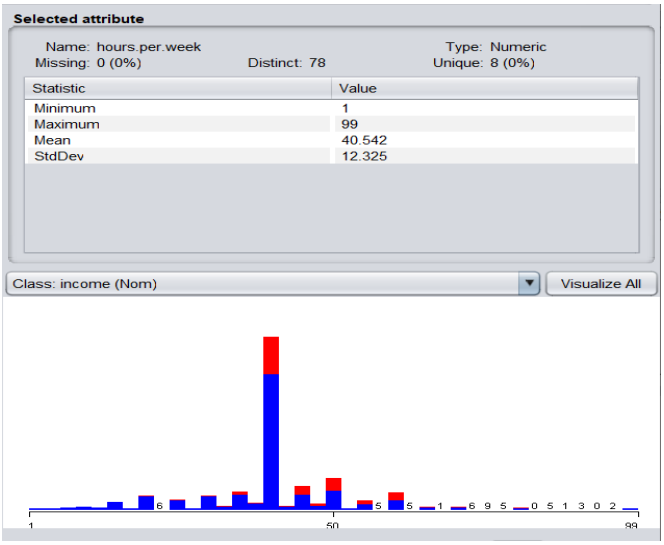


Figure 17hours.per.week

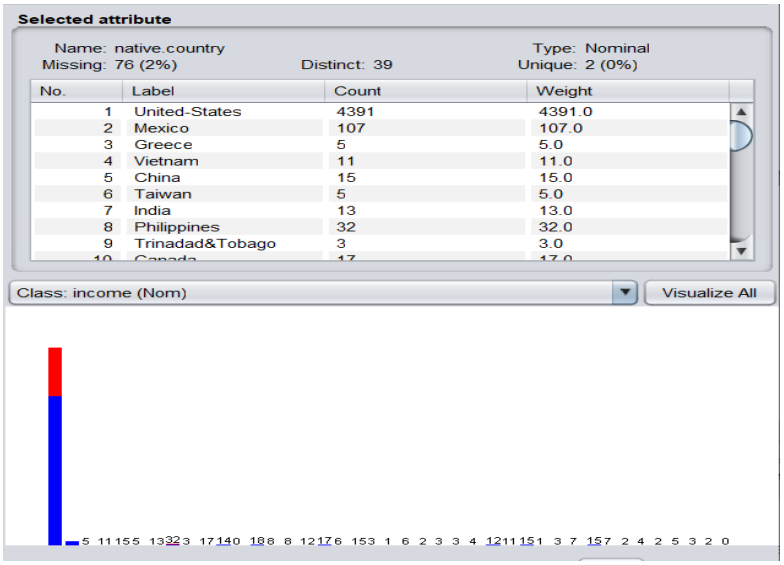


Figure 18native.country

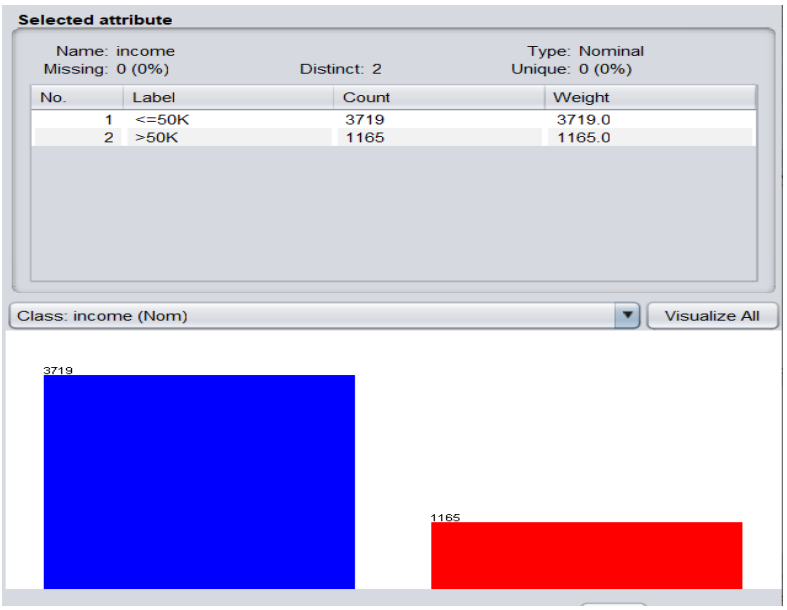


Figure 19income

Observations from attribute visualizations

- There are 15 multivariate attributes.
- Three attributes have very small percentage of missing values- 'workclass', 'occupation' and 'native.country'. None of the attributes have huge number of missing or noisy values.
- There are 6 numeric attributes and 9 nominal attributes.
- The dataset consists of 3719 instances with income equal to or less than 50,000 and 1165 instances with income exceeding 50,000. 'income' is the class that we intend to predict.

- **Objective**:

  1. To predict an individual's income through classification techniques such as MLP and J48,after understanding the effects of different factors such as age, gender,

occupation etc. through literature review and detailed study of the Adult Census Data Set and its attributes.

2. To apply various feature selection methods in Weka to identify the best features and to vary the hyperparameters to find and evaluate the models with better performance and accuracy after training the model and testing on the trained model.

3. To implement and evaluate ensemble methods such as Bagging, Boosting in Weka Experimenter to understand the effect of these ensemble methods on the model

- **Summary of Findings**:

Preprocessingtechniques suchasfeature encoding, feature selection, replacement of missing values, conversion of datatypes, outliers, normalization, discretization, class imbalance etc. were analysed for effectiveness on thisdataset.

1. **Preprocessing**

- 'fnlwgt' contains all 4508 unique values, thus they arenotusefulforclassification. It is thus not useful for data mining and is removedfrom the dataset.
- Normalization can be performed using Weka's normalize filter.
- Age is discretized into [min-31.6],[31.6-46.2],[46.2-60.8],[60.8-75.4],[75.4-max]
- Data Type Conversion is performed using NumericToNominal and StringToNominal to convert all the 14 attributes to Nominalvalues.
- Outliers and extreme values in the dataset are detected using Weka's unsupervised attribute InterQuartileRange Filter.
- Applying SMOTE filter increased the performance of classification algorithm J48inthisdatasetdramatically.ThisisbecausetheSMOTEfilterreducesclass imbalance.
- However, in this dataset class imbalance is not a problem and hence class balancing need not be performed as it will reduce the performance.
- All missing values are replaced with modes and means accordingly using ReplaceMissingValues Filter. Preprocessed file is saved as 'dataset.arff'.

2. **Splitting**

- The dataset is split into training and test dataset using 9:1 ratio and Resample filter. These files are stored in 'trainigSet.arff' and 'testingSet.arff'.

3. **Classification J-48**

- Threedifferentexperimentswerecarriedoutbysupplyingthetestsetand by varying the hyperparameters to see the performance of each model, the detailed description of each model and its results are included in Experiments sectionbelow.
- Infirstexperiment,confidencefactor0.25andminNumObjissetas2,with pruned tree and 84.86% of instances were correctly classified by this model.From,

the confusion matrix it is understood that the model is more good at identifying the people with income less than 50k (a<=50k) – correctly classified 352 out of 371 and only incorrectly classified it as b >50k only 19 instances out of 371, whereas in identifying patients with income exceeding, it only classified 63 correctly and 55 were classified incorrectly.

- In second experiment, using confidence factor 0.05 and minNumObj as 4, with pruned tree – 84.0491% of instances were correctly classified by this model.From, the confusion matrix it is understood that the model is more good at identifying the people with income less than 50k (a<=50k) – correctly classified 350 out of 371 and only incorrectly classified it as b >50k only 21 instances out of 371, whereas in identifying patients with income exceeding, it classified 57 correctly and 61 were classified incorrectly.
- **Highest accuracy achieved with J48 is 85.0716 in third experiment**. This is achieved due to the parameters are varied to tune the model and it increased the model accuracy to reach 85.0716%. However Root Mean Squared Error(RMSE) decreased to 0.3282 from 0.3292.From, the confusion matrix it is understood that the model is getting better at identifying the people with income exceeding 50k also as compared to patients with risk of readmission (b=1)– correctly classified 223 out of 539 and only 316 instances out of 539 incorrectly classified it as 1, whereas in identifying patients with risk of readmission (b=1), it classified 814 instances correctly and 134 were classified incorrectly.
- TheresultbuffersandmodelsofalltheJ48experimentsarestoredin'J48 RESULT BUFFERS AND MODELS' with appropriate and self-explanatory names.
- 'Experiment1J48tree.arff','Experiment2J48tree.arff', Experiment3J48tree.arff' are the files that contain J48 visualization with predicted class.

4. Classification- MLP

- In Experiment 1, the model seems to have performed better when we see the confusion matrix, 81 instances are correctly classified as people with less than or equal to 50k income and none of instances are incorrectly classified as people exceeding 50k. This is a really good sign. However, 22 people are incorrectly classified as people with income less than or equal to 50k and only 7 instances are correctly classified as people exceeding 50k. This means the model is better at identifying people with income less than or equal to 50k.
- With hidden layer= "a", (attributes + classes)/2=(14+1)/2 layers are used while developing the neural network. It produced an accuracy of 80% and mean absolute error of 0.1964.
- In Experiment 2, accuracy is reduced to 88.8%. Mean absolute error also increased to 0.1444 from 0.08. This can be due to the reduction of number of hiddenLayers and a lightweight MLP with less layers.
- The confusion matrix shows that 75 instances are correctly classified as people with less than or equal to 50k income and 4 instances are incorrectly classified as people exceeding 50k. This is a really good sign. However, 7 people are incorrectly classified as people with income less than or equal to

50k and only 11 instances are correctly classified as people exceeding 50k. This means the model is better at identifying people with income less than or equal to 50k.

- In Experiment 3, accuracy is dramatically increased to 96.9072 and mean absolute error is just 0.032. With hidden layer =i, we have 14 layers.
- MultiLayer Perceptron seems to be working really well with this dataset which has numerical attributes. It is comparatively slower, but there are less errors in classification which is really good for adult census dataset.

2. Preprocessing – 10%

1. Selecting or filtering the attributes

Feature selection is an extremely important step in any data mining process. It not only reduces the dimensions, but also improves the performance of the algorithms to a great extent. The dataset contains 14 attributes, thus removing or selecting certain features based on their importance and correlation, increases the ease for processing, improves accuracy and overall provides better results. To select the predictor variables for readmitted class, it is essential to remove the features that are futile for the task.

- Most of the features seems to be important to understand an individual's income and predicting the class.
- There are no features with more than 40% noise or missing values, which is the general rule of thump to remove features, hence features need not be removed based on that criteria.

Various feature selection methods are also used to select the best features.

ClassifierAttributeEval with Ranker

It evaluates the worth of an attribute by using a user-specified classifier.

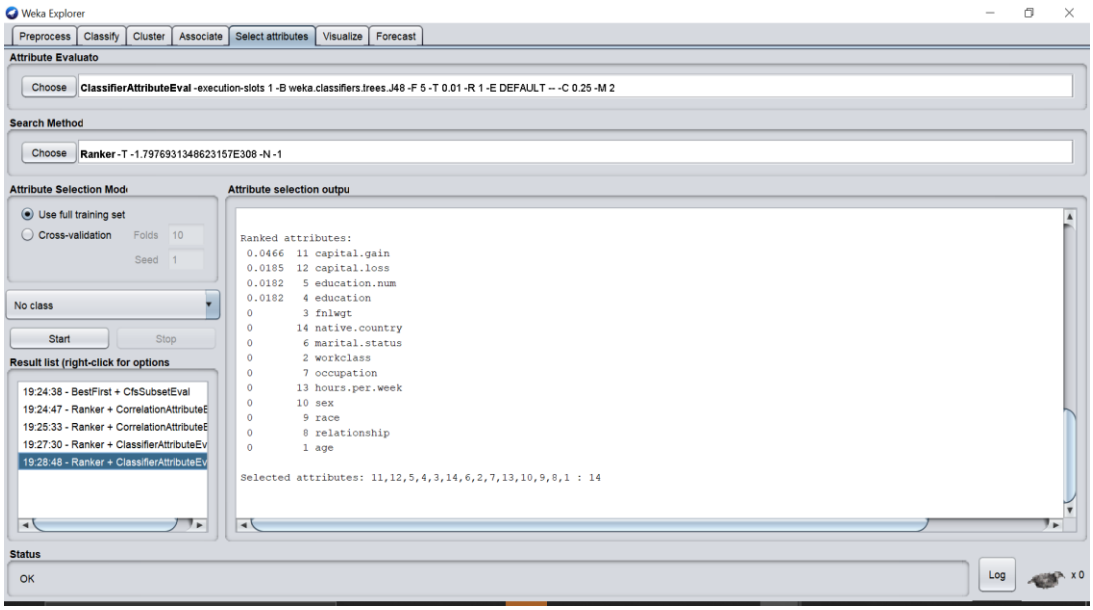


Figure 20ClassifierAttributeEval

ClassifierAttributeEval with Ranker ranked 1 features useful to classify income and more rank is given for ‘capital gain’, ‘capital loss’, ‘education and education.num’. ‘fnlwgt’,‘age’ etc are the least and equally ranked attributes.

InfoGainAttributeEvalwith Ranker

According to Weka Documentation, this feature selection method evaluates the worth of an attribute by measuring the information gain with respect to the class.

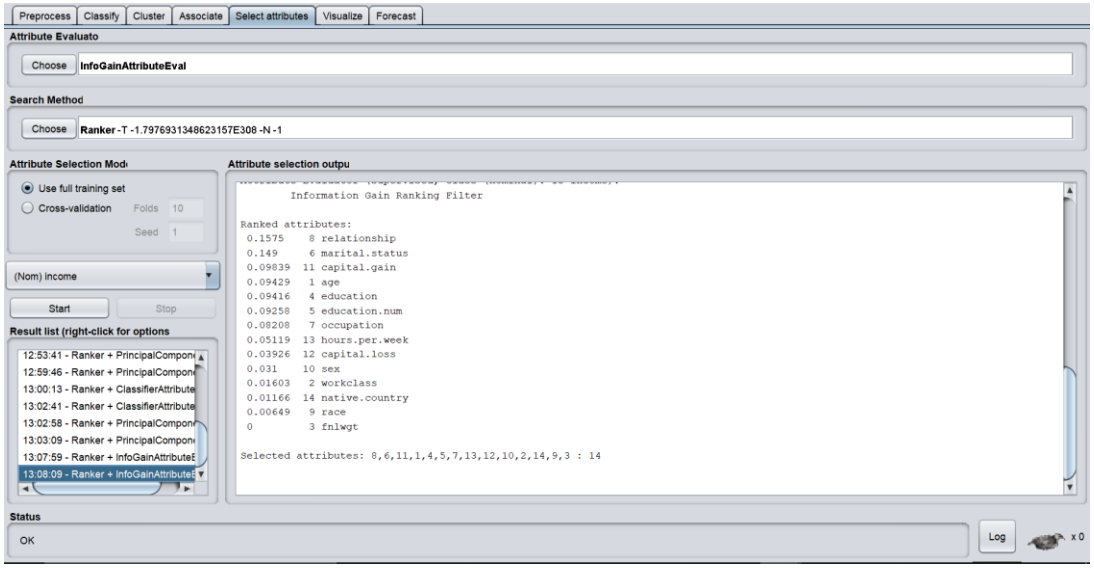


Figure 21InfoGainAttribute Eval

The output of this feature selection method shows relationship, marital status, capital gain and age as the most important attribute. ‘fnlwgt’ is the least important attribute in terms of information gained about the class.

Visualization

Visualizing all attributes together with the ‘income’ class can be useful for Exploratory Data Analysis and feature selection.

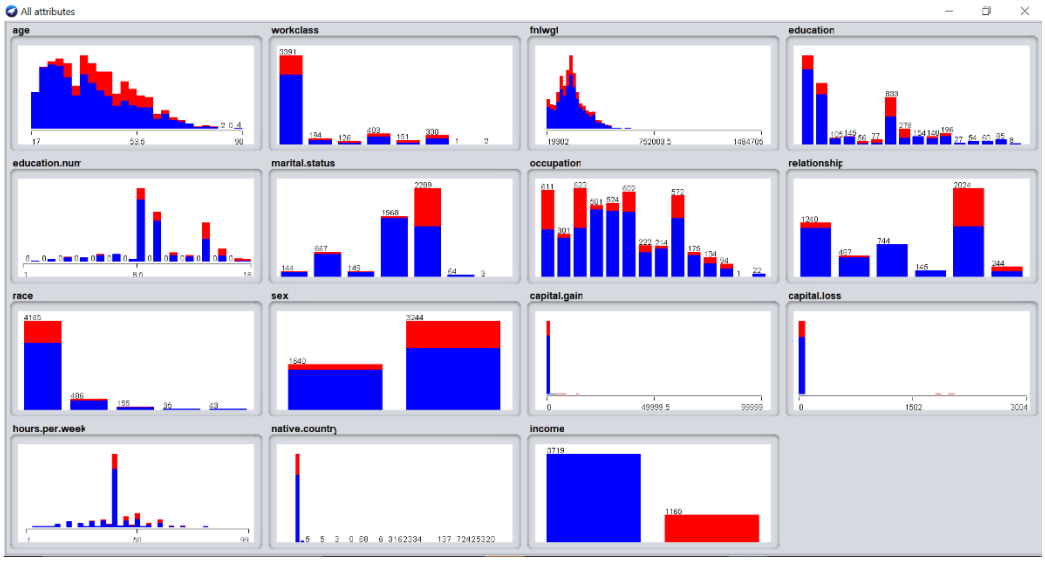


Figure 22 Visualization of all attributes

After feature selection and careful investigation of the features, 14 features are selected for performing experiments. The only feature that is removed is 'fnlwgt' as it doesn't contribute much in predicting the income of an individual, it is the number of units in the target population that the particular unit represents.

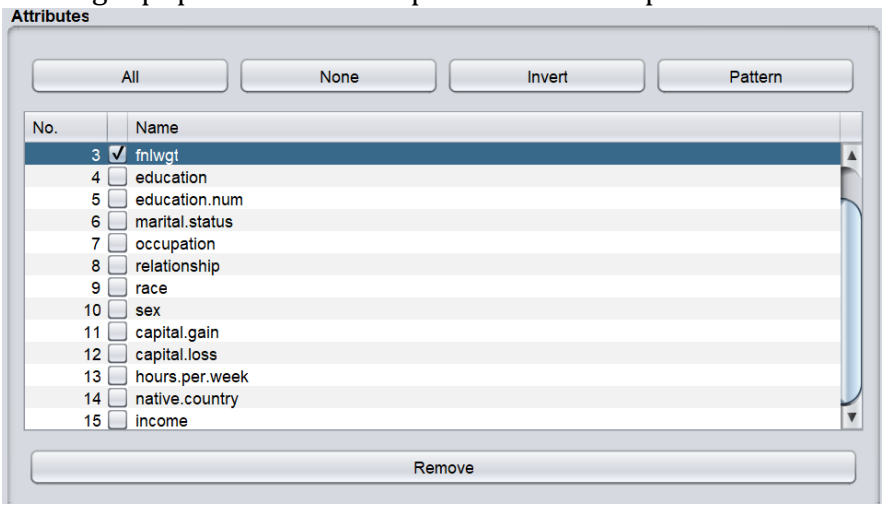


Figure 23 Removing 'fnlwgt'

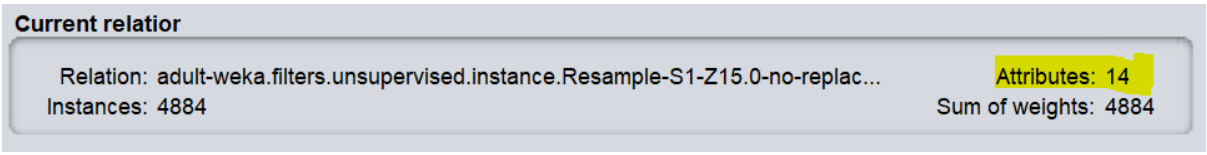


Figure 24 After removal

2. Discretization (Binning)

Binning may improve the accuracy of predictive models by reducing the non-linearity or noise and it is also useful for certain classifiers. After binning, outliers, missing or invalid values can be easily identified. This can be performed using Weka's Discretize filter. In the dataset, age has 67 distinct values which can be discretized to 4 bins.

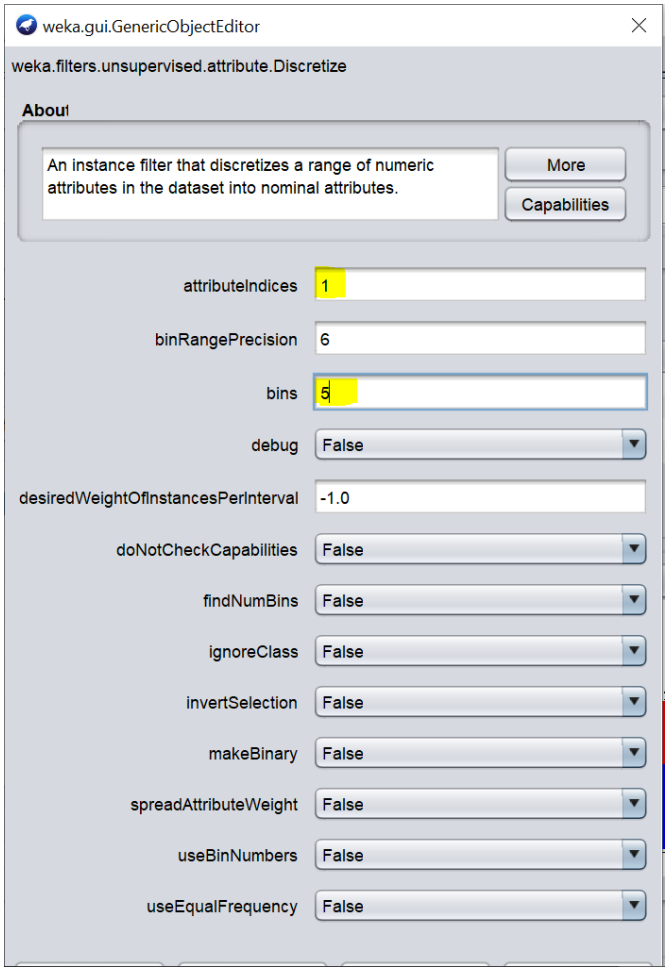


Figure 25 Discretize filter

After discretization, it is useful to replace the bin range names with more readable and comprehensible names using a text editor and it is done as shown below.

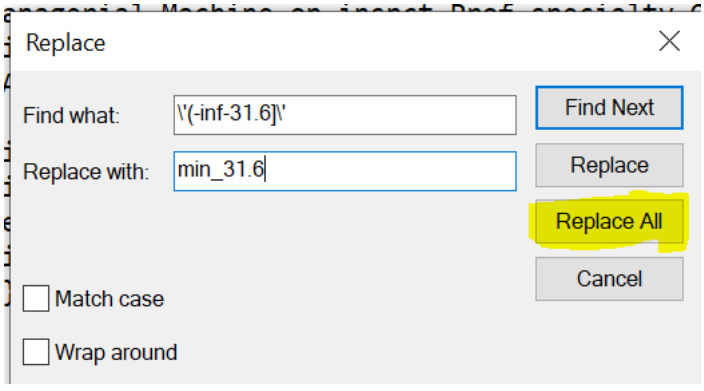


Figure 26 Replacing with proper labels.

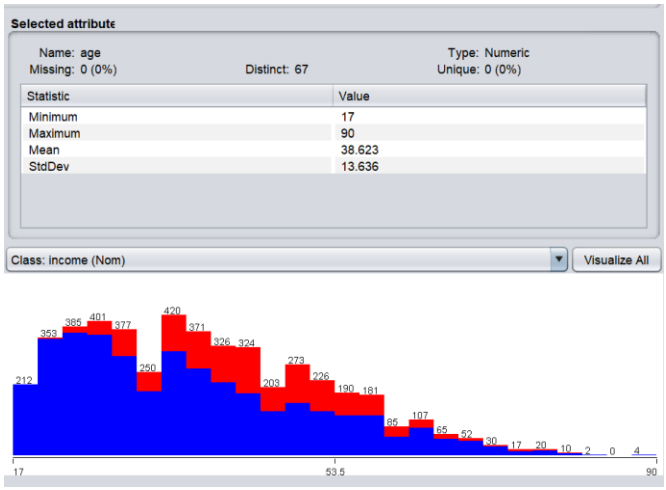


Figure 27 Before discretization

Figure 28 After discretization

3. Normalization

Normalization can be performed using Weka's normalize filter. This ensures that numerical values are scaled by removing the different min and max values and scaling them down to a standard range which makes certain machine learning algorithms work better (Dr.AbubakrSiddig- Datasets,EDA and altering data structure lecture , 2020).

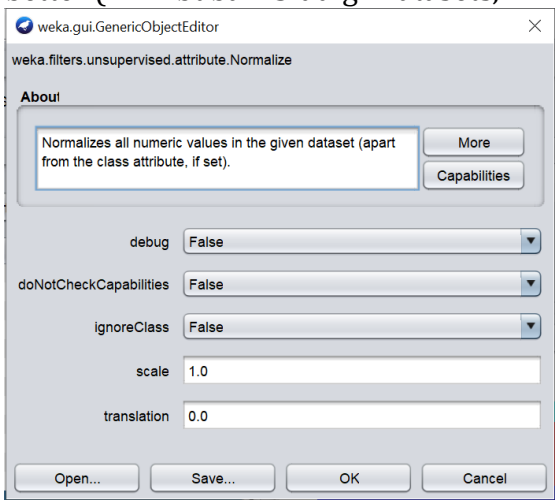


Figure 29 Normalize filter

However, for this dataset, most of the selected features are not benefitted from normalizing. As there are hours per week, education number etc. Normalizing them is not beneficial. Applying normalization was also not siginificantly beneficial when tested using 10-fold crossvalidation. Hence the preprocessed data is not normalized.

4. Data Type Conversion

Data type conversion is necessary for certain algorithms like J48. J48 requires all attributes to be nominal. Hence conversion is performed using NumericToNominalfilters.

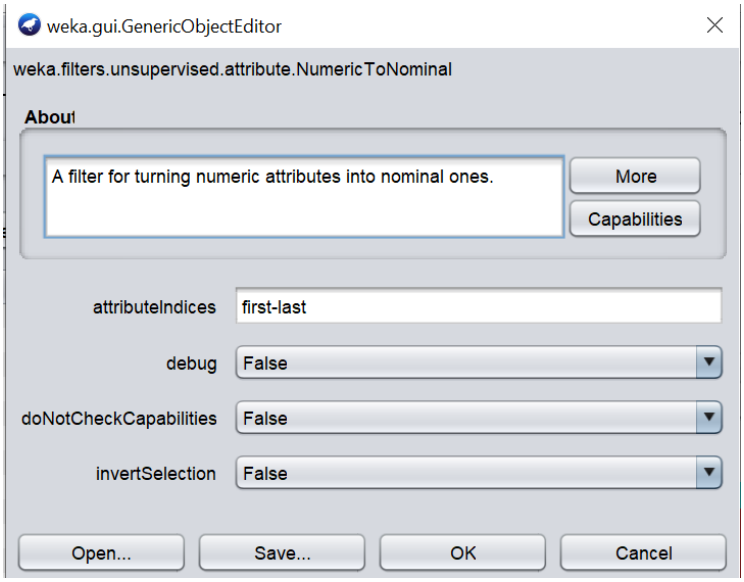


Figure 30NumericToNominal Filter

Figure 31 Before conversion

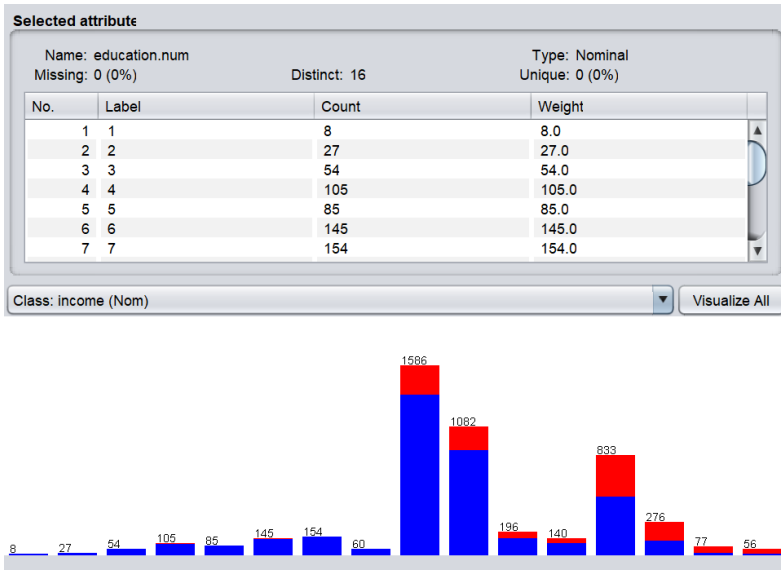


Figure 32 after conversion

5. Missing Values

Data can be missing due to a variety of reasons- hesitance of the respondents to provide complete information, malfunctioning of equipments, errors when entering the data in to the database, sudden changes etc etc(Dr.AbubakrSiddig- Datasets,EDA and altering data structure lecture , 2020). A small amount of missing value is almost unavoidable in large datasets. However, a significant percentage of missing values can be problematic. There are only three attributes with missing values; 'workclass'-6% missing values, 'occupation' -6% missing values, Missing values of these attributes are removed by replacing with modes and means using ReplaceMissingValues Filter.

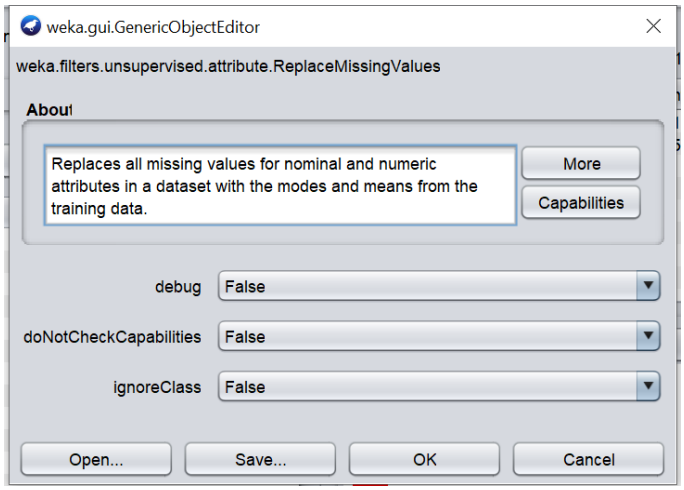


Figure 33 Replace Missing Values Filter

6. Outliers

Outlier is a data point that differs significantly from other observations which can be due to error in measurements or exceptional cases (Wikipedia Outlier, 2020). Outliers and extreme values in the dataset are detected using Weka's unsupervised attribute InterQuartileRange Filter. It gives us the middle spread of the data. This filter skips the

class values. It creates two new features Outlier and ExtremeValue with two distinct values 'No' and 'Yes' for all instances.

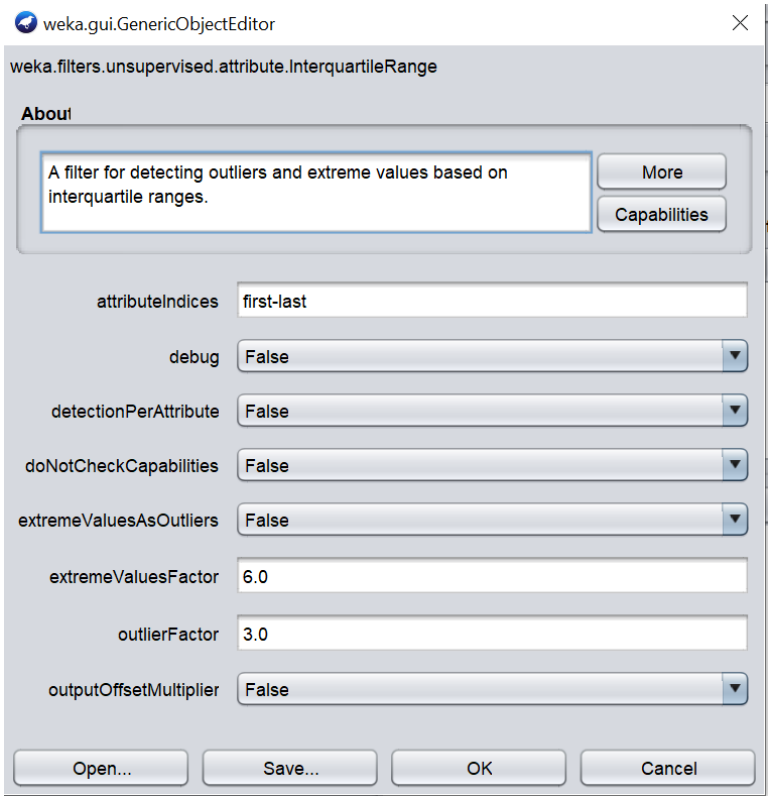


Figure 34InterQuartileRange Filter

This dataset do not contain any outliers or extreme values.

7. Class Imbalance handling with SMOTE(Synthetic Minority Oversampling Technique)

Class imbalance is a problem where "the total number of a class of data (positive) is far less than the total number of another class of data (negative)". Machine learning algorithms tend to work better with roughly equal classes (Chioka, 2013). Weka's SMOTE filter can be used to balance the classes. SMOTE filter was installed using Weka's Tools Option with PackageManager.
However, in this dataset class imbalance is not a problem and hence class balancing need not be performed as it will reduce the performance.

8. Reordering of Attributes

Weka's Reorder filter can be used to reorder the attributes. However, in this dataset the class attribute 'income' is already set as the last attribute and there is no need to rearrange the order of the attributes. Hence this will not be performed.

All of the above steps and their effects on the data is checked during the preprocessing stage. Not all of these methods are applied on the data as some of them are not effective on this dataset as clearly mentioned in each of the cases above. The preprocessed data is stored in 'dataset.arff'.

3. Divide your dataset into training and test set – 0%

Divide the test into a training and testing set in the ration of (9:1).
The files generated as part of this process should be saved and submitted as the following

- trainingSet.arff and
- testingSet.arff

Screen shots of these files should be included.

The dataset is divided into training and test dataset. It was also initially divided into crossvalidation set for evaluation. However, only training and test data sets are submitted. Training data is the sample of data that is used to fit the learning model and test data is defined as "The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset" (Medium, 2017).

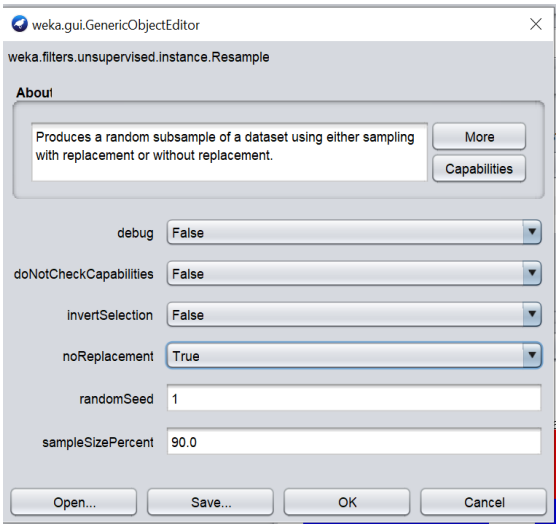


Figure 35 Creating Training Data

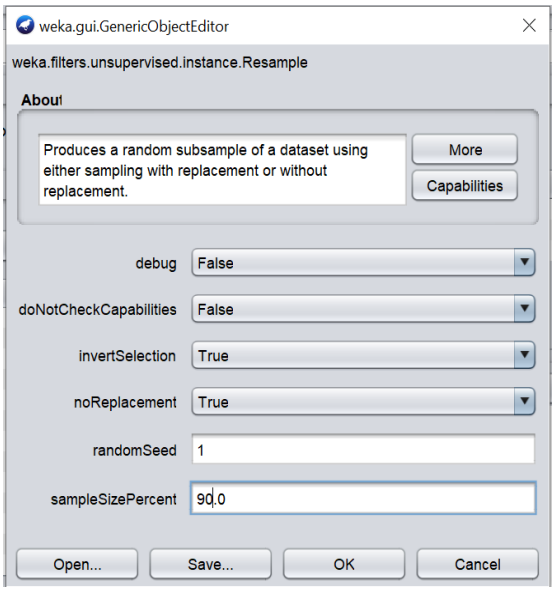


Figure 36 Creating test data

While creating test dataset, previous operation is undone and invertSelection is set to True. This is because we want our dataset to be properly split between training and test data.
These files are saved in trainingSet.arff and testingSet.arff

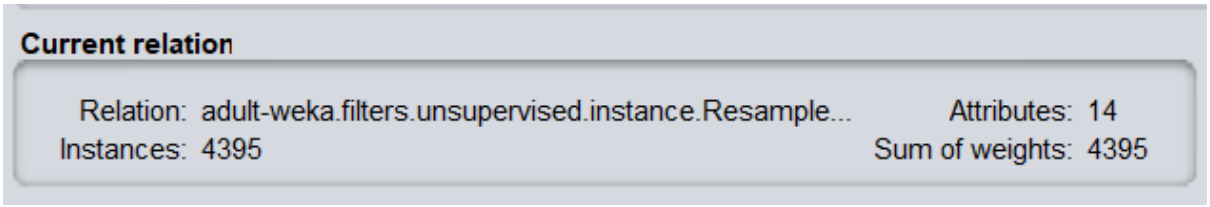


Figure 37 Training data

Figure 38 Training data file

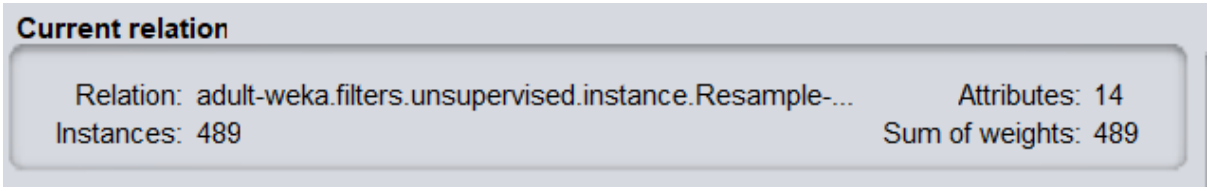


Figure 39 Test data

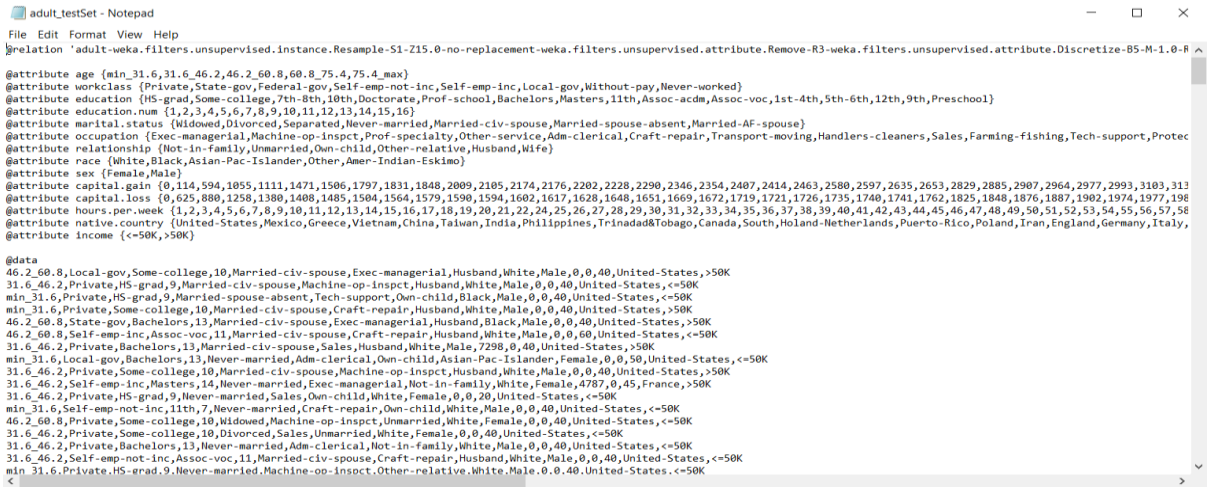


Figure 40 Test data file

Data Mining Techniques

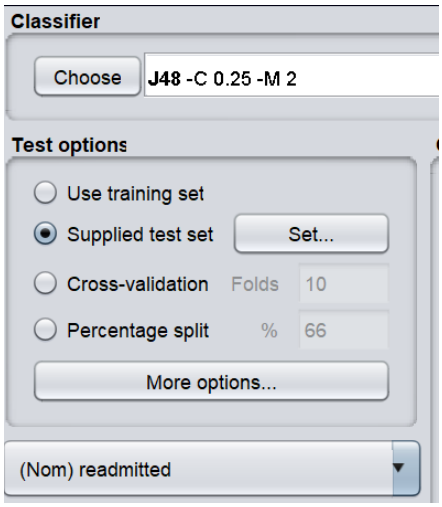
4. Classification/ Association: J48 Tree or Association Rules – 10%

Classification- J48 Tree

For the classification task, J48 decision tree is used. Decision trees are versatile and they can fit into complex datasets with a divide and conquer approach. (Dr.AbubakrSiddig- Decision Trees- lecture , 2020). J48 provides the best performance, the diagonal elements will be evenly distributed.

Experiments using training and test data

Weka Explorer is used to test the trained model on a supplied test set. This is a very useful feature. Parameters are varied to see the effect in performance. All experiments are performed with J48 algorithm. Parameters will restrict the freedom of the tree thus preventing overfitting (Dr.AbubakrSiddig- Decision Trees-2 lecture , 2020)



The parameters that are significant when performing J48 (As per the Weka definitions) are:

1. minNumObj - The minimum number of instances per leaf. Increasing this parameter will decrease the treesize as we will see below.
2. unpruned- Whether pruning is performed. Both pruned and unpruned options will be tested.
3. confidenceFactor - The confidence factor used for pruning (smaller values incur more pruning).
4. subtreeRaising - Whether to consider the subtree raising operation when pruning.

All of the experiments below are performed using supplied test set -testingSet.arff

Experiment 1

For the first experiment, the parameters chosen are:

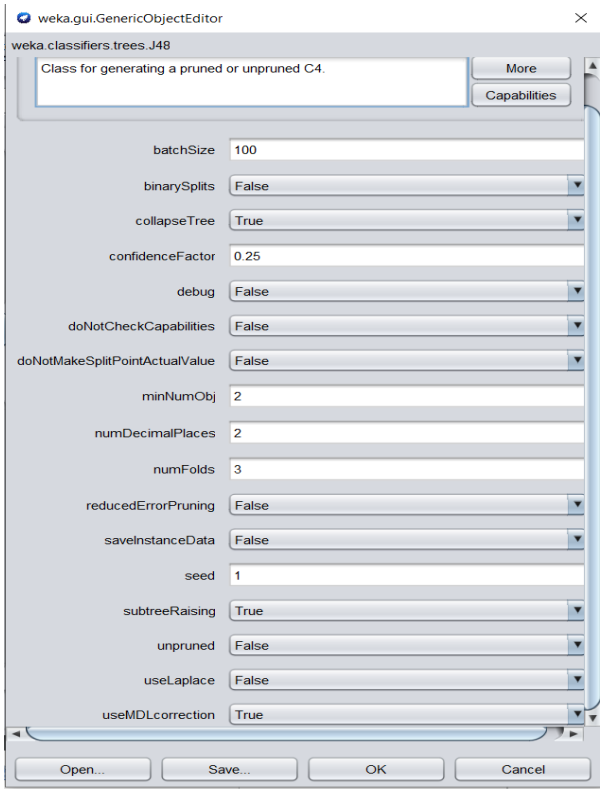


Figure 41Experiment 1-J48-parameters

Leaving parameters unstrained will fit the tree very closely and mostly overfit the data (Dr.Abubakr Siddig- Decision Trees-2 lecture, 2020). However, the parameters are kept set in their default options initially, to see the performance of the algorithm with these options.

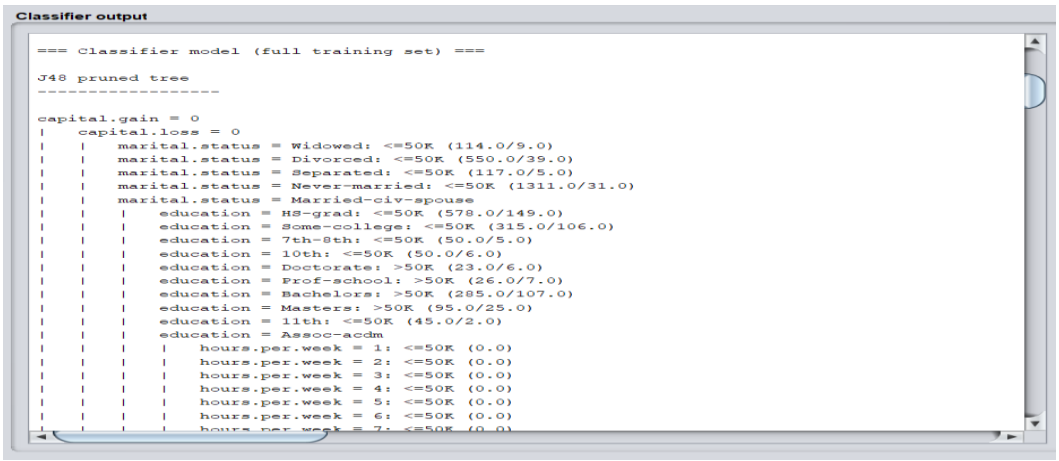


Figure 42 Experiment 1- J48 model beginning

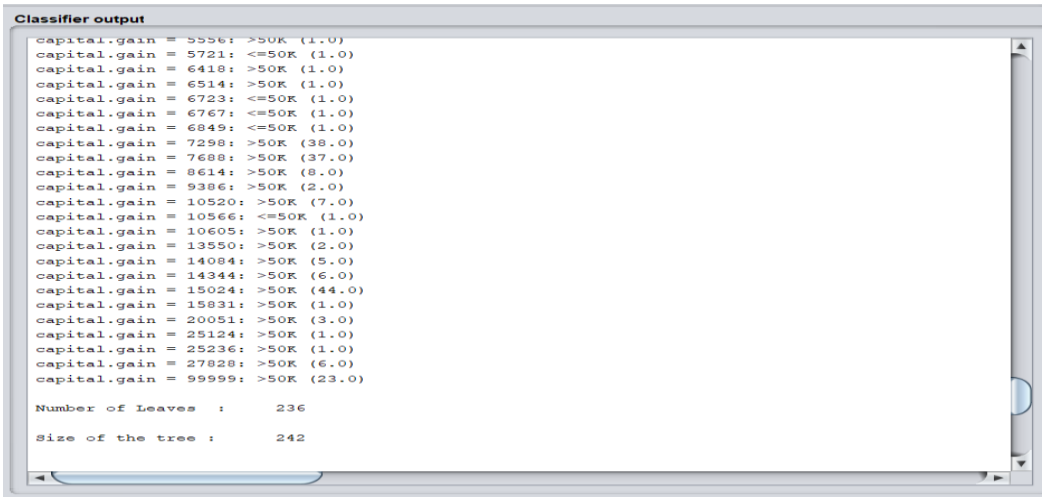


Figure 43Experiment 1- J48 model end

Results with confidence values, confusion matrix, tree, rules

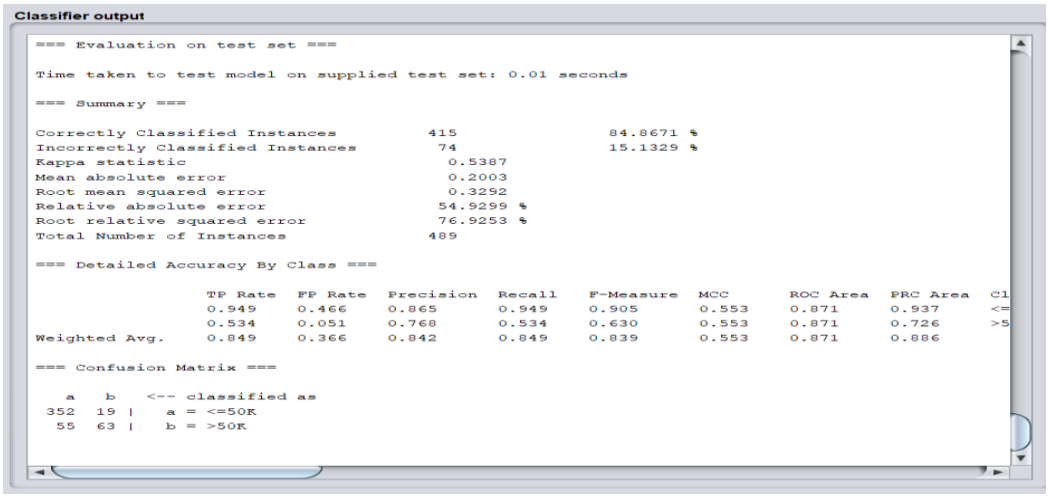


Figure 44 Experiment 1- J48 –Results

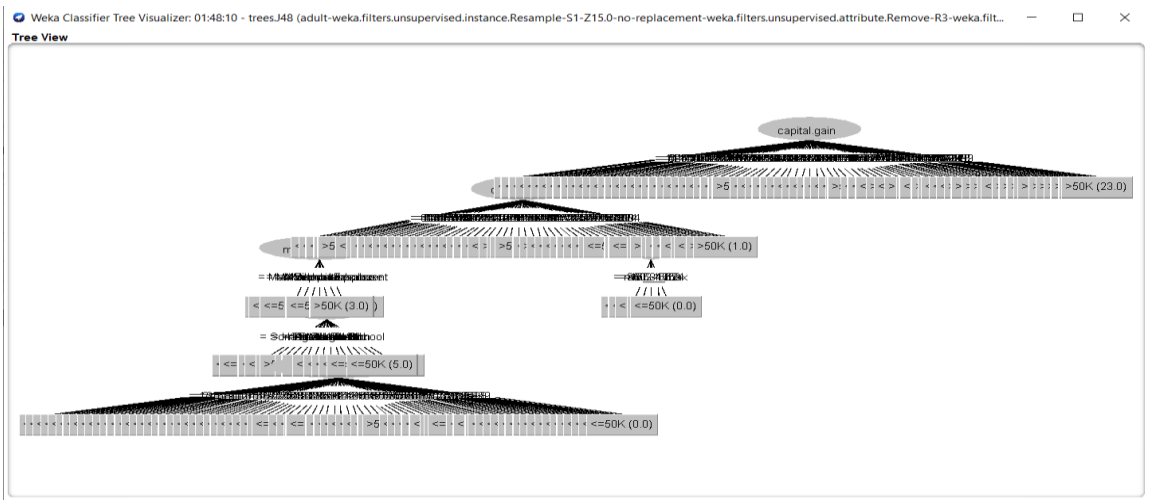


Figure 45 Experiment 1- J48 -Tree

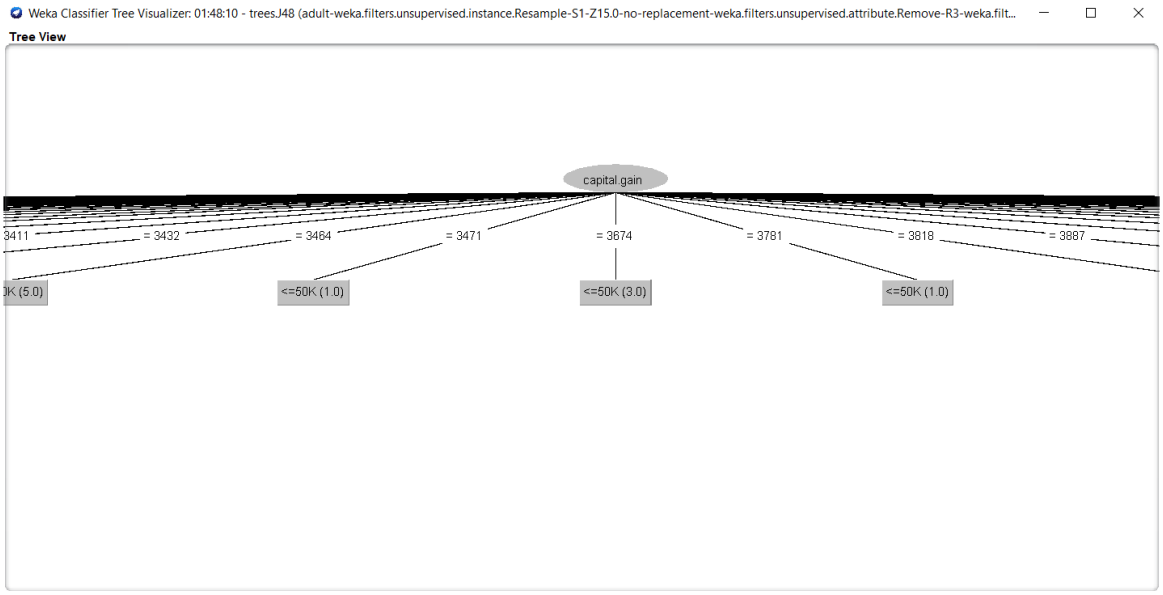


Figure 46 Experiment 1-J48-Tree

Findings

1. Using confidence factor 0.25 and minNumObj as 2, with pruned tree – 84.8671% of instances were correctly classified by this model.
2. Pruned tree increased the accuracy. The model used a pruned tree because pruning mostly reduces the complexity of the final classifier, reduce overfitting to increase predictive accuracy. (Wikipedia Decision Tree Pruning, 2020)
3. From, the confusion matrix it is understood that the model is more good at identifying the people with income less than 50k (a<=50k) – correctly classified 352 out of 371 and only incorrectly classified it as b >50k only 19 instances out of 371, whereas in identifying patients with income exceeding, it only classified 63 correctly and 55 were classified incorrectly.
4. From the visualization of the tree, it is evident that the tree gained more information from the 'capital.gain' attribute and further branching was made on this attribute. This also seems correct as this variable was the most ranked during feature selection. 'capital.gain' is the best feature on which the data is further split.

Experiment 2

In experiment 2, confidenceFactor, binarySplits and minNumObj is varied. confidenceFactor is decreased for more pruning (smaller values incur more pruning). •

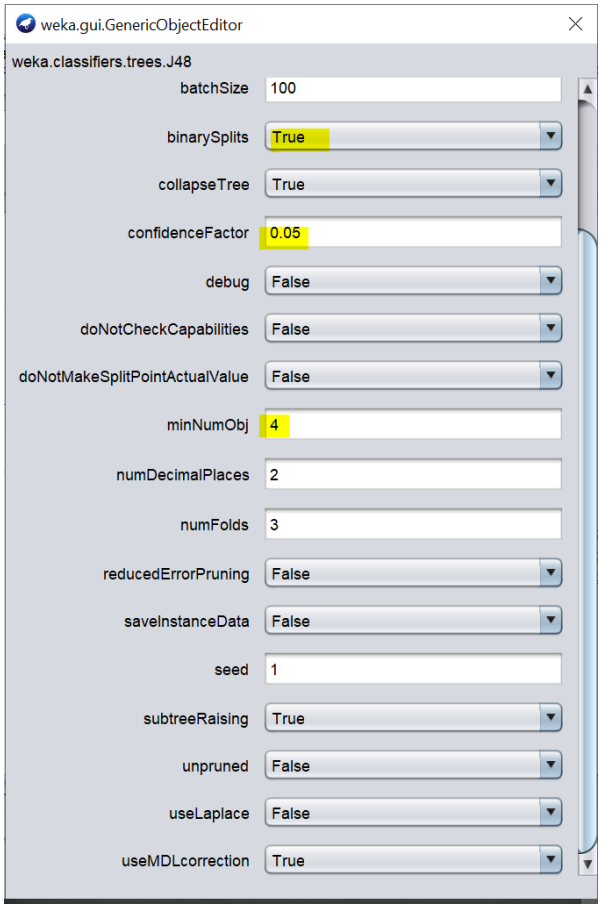


Figure 47 Experiment 2 J48- parameters

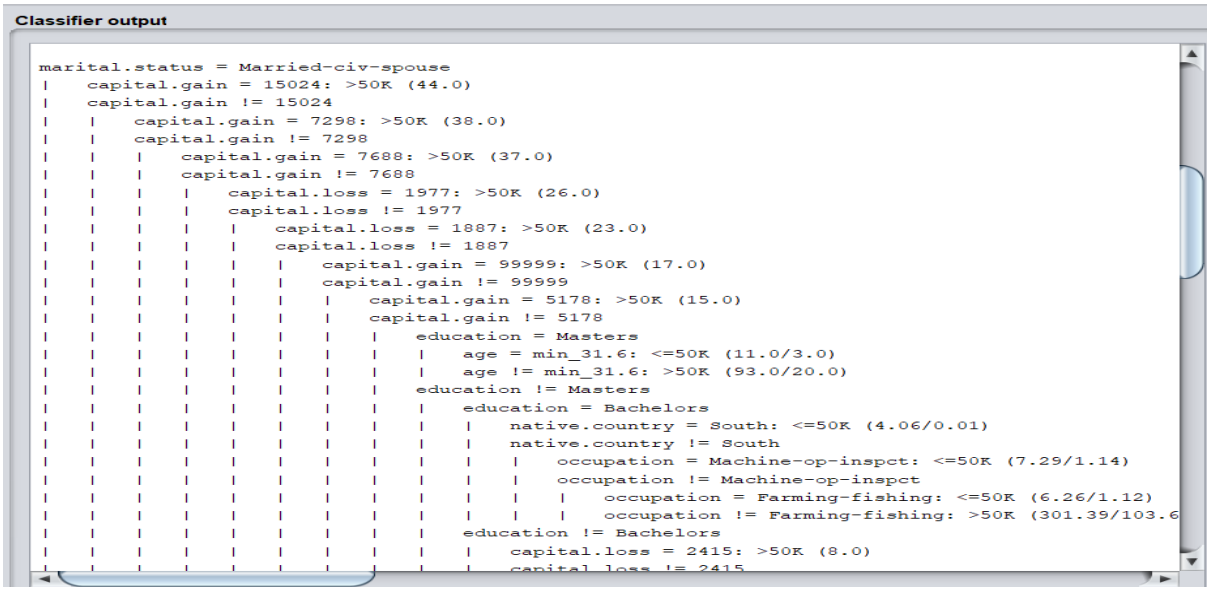


Figure 48 Experiment 2-Model(small part)

Results with confidence values, confusion matrix, tree, rules

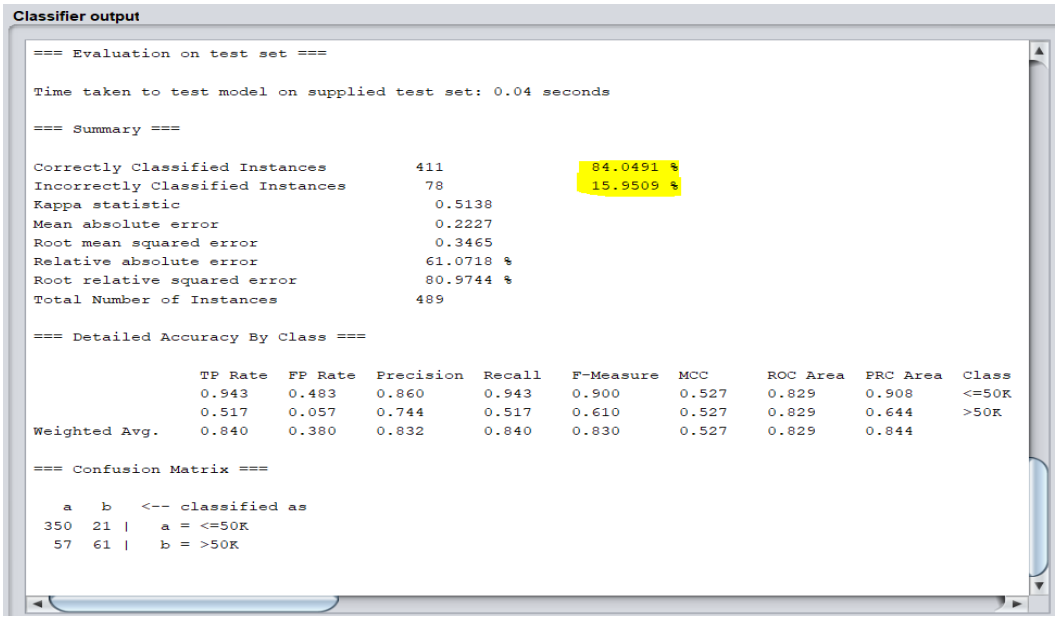


Figure 49 Experiment 2 Result J48

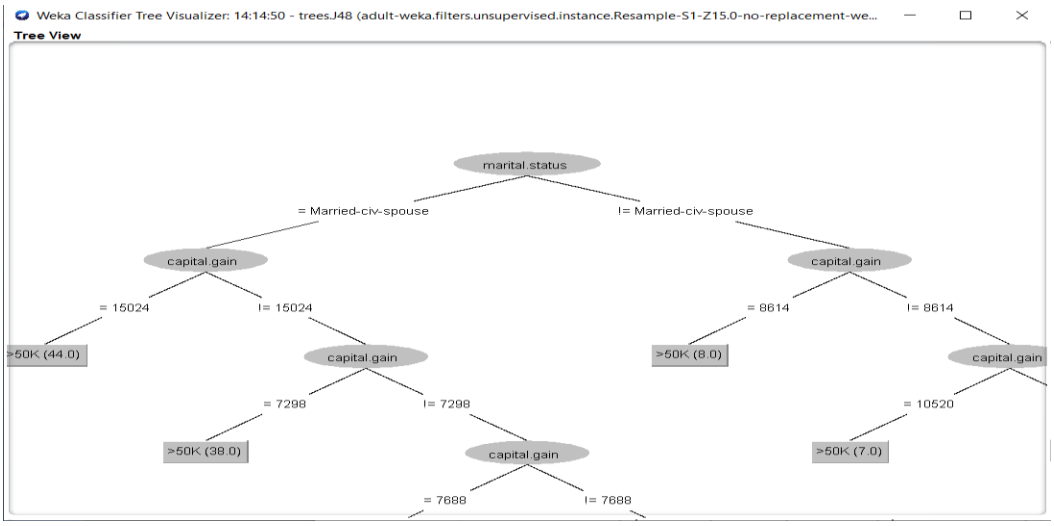


Figure 50 Experiment 2-J48 –Tree root

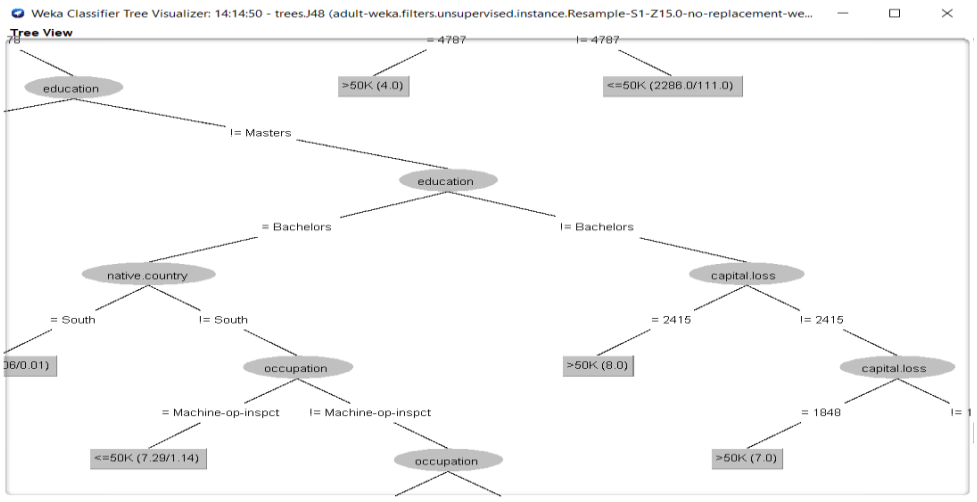


Figure 51 Experiment 2-J48 tree Auto scaled

Figure 52 J48 Full tree after binary split

Findings

1. Using confidence factor 0.05 and minNumObj as 4, with pruned tree – 84.0491% of instances were correctly classified by this model.
2. However, using more confidence didn't benefit the performance, due to over fitting of the model.
3. minNumObj was increased to 4 to reduce the tree size. binarySplits is set to true to use binary splitting on nominal variables while building tree.
4. From, the confusion matrix it is understood that the model is more good at identifying the people with income less than 50k (a<=50k) – correctly classified 350 out of 371 and only incorrectly classified it as b >50k only 21 instances out of 371, whereas in identifying patients with income exceeding, it classified 57 correctly and 61 were classified incorrectly.
5. From the visualization of the tree, it is evident that the tree continues to gain more information from the 'marital.status' attribute and further branching was made on this attribute. marital.status'in Experiment2 provided more information on which

the data is further split. The tree has changed shape due to binary splitting on nominal variables. However, this parameter increased the number of correctly classified instances.

Experiment 3

The parameters are set to the following as shown below. The changes from previous experiment are highlighted. confidenceFactor is set to 0.5 for less pruning (smaller values incur more pruning).

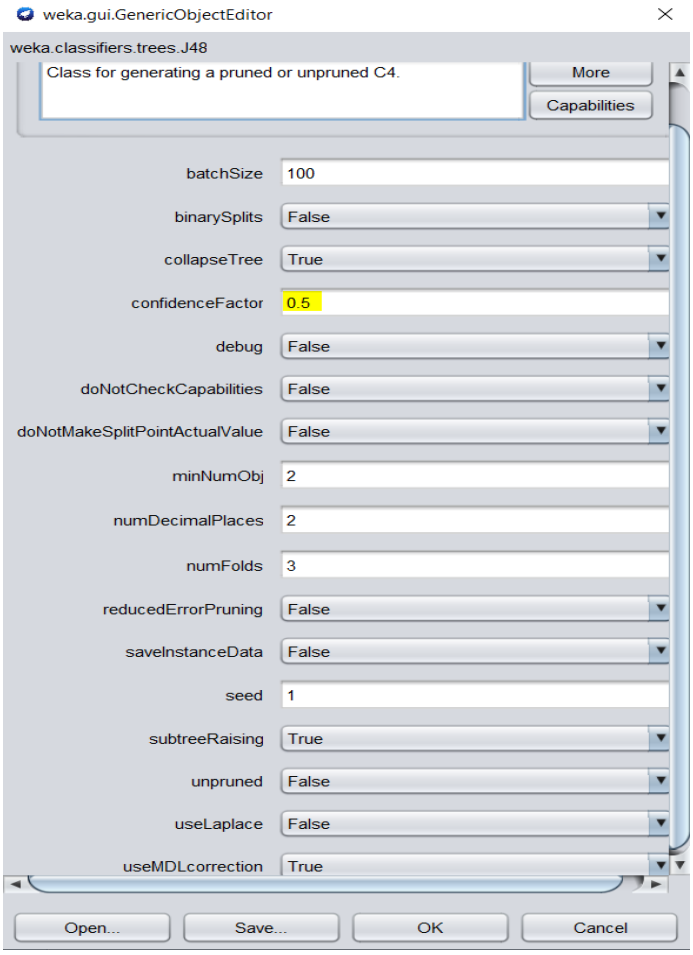


Figure 53 Experiment 3 -J48 parameters

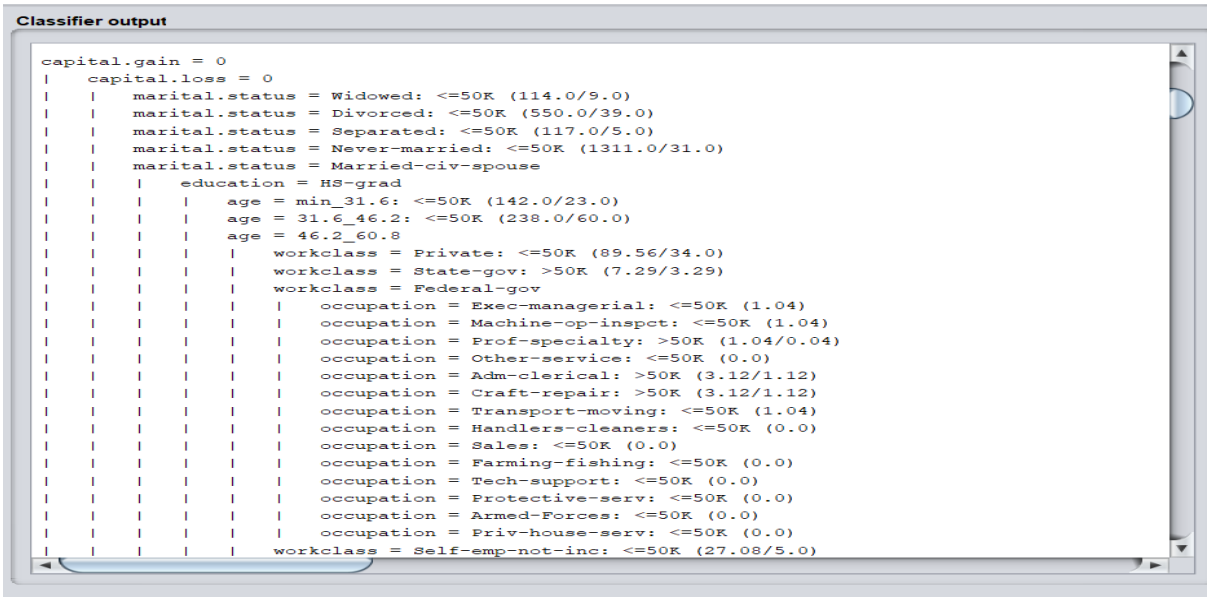


Figure 54Experiment 3-Model(small part)

Results with confidence values, confusion matrix, tree, rules

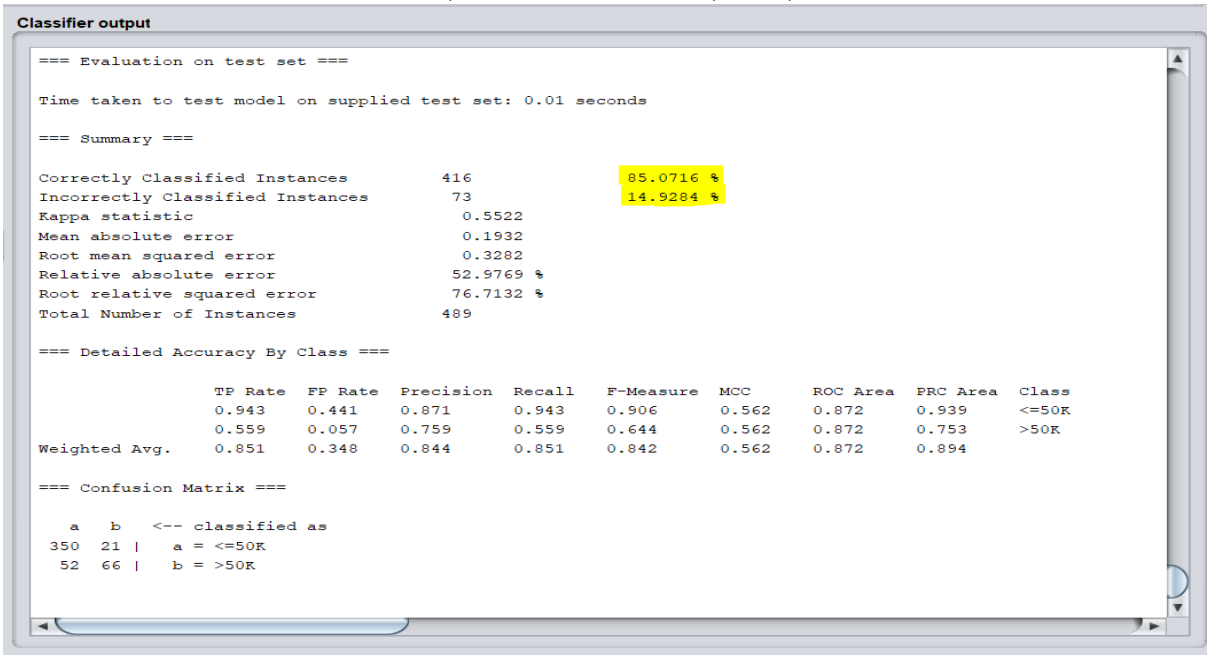


Figure 55Experiment 3 Result J48

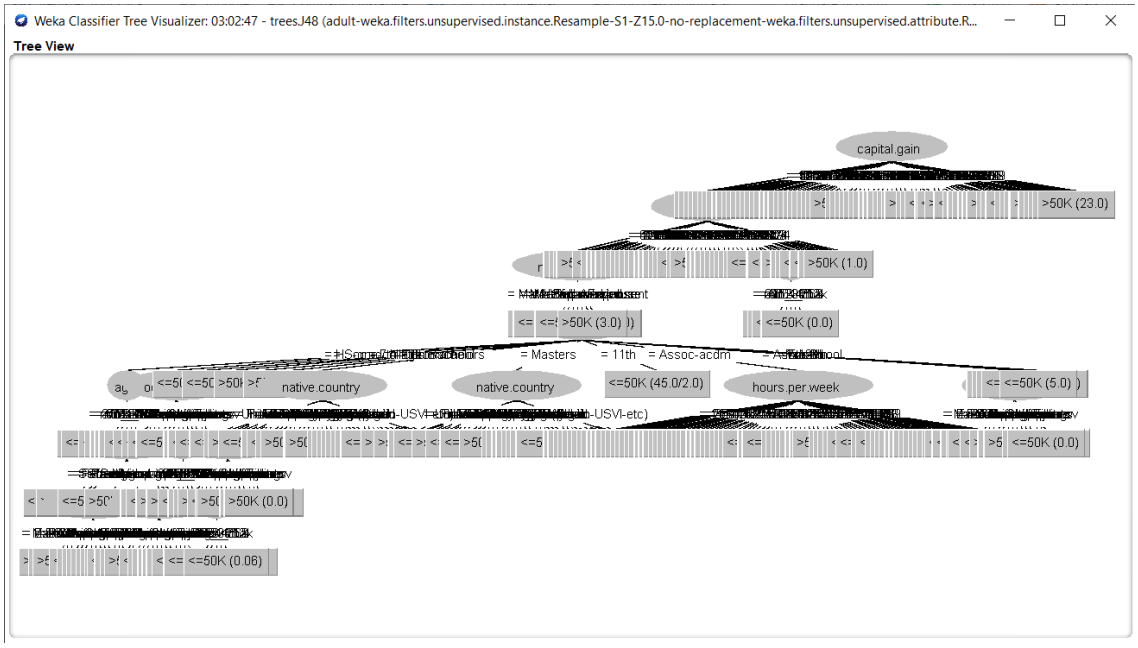


Figure 56 Experiment 3-J48 Tree

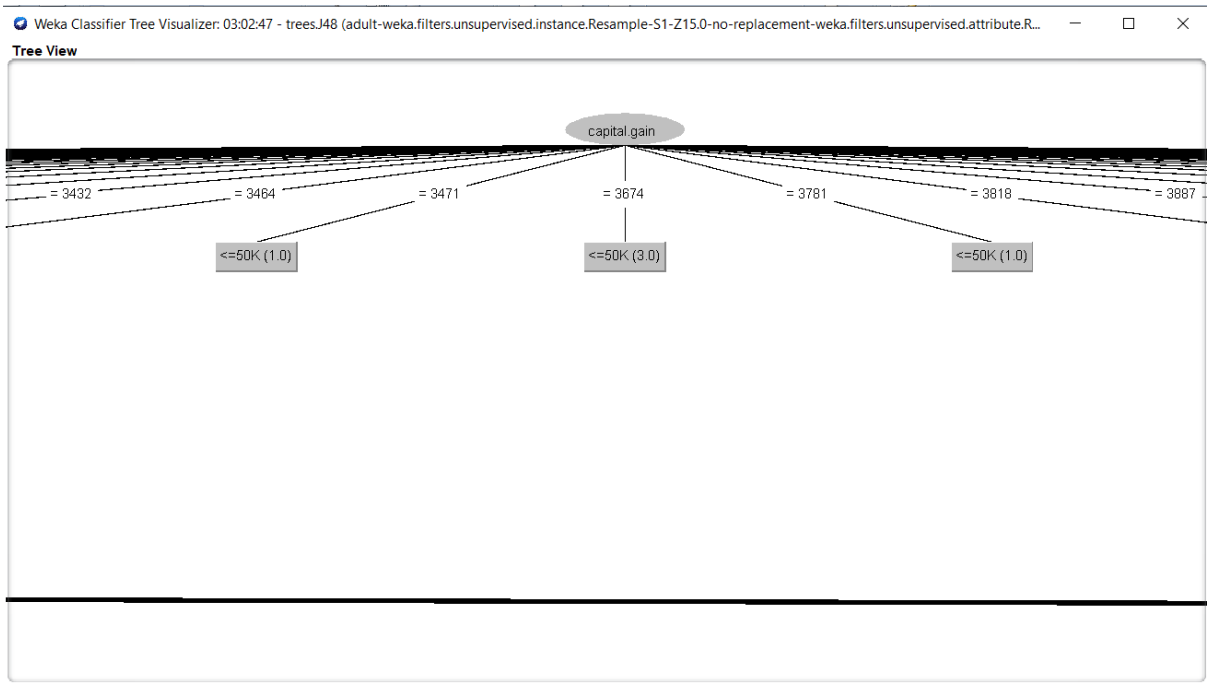


Figure 57 Experiment 3 J48 Tree

Findings

1. In this experiment, the parameters are varied to tune the model and it increased the model accuracy to reach 85.0716%. However Root Mean Squared Error(RMSE) decreased to 0.3282 from 0.3292.
2. The parameters that had a effect on this are:
    - unpruned is set as False, because pruned decision trees prevent the overfitting of the training data. (Dr.Abubakr Siddig- Decision Trees-2 lecture , 2020).

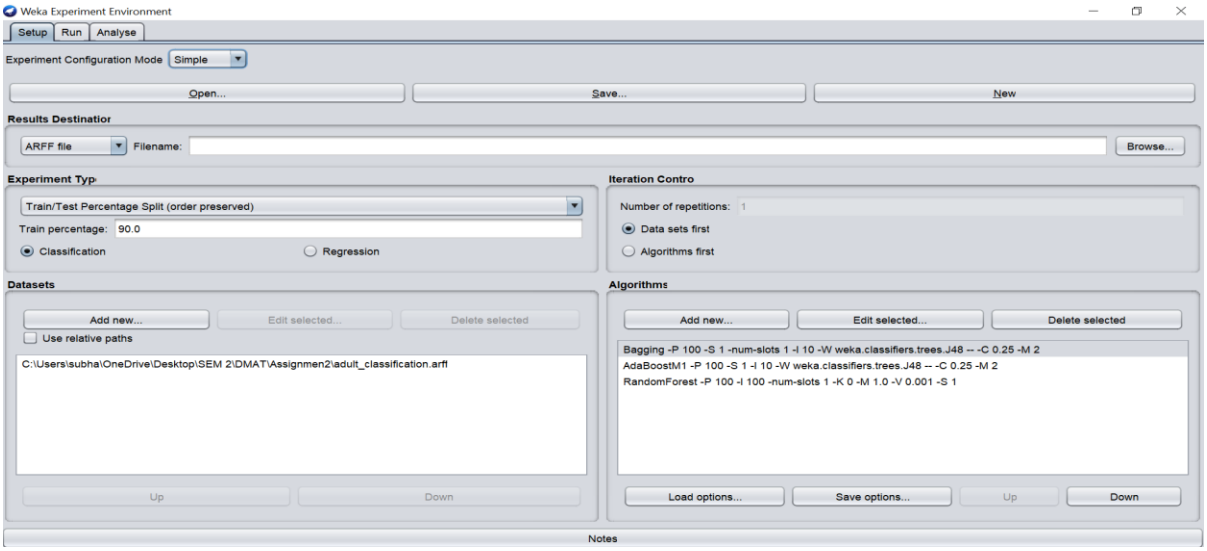
- confidenceFactor is set to 0.05 for pruning (smaller values incur more pruning)

3. From, the confusion matrix it is understood that the model is getting better at identifying the people with income exceeding 50k also as compared to patients with risk of readmission (b=1)– correctly classified 223 out of 539 and only 316 instances out of 539 incorrectly classified it as 1, whereas in identifying patients with risk of readmission (b=1), it classified 814 instances correctly and 134 were classified incorrectly.
4. From the visualization of the tree, it is evident that the tree began from the 'capital.gain' attribute. Further splitting was based on this attribute.

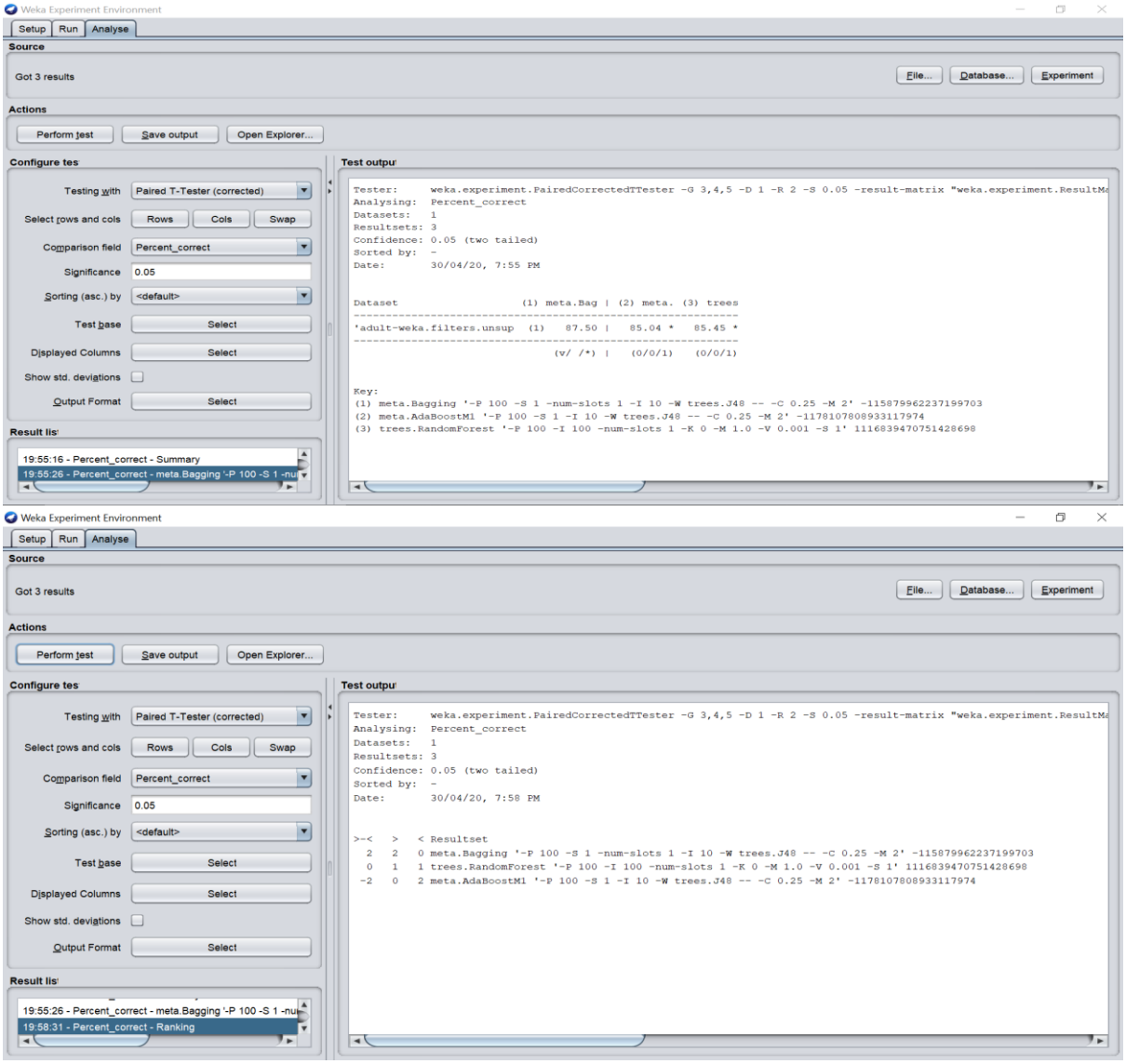
Ensemble learning

Ensemble learning can provide better predictions compared to individual predictors. This is due to "wisdom of crowd" (Dr.Abubakr Siddig- Ensemble Learninglecture , 2020).

These techniques are used to alter variance and bias so that model can perform better.
For the classification data, Bagging, Boosting and RandomForest algorithms are tested using Weka's Experimenter.



Result



Findings

1. The results show that Bagging worked much better than AdaBoost and RandomForest with this dataset. Bagging achieved an accuracy of 87.50% which is the highest accuracy achieved using this dataset. It won over the other 2 algorithms as shown by the ranking test base.
2. This can be due to the fact that Bagging works really well on unstable learners like decision trees and performs bias-variance of decomposition. (Dr.AbubakrSiddig- Ensemble Learning lecture , 2020).
3. AdaBoostM1 produced the least performance 85.04% because of overfitting which was not the case with Bagging. RandomForest performed slightly better than AdaBoostM1 with 85.45% accuracy.

5. Classification: MLP or a similar advanced technique from Weka – 15%

Classification using Multi-LayerPerceptron-MLP

MLP is an advanced classification technique that uses feed-forward network and back propagation algorithm. (Dr.AbubakrSiddig- Neural Network and Deep Learninglecture , 2020).

Experiments performed using Neural Network based algorithms are comparatively slower than J48. Testing on huge datasets on a normal machine is difficult. Hence the dataset is resampled for performing MLP experiments. Resampling is done on de identified dataset; hence performance is not affected.

In order to understand Weka MLP better, a tutorial video by University of Waikato was studied.(UK Government, 2020).Weka supports building network by hand or using heuristics. The network parameters can be changed during training process.Weka uses sigmoid function.Numerical attributes are not converted to Nominal as MLP doesn't require that. All of the experiments are performed on the test set- adult_MLP_testSet.arff.

Experiment 1

For experiment 1, the default settings of MLP is used as showed below. Hyper parameters are not varied here. It helps us to understand how the algorithm works generally for the dataset.

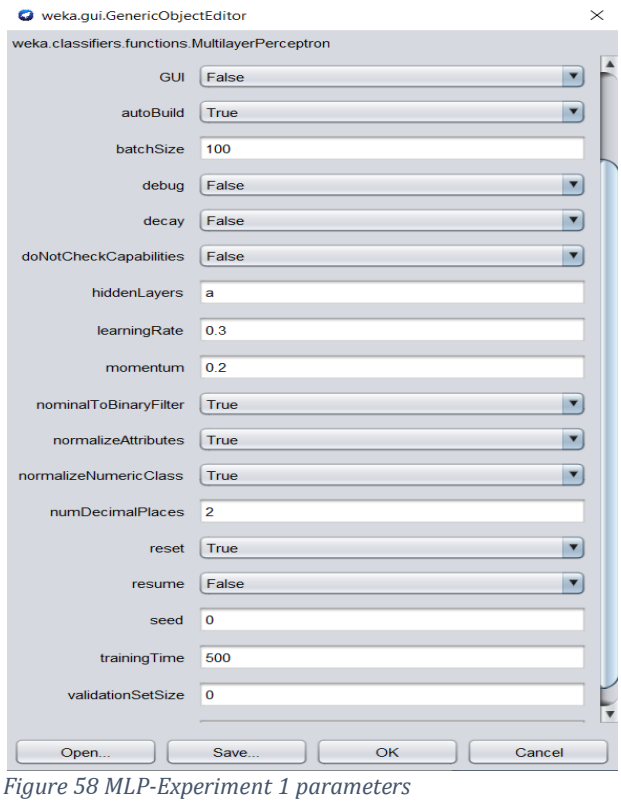


Figure 58 MLP-Experiment 1 parameters

Result

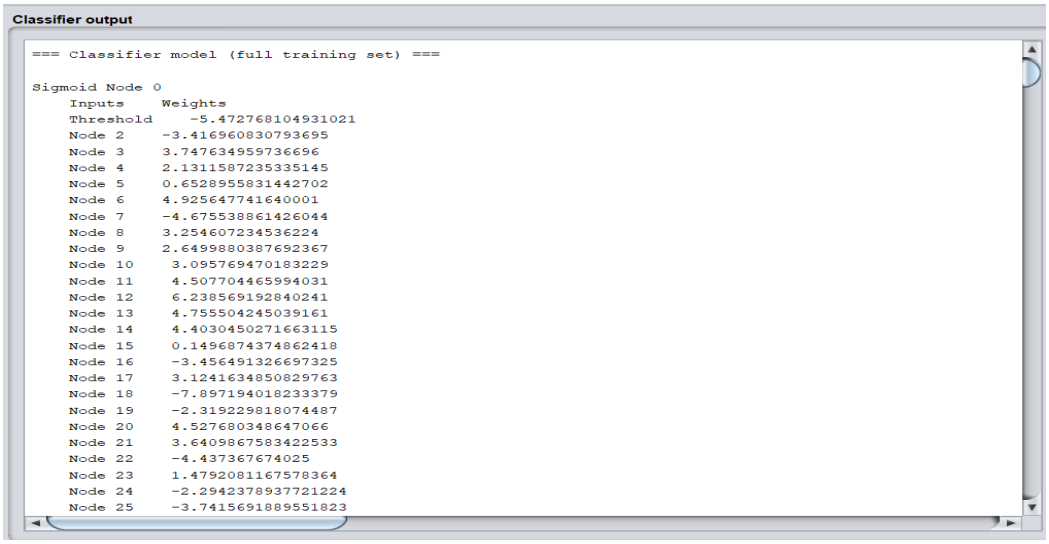


Figure 59 MLP-Experiment 1 Result

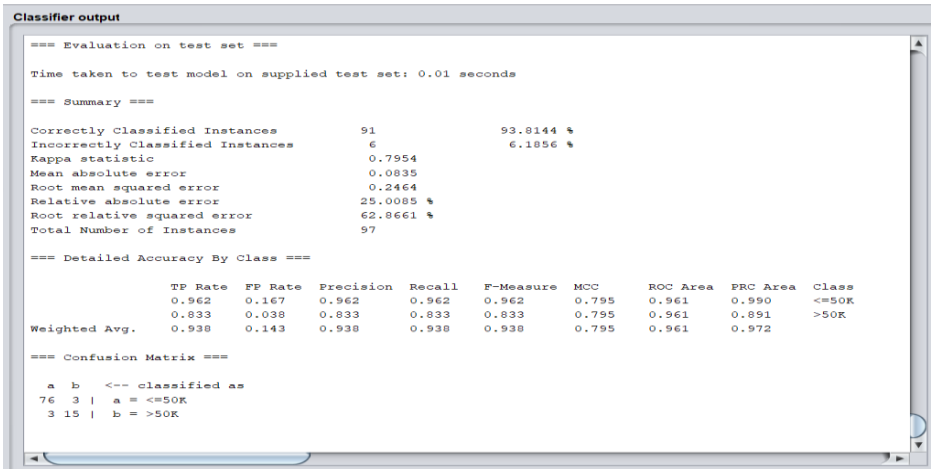


Figure 60 MLP-Experiment 1-test set Results

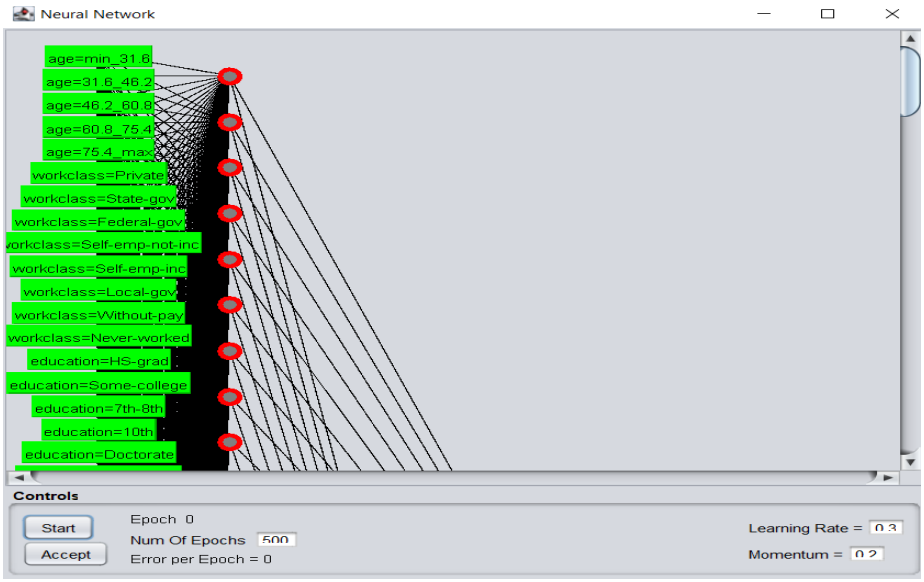


Figure 61 MLP-Experiment 1-Neural Network(Top view)

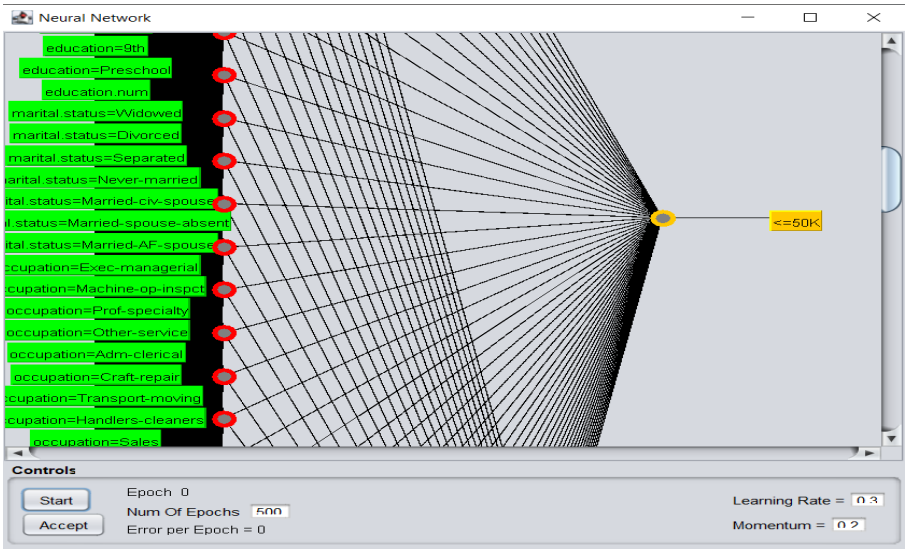


Figure 62 –MLP-Experiment 1-Neural Network<=50k (Middle view)

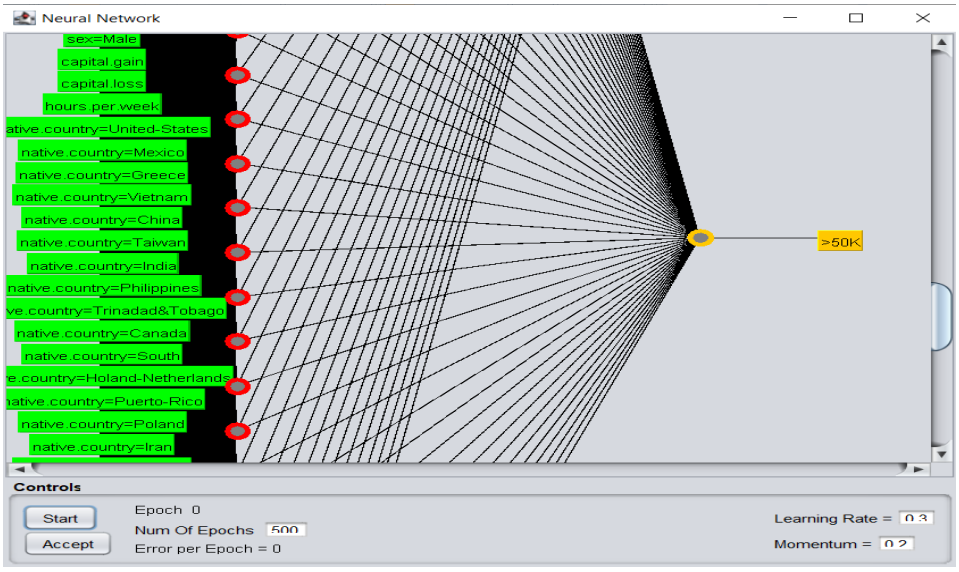


Figure 63MLP-Experiment 1-Neural Network >50k (Middle view)

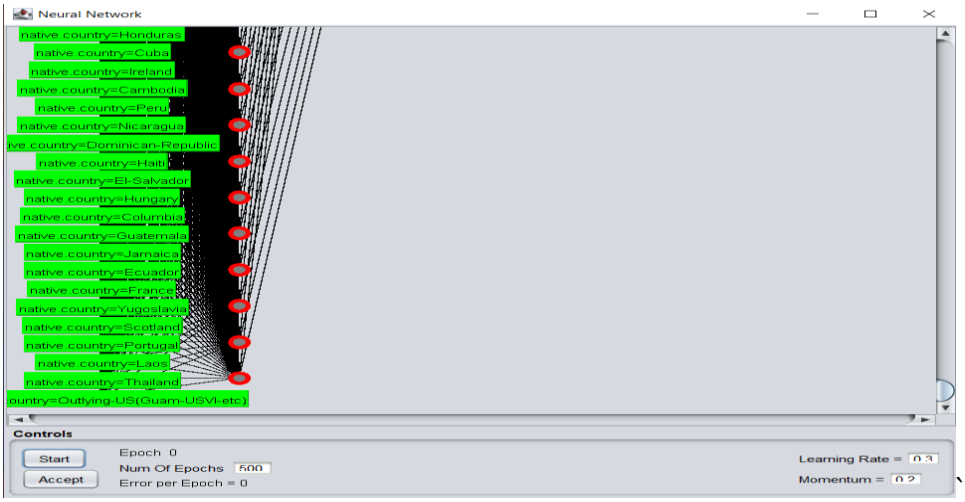


Figure 64 MLP-Experiment 1-Neural Network(End view)

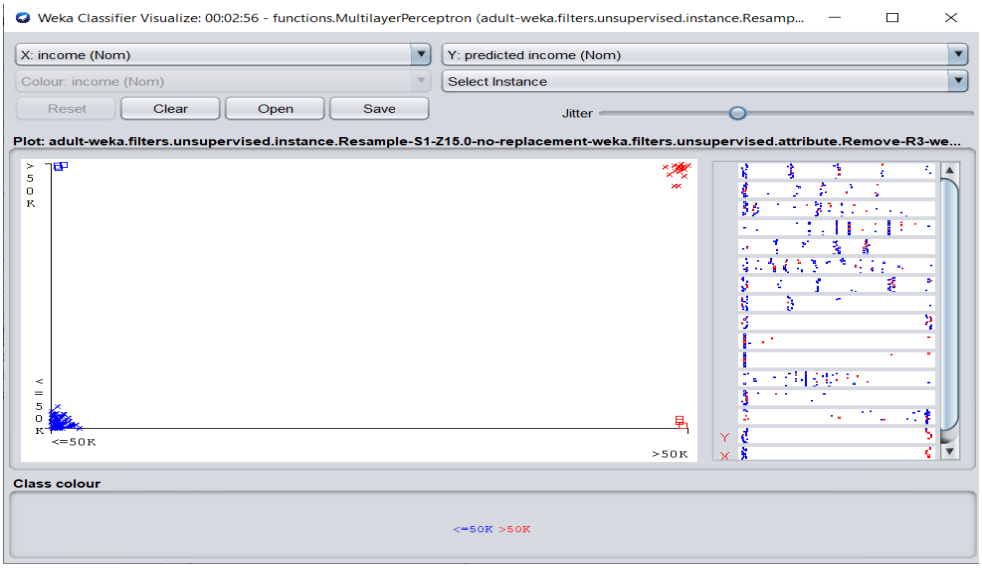


Figure 65 Classifier income Vs predicted income

Findings

1. With hidden layer= "a", (attributes + classes)/2=(14+1)/2 layers are used while developing the neural network. It produced an accuracy of 80% and mean absolute error of 0.1964.
2. Learning rate used to generate this is 0.5. Number of epochs is 500. The model was not varied during its creation and it's generate using heuristics.
3. The model seems to have performed better when we see the confusion matrix, 81 instances are correctly classified as people with less than or equal to 50k income and none of instances are incorrectly classified as people exceeding 50k. This is a really good sign. However, 22 people are incorrectly classified as people with income less than or equal to 50k and only 7 instances are correctly classified as people exceeding 50k. This means the model is better at identifying people with income less than or equal to 50k.
4. Error is minimized iteratively by using steepest descent and training is continued till error reduced.

Experiment 2

In experiment 2, the hidden layers are varied. hiddenLayers is set to 2,3, training time is set to 1000 and seed is set to 20.

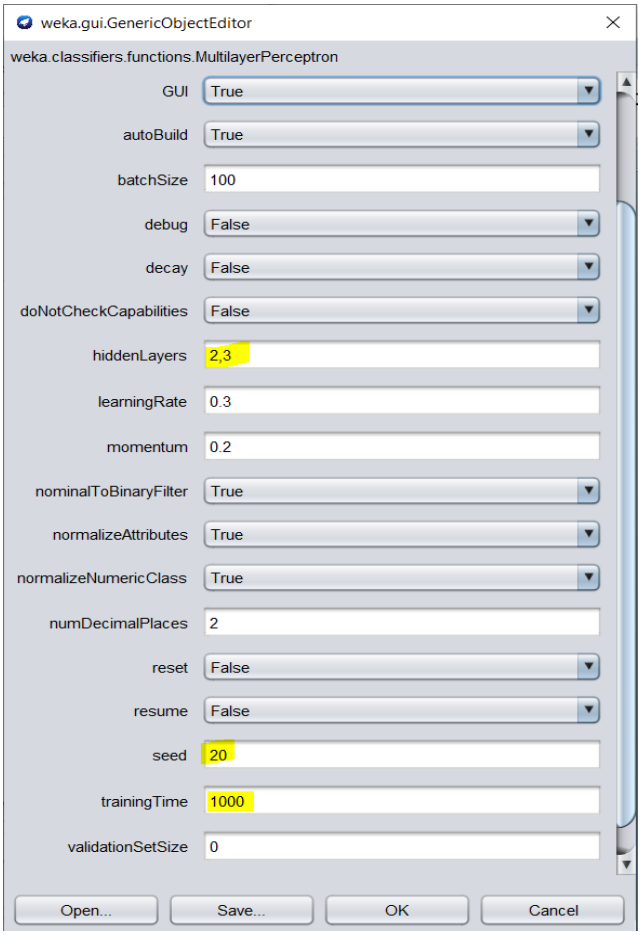


Figure 66 MLP-Experiment 2 parameters

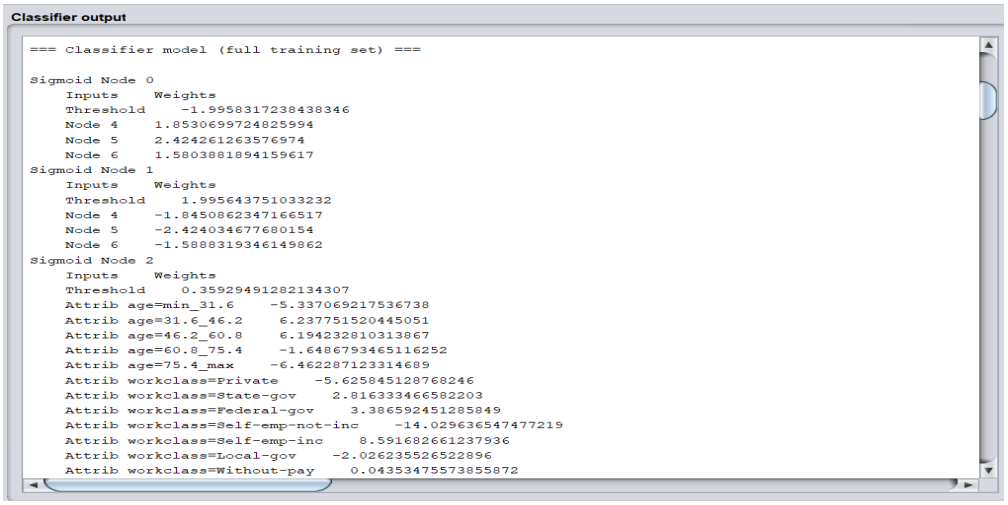


Figure 67 MLP-Experiment 2-Training set Results

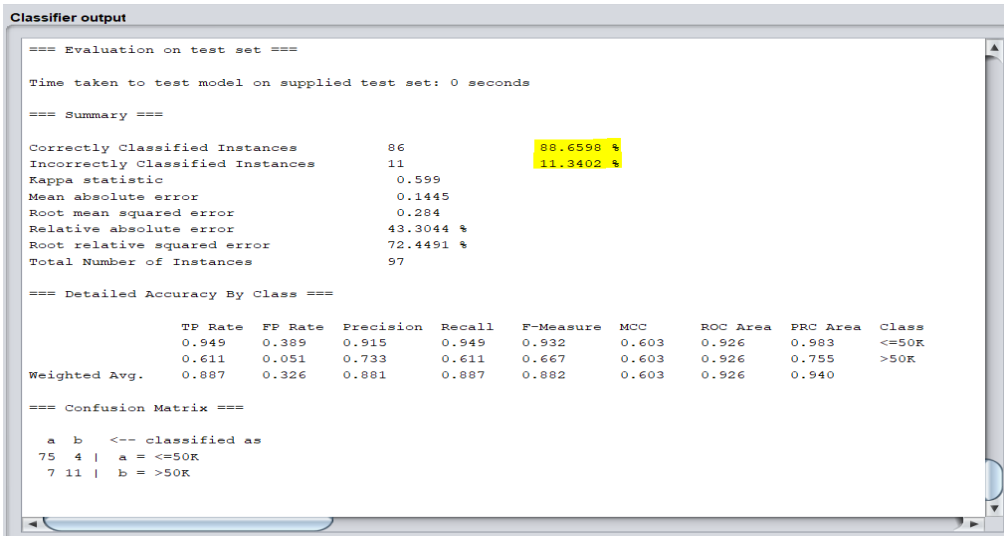


Figure 68 MLP-Experiment 2-Test set results

Figure 69 MLP-Experiment 2-Neural Network(Top view)

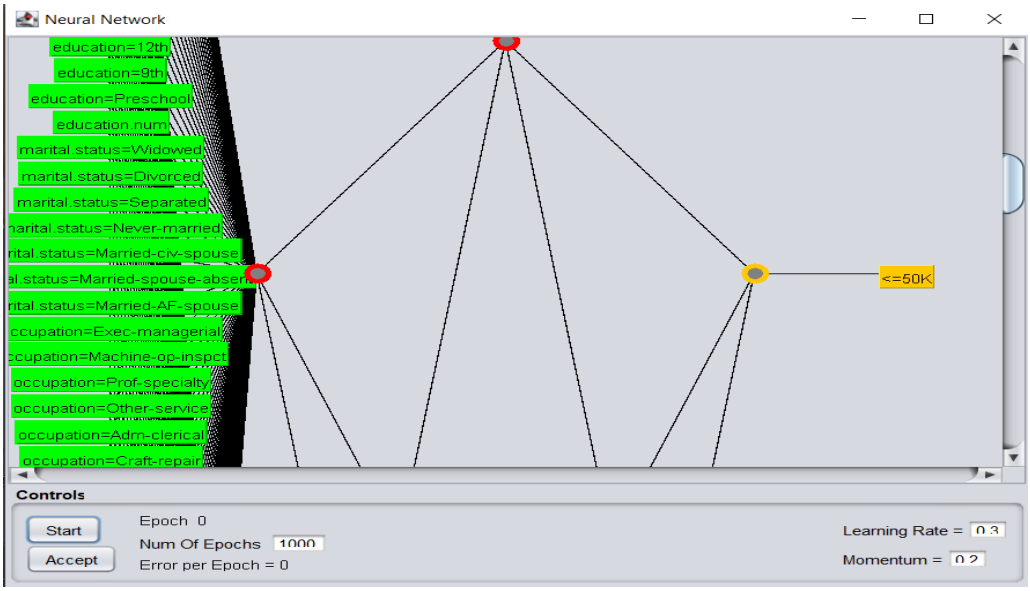


Figure 70MLP-Experiment 2-Neural Network<=50k(Middle view)

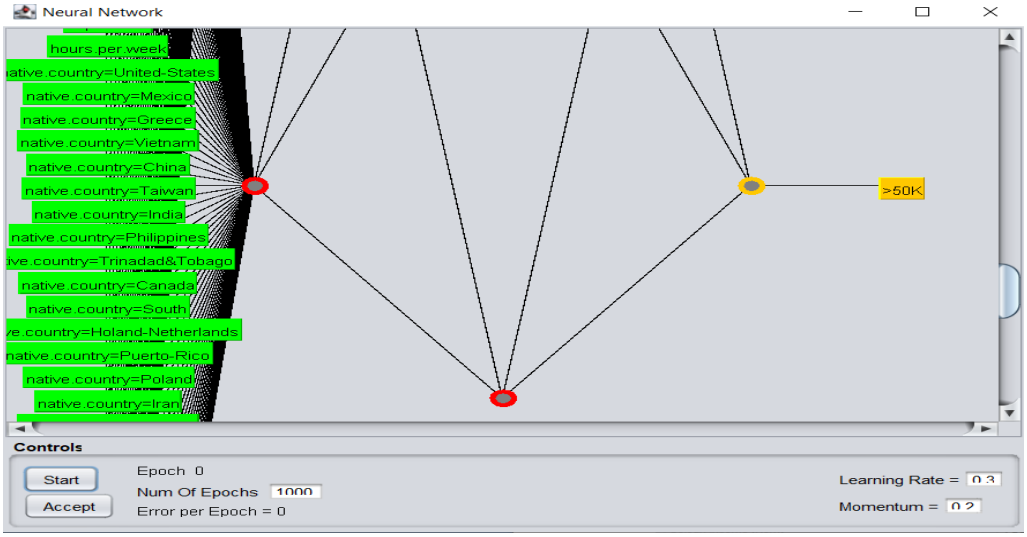


Figure 71MLP-Experiment 2-Neural Network >50k(Middleview)

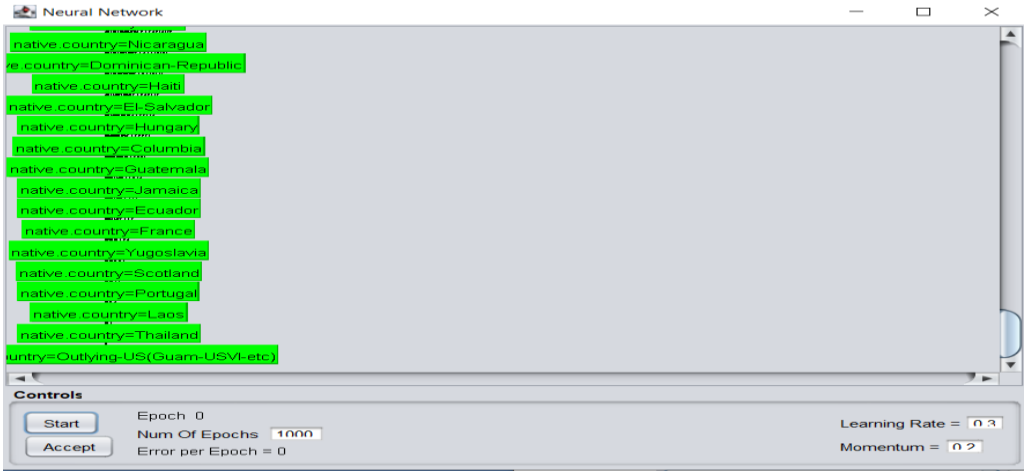


Figure 72MLP-Experiment 2-Neural Network(End view)

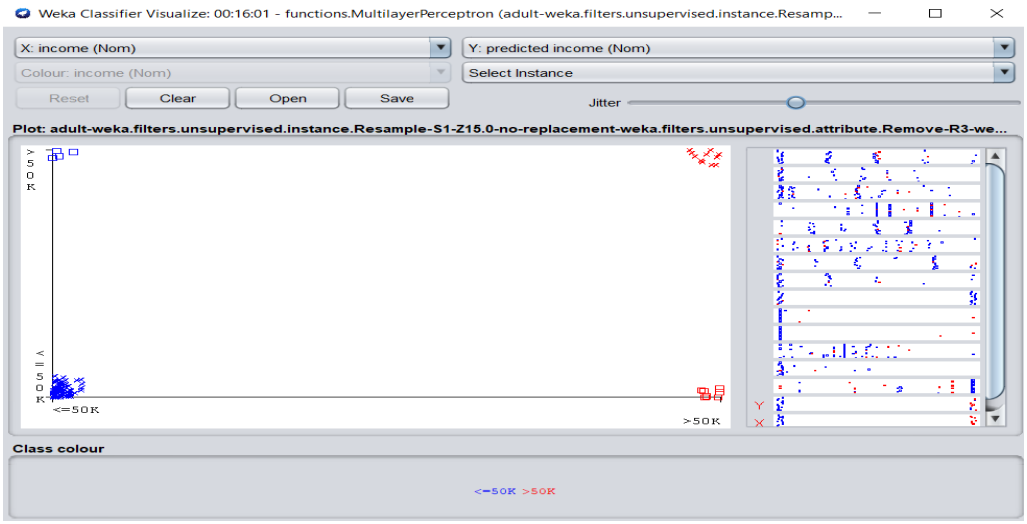


Figure 73 Classifier visualization with income on X axis and predicted income on Y axis

Findings

1. Accuracy is reduced to 88.8%. Mean absolute error also increased to 0.1444 from 0.08. This can be due to the reduction of number of hiddenLayers and a lightweight MLP with less layers.
2. A 2-layer, 3 unit network is created with number of epochs as 1000.Ideally,training should stop when validation error is increased to reduce loss.
3. Increasing training time to 1000, may have also caused reduced performance as
5. The confusion matrix shows that 75 instances are correctly classified as people with less than or equal to 50k income and 4 instances are incorrectly classified as people exceeding 50k. This is a really good sign. However, 7 people are incorrectly classified as people with income less than or equal to 50k and only 11 instances are correctly classified as people exceeding 50k. This means the model is better at identifying people with income less than or equal to 50k.

Experiment 3

In third experiment, hidden layer is set to 'i'- i.e, number of attributes. Seed is set to 10 and training time remains as 1000.

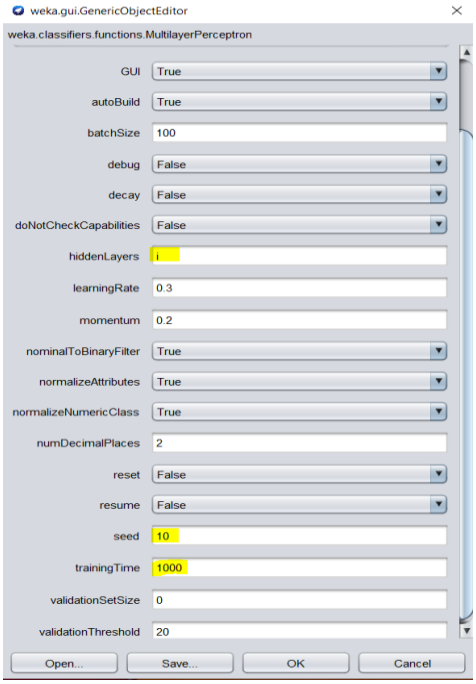


Figure 74 MLP-Experiment 3 parameters

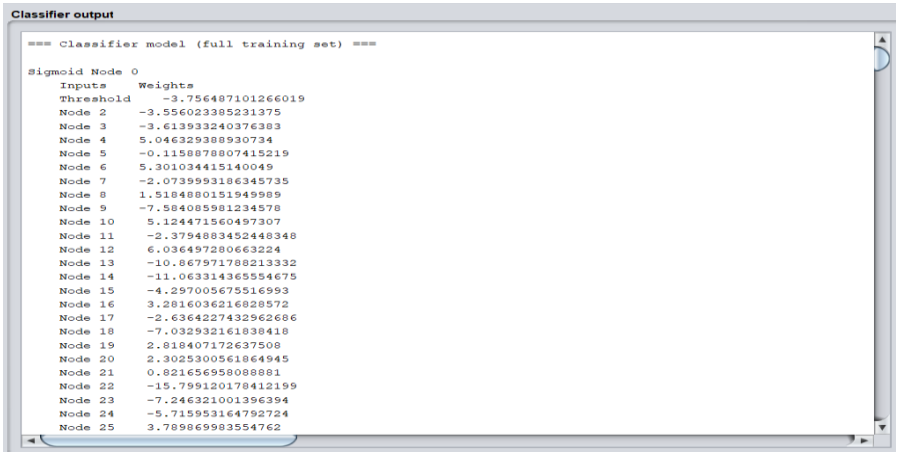


Figure 75 MLP-Experiment 3 Training set Results

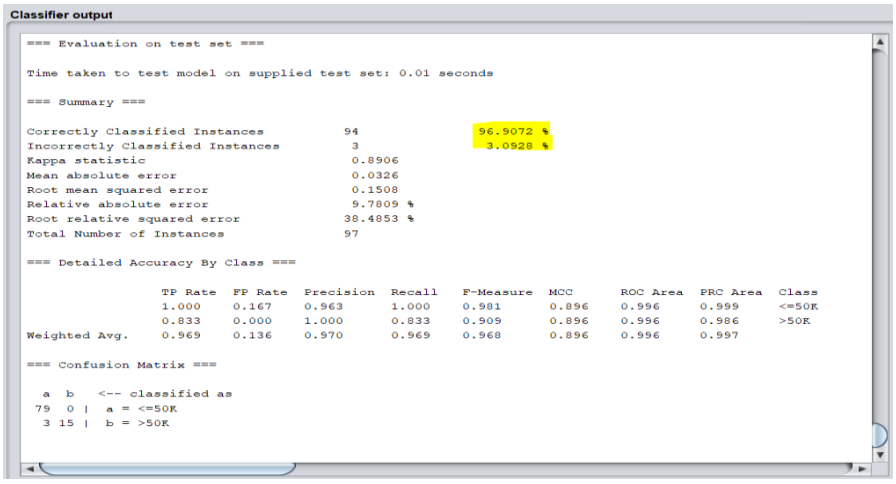


Figure 76 MLP-Experiment 3-Test set Results

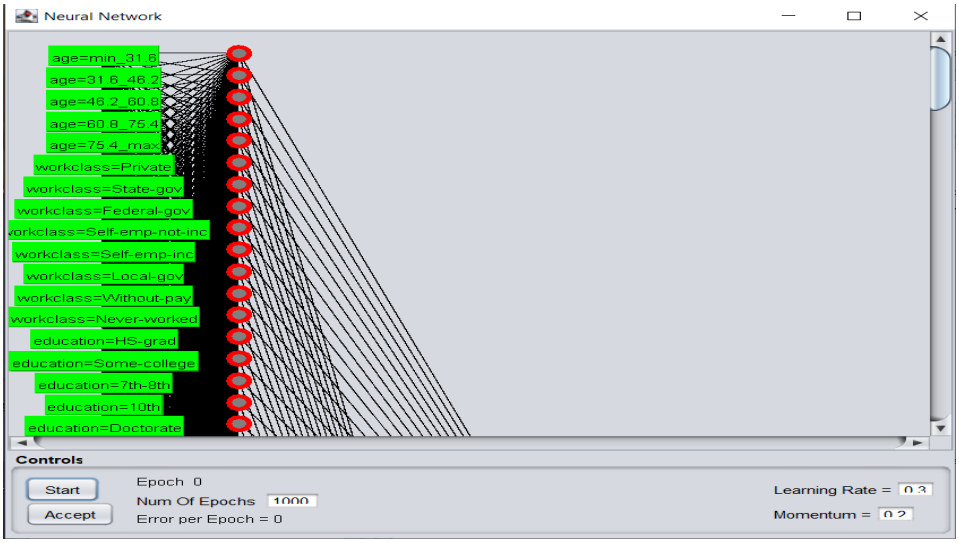


Figure 77 MLP-Experiment 3-Neural Networks(Top view)

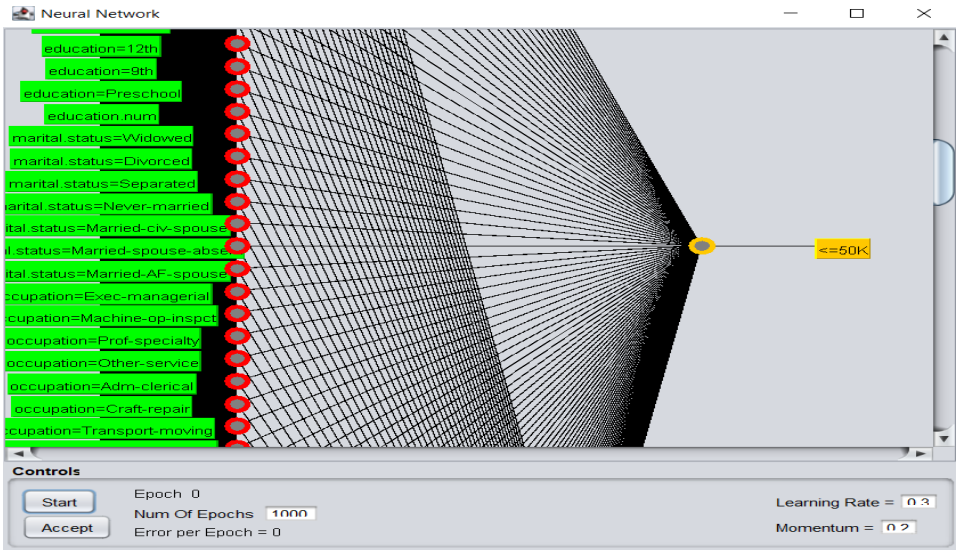


Figure 78 MLP-Experiment 3- Neural Network<=50k(Middle view)

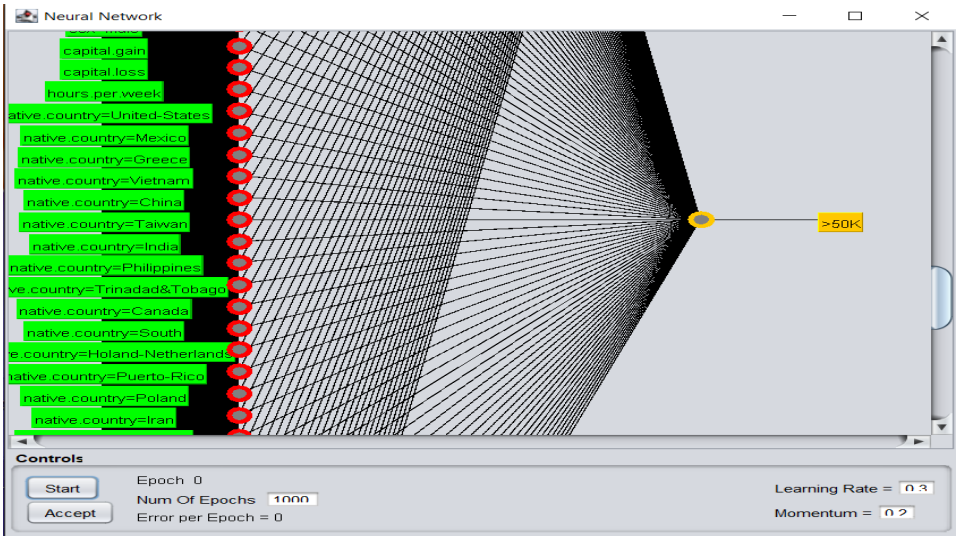


Figure 79MLP-Experiment 3- Neural Network >50k(Middle view)

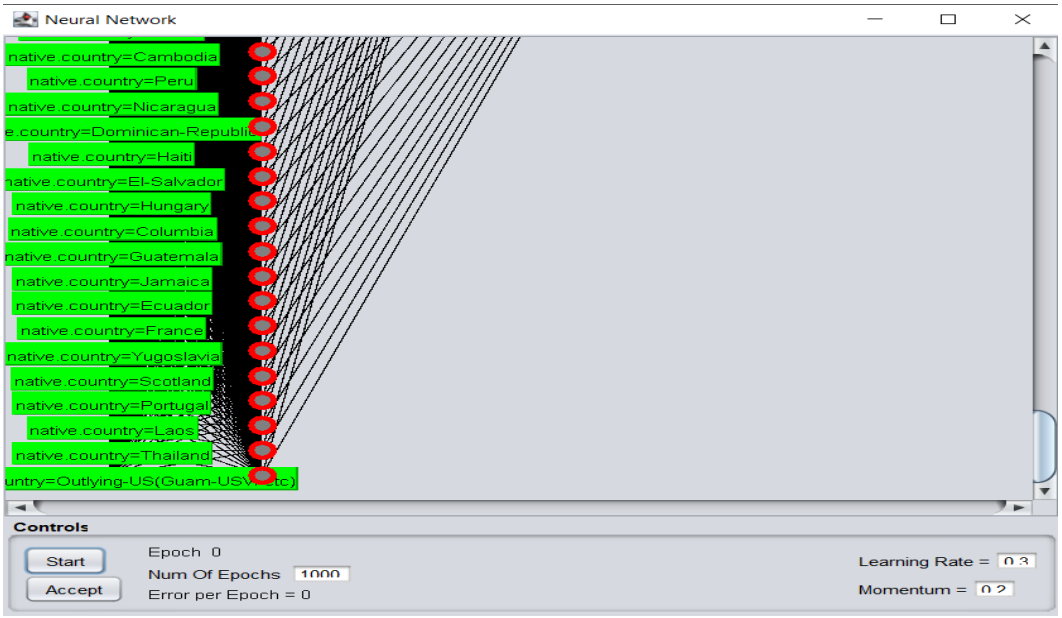


Figure 80 MLP-Experiment 3-Neural Network(End view)

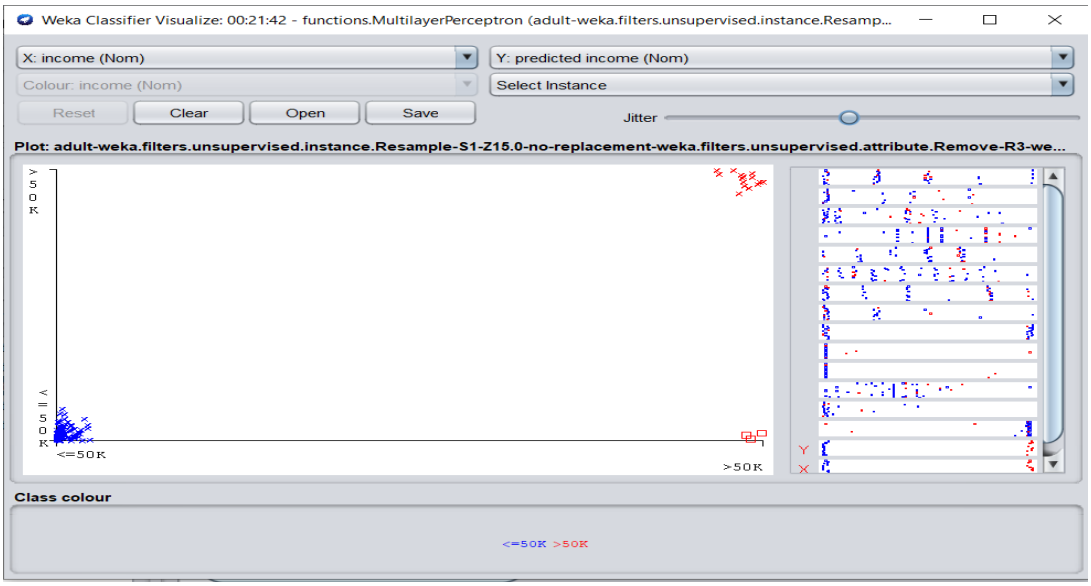


Figure 81 Classifier with income Vs predicted income

Findings

1. Accuracy is dramatically increased to 96.9072 and mean absolute error is just 0.032. With hidden layer =i, we have 14 layers.
2. The confusion matrix shows that 79 instances are correctly classified as people with less than or equal to 50k income and none of the instances are incorrectly classified as people exceeding 50k. This is a really good sign. However, 3 people are incorrectly classified as people with income less than or equal to 50k and 15 instances are correctly classified as people exceeding 50k.
3. MultiLayer Perceptron seems to be working really well with this dataset which has numerical attributes. It is comparatively slower, but there are less errors in classification which is really good for adult census dataset.

6. Clustering

- **Title**: K-Means Clustering with Adult Census Dataset
- **Data description**: The dataset used is Adult Census Dataset after resampling which is described in Part 1.
- **Objective:**

1. To partition the data to good quality clusters from Adult Census Dataset using K-Means Clustering with predefined number of clusters.
2. To understand useful patterns from the clusters generated by K-Means

- **Summary of Findings**

Clustering: K-Means or DBSCAN – 10%

Summary of Findings

1. Preprocessing

1. Preprocessing steps used in clustering is not the same as the steps used in the classification problem. Hence preprocessing steps are performed again on the original resampled dataset adult_resampled.Arff
2. The dataset contains 14 attributes, thus removing or selecting certain features based on their importance and correlation, increases the ease for processing, improves accuracy and overall provides better results.
3. Binning may improve the accuracy of predictive models by reducing the non-linearity or noise and it is also useful for certain classifiers..In the dataset, age has 67 distinct values which can be discretized to 4 bins.
4. MissingValuesareremovedusingReplaceMissingValuesfiltersimilar to what is done in PartA.
5. Preprocessed file is saved as dataset_clustering.arff

2. K Means Clusering

- Three experiments were performed using K-Means clustering on the dataset dataset_clustering. arff in Weka Explorer.Classes to Cluster Evaluation was used in all three of the experiments, to evaluate the performance of eachcluster.
- In Experiment 1, K-Means Algorithm with Euclidean distance measure, seed set to 10 and numClusters=2, 2 clusters are generated. Random initial points was used.Classes To Cluster evaluation is used (cs.ccsu.edu, nd), and it is identified that only 2868 classes with income less than or equal to 50k is assigned to cluster 0 and 851 instances is assigned to cluster 1. However, 459 instances of people with income exceeding 50k is assigned to cluster 0 and 706 is assigned to cluster 1. 26.8223% of instances were incorrectly clustered. Thus this has to be improved.From the visualization of cluster with colour as 'income', cluster 1 is dominated by patients with income less than 50k (851 instances), but it almost roughly equal in the visualization compared to the 706 instances of patients with income exceeding 50k and cluster 0 is clearly dominated by patients with income less than or equal to 50k, however this is slightly difficult to identify from the cluster visualization due to roughly same distribution.

- In Experiment 2, K-Means Algorithm with Euclidean distance measure, seed set to 20 and numClusters=3, 3 clusters are generated. Random initial points was used. The algorithm iterated 11 times.
- Cluster 0- A male with Private workclass and Bachelors education, specialising in Prof-specialty with capital gain of 2361.5427.
  Cluster 1- A male with Private workclass and 10th education, specialising in other service with capital gain of 305.3263.
  Cluster 2- A male with Private workclass and HS-Grad, specialising in craft-repair with capital gain of 536.6745.
- From the cluster visualization, most of the people with high capital gain and income exceeding 50k belong to cluster 0.
- In the experiment 3, canopy initialization method was used instead of K-Means clustering which slightly improved the performance with radius T2 radius: 1.583 and T1 radius: 1.979.
- With canopy preclustering, final centroid of cluster 0 is HS- Grad educated person and in cluster 1 centroid education is Bachelor's.

K-Means is a relatively efficient algorithm that often terminates at local optimum and handles numerical data. For categorical data, it replaces means of clusters with modes. Number of clusters need to be specified earlier. For a mixture of categorical and numerical data, k-prototype method is used. (Dr.AbubakrSiddig- Clustering lecture , 2020)

Preprocessing

The preprocessing steps used in Clustering slightly differ from the ones used in Classification because the objective of clustering is different from classification. The dataset that is used for clustering is adult_resampled.arff with 4884 instances.

All of the preprocessing techniques used in Classification part except data type conversion is performed for clustering and the dataset is stored as dataset_clustering.arff.

1. Selecting or filtering the attributes

Feature selection is an extremely important step in clustering also. It not only reduces the dimensions, but also improves the quality of the clusters to a great extent. The dataset contains 14 attributes, thus removing or selecting certain features based on their importance and correlation, increases the ease for processing, improves accuracy and overall provides better results. There are no features with more than 40% noise or missing values, which is the general rule of thumb to remove features, hence features need not be removed based on that criteria.

For clustering, CorrelationAttributeEval can be used to select features. It evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Correlation is a useful statistic measure, especially in this dataset with 50

attributes. Ranker is very useful because it gives the rank of the features. Features can be selected depending upon this information.

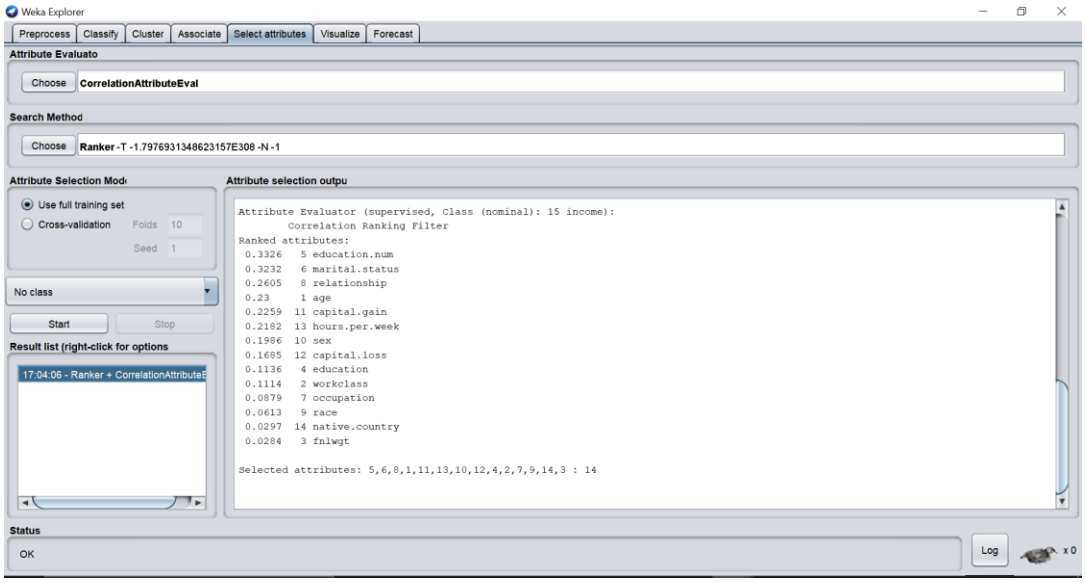


Figure 82CorrelationAttributeEval

'fnlwgt' is ranked the least and it is removed.

2. Discretization (Binning)

Binning may improve the accuracy of predictive models by reducing the non-linearity or noise and it is also useful for certain classifiers. After binning, outliers, missing or invalid values can be easily identified. This can be performed using Weka's Discretize filter. In the dataset, age has 67 distinct values which can be discretized to 4 bins.
After discretization, it is useful to replace the bin range names with more readable and comprehensible names using a text editor and it is done as shown below.

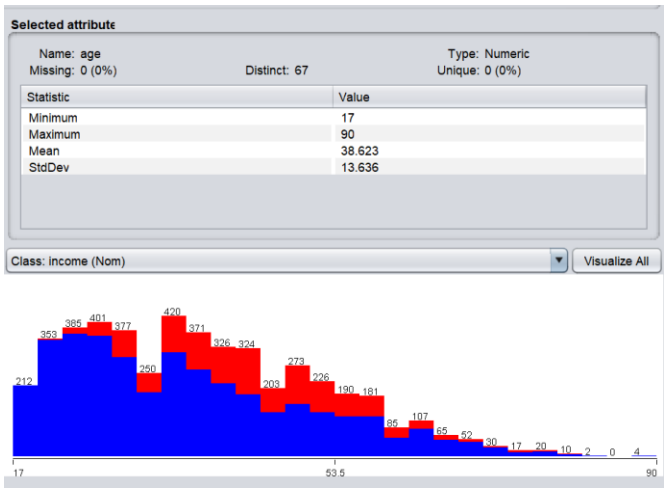


Figure 83 Before discretization

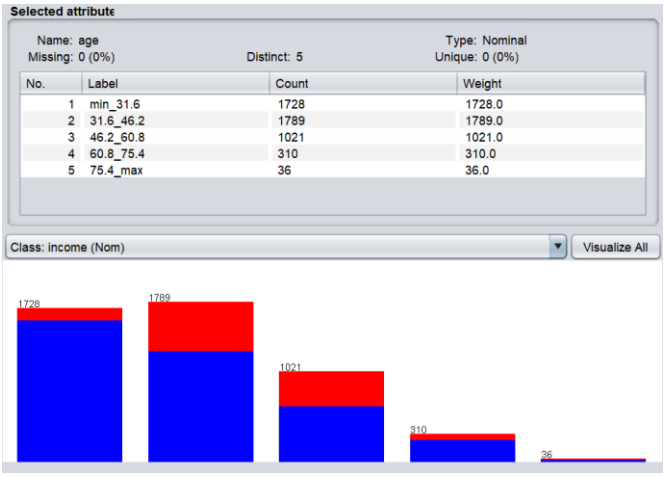


Figure 84 After discretization

3. Missing Values

Data can be missing due to a variety of reasons- hesitance of the respondents to provide complete information, malfunctioning of equipments, errors when entering the data in to the database, sudden changes etc etc (Dr.AbubakrSiddig- Datasets,EDA and altering data structure lecture , 2020). A small amount of missing value is almost unavoidable in large datasets. However, a significant percentage of missing values can be problematic. There are only three attributes with missing values; 'workclass'-6% missing values, 'occupation' -6% missing values, Missing values of these attributes are removed by replacing with modes and means using ReplaceMissingValues Filter.

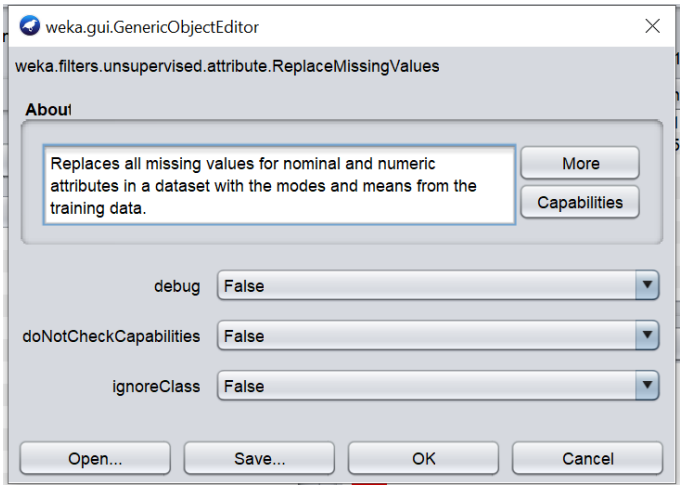


Figure 85ReplaceMissingValues

Experiment 1

For the first experiment, default settings of SimpleK-Means will be used.

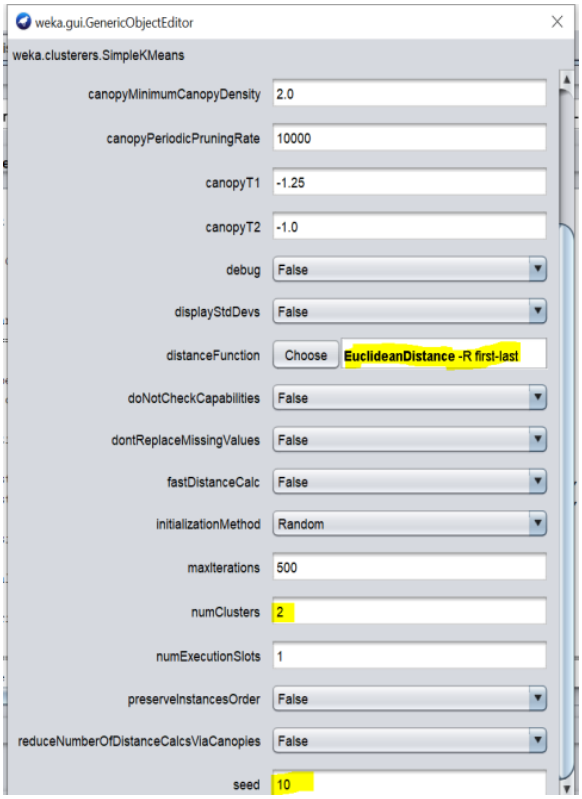


Figure 86 K-Means Experiment 1

The distance function used is Euclidean distance is straight-line distance between two points in Euclidean space. Euclidean distance here is used for the numerical attributes with indices as shown above. dontNormalize is set to False because we need these numerical attributes to be scaled, since that is not done beforehand. Seed is set to 10 and numClusters is set to 2 for generating 2 clusters.

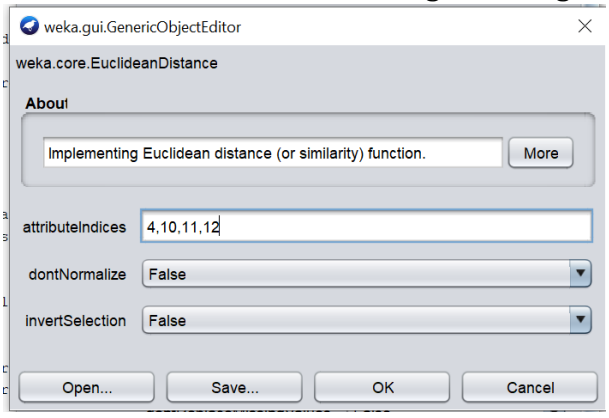


Figure 87 Euclidean distance

Result

Clusterer output

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 249.55194054965827

Initial starting points (random):

Cluster 0: 31.6_46.2,Private,HS-grad,9,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,0,40,United-States
Cluster 1: 46.2_60.8,Self-emp-inc,Bachelors,13,Married-civ-spouse,Sales,Husband,White,Male,0,0,60,United-States

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                               Full Data           0                1
                                         (4884.0)    (3327.0)         (1557.0)
=============================================================================================
age                                     31.6_46.2   min_31.6          31.6_46.2
workclass                                 Private     Private            Private
education                                 HS-grad     HS-grad          Bachelors
education.num                             10.0678      8.7099            12.9692
marital.status                 Married-civ-spouse Married-civ-spouse Married-civ-spouse
occupation                          Prof-specialty      Craft-repair     Prof-specialty
relationship                               Husband     Husband            Husband
race                                         White       White              White
sex                                           Male        Male               Male
capital.gain                            1070.2797    488.312          2313.8292
capital.loss                              92.9134     10.5416           268.9255
hours.per.week                             40.542     39.2771            43.2447
native.country                      United-States United-States      United-States



Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       3327 ( 68%)
1       1557 ( 32%)
```

Figure 88KMeans Experiment 1-Result

```
Class attribute: income
Classes to Clusters:

    0    1  <-- assigned to cluster
 2868  851 | <=50K
  459  706 | >50K

Cluster 0 <-- <=50K
Cluster 1 <-- >50K

Incorrectly clustered instances :        1310.0    26.8223 %
```

Figure 89KMeans -Experiment-1 Classes to Cluster evaluation

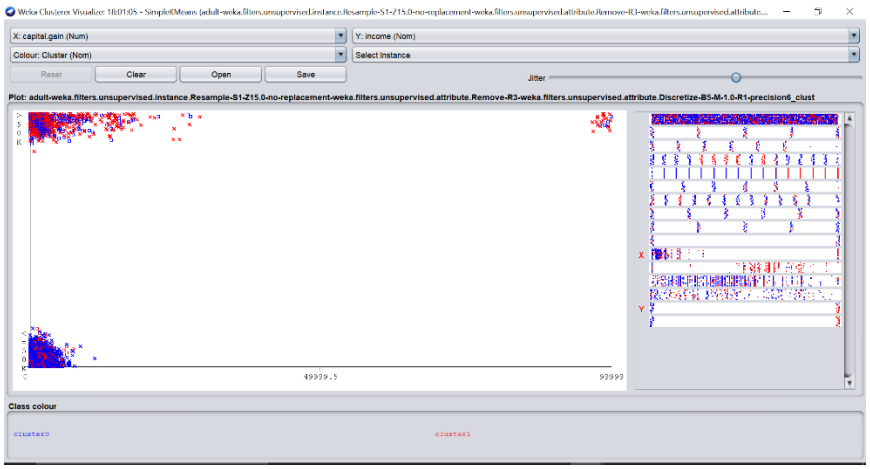


Figure 90 Cluster Visualizations with capital gain on X axis and income on Y axis

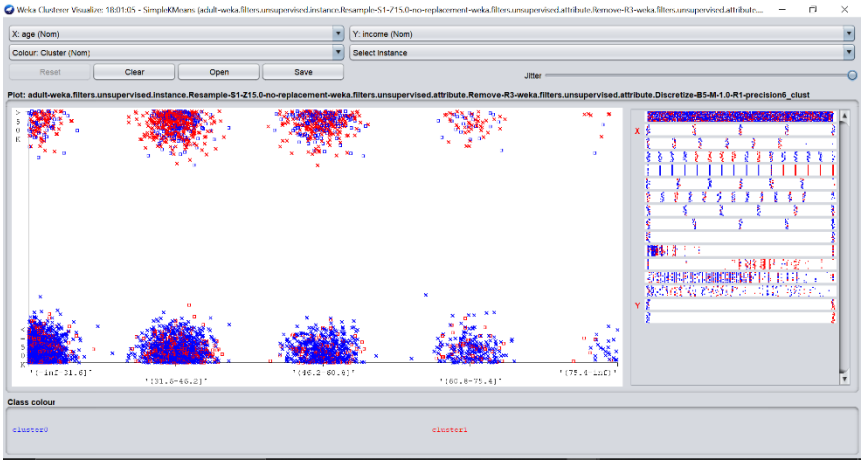


Figure 91Cluster : age Vs income

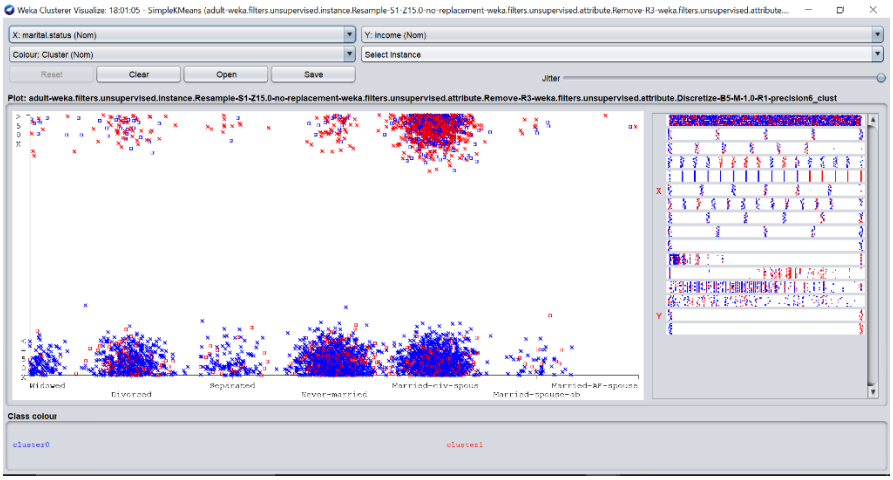


Figure 92 marital-status clusters

Findings

1. Using K-Means Algorithm with Euclidean distance measure, seed set to 10 and numClusters=2, 2 clusters are generated. Random initial points was used.

   Cluster0:31.6_46.2,Private,HS-grad,9,Married-civ-spouse, Prof-Specialty,Husband,White,Male,0,0,40,United-States
   Cluster1:46.2_60.8,Self-emp-inc,Bachelors,13,Married-civ-spouse,Sales,Husband,White,Male,0,0,60,United-States
2. Sum of Within cluster sum of squared errors: 249.55. This error have to be reduced in further experiments. A good clustering method will produce high quality clusters with high intra-class similarity(Dr.AbubakrSiddig- Clustering lecture , 2020).
3. Classes To Cluster evaluation is used (cs.ccsu.edu, nd), and it is identified that only 2868 classes with income less than or equal to 50k is assigned to cluster 0 and 851 instances is assigned to cluster 1. However, 459 instances of people with income exceeding 50k is assigned to cluster 0 and 706 is assigned to cluster 1. 26.8223% of instances were incorrectly clustered. Thus this has to be improved.
4. Cluster centroids are the mean vectors for each cluster. In the final cluster centroids, some of the factors that are evident are:

- Cluster 0- A male with Private workclass and HS Grad education, specialising in craft repair.
- Cluster 1- A male with Private workclass and Bachelor's education, specialising in Prof-Speciality.

5. From the visualization of cluster with colour as 'income', cluster 1 is dominated by patients with income less than 50k (851 instances), but it almost roughly equal in the visualization compared to the 706 instances of patients with income exceeding 50k and cluster 0 is clearly dominated by patients with income less than or equal to 50k, however this is slightly difficult to identify from the cluster visualization due to roughly same distribution.
6. As shown by the final centroid and also in the visualization it can be identified that cluster 0 is dominated by people with HS Gradeducationand cluster 1 is dominated by people with Bachelor's education.

Experiment 2

In experiment 2, the number of iterations is increased to 3000 and the number of clusters is also increased to 3 to see the performance. The seed is increased to 20 to see the difference in performance.

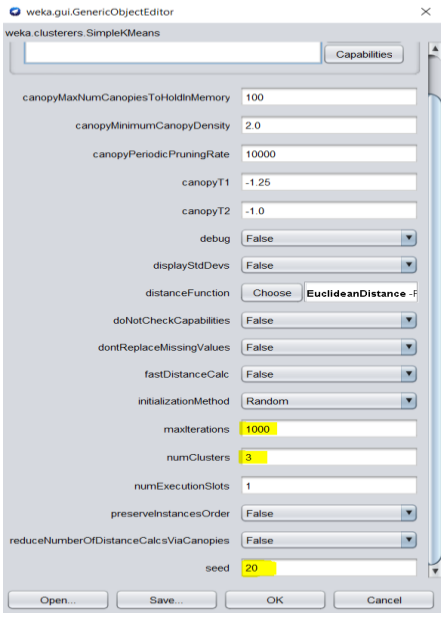


Figure 93 K-Means-Experiment 2 parameters

Result

Clusterer output

```
kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 210.68026763259763

Initial starting points (random):

Cluster 0: 31.6_46.2,Private,Masters,14,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,0,46,United-States
Cluster 1: 46.2_60.8,Private,7th-8th,4,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,0,24,United-States
Cluster 2: 31.6_46.2,Private,Some-college,10,Married-civ-spouse,Tech-support,Husband,White,Male,0,0,40,United-States

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                          Full Data           0                1                2
                                    (4884.0)    (1500.0)         (567.0)         (2817.0)
=============================================================================================================
age                                31.6_46.2   31.6_46.2        min_31.6        min_31.6
workclass                            Private     Private         Private         Private
education                            HS-grad   Bachelors            10th         HS-grad
education.num                        10.0678     13.0367          5.1852          9.4696
marital.status       Married-civ-spouse Married-civ-spouse Married-civ-spouse Married-civ-spouse
occupation             Prof-specialty Prof-specialty Other-service Craft-repair
relationship                    Husband     Husband         Husband         Husband
race                              White       White           White           White
sex                                Male        Male            Male            Male
capital.gain                   1070.2797   2361.5427        305.3263        536.6745
capital.loss                     92.9134     283.5073         45.3245          1.0043
hours.per.week                   40.542      42.8973         36.6455         40.0721
native.country           United-States United-States United-States United-States




Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      1500 ( 31%)
1       567 ( 12%)
2      2817 ( 58%)
```

Figure 94 K-Means Experiment-2 Result

```
Class attribute: income
Classes to Clusters:

    0    1    2  <-- assigned to cluster
  810  537 2372 | <=50K
  690   30  445 | >50K

Cluster 0 <-- >50K
Cluster 1 <-- No class
Cluster 2 <-- <=50K

Incorrectly clustered instances :        1822.0    37.3055 %
```

Figure 95 K-Means- Experiment 2- Classes to Clusters

Figure 96Cluster : marital-status Vs income

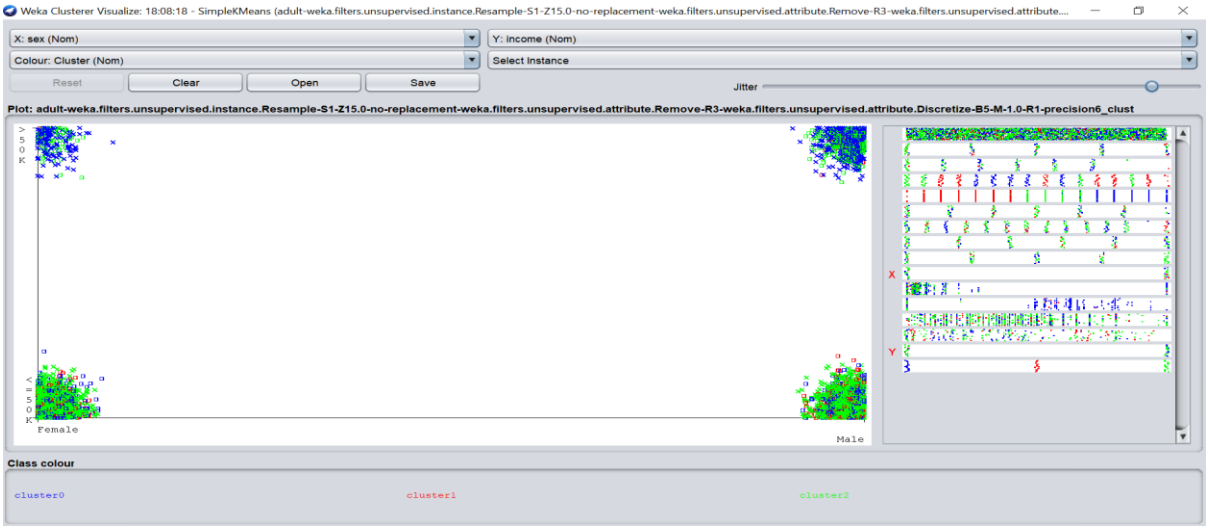


Figure 97 Cluster visualizations with sex on X axis and income on Y axis

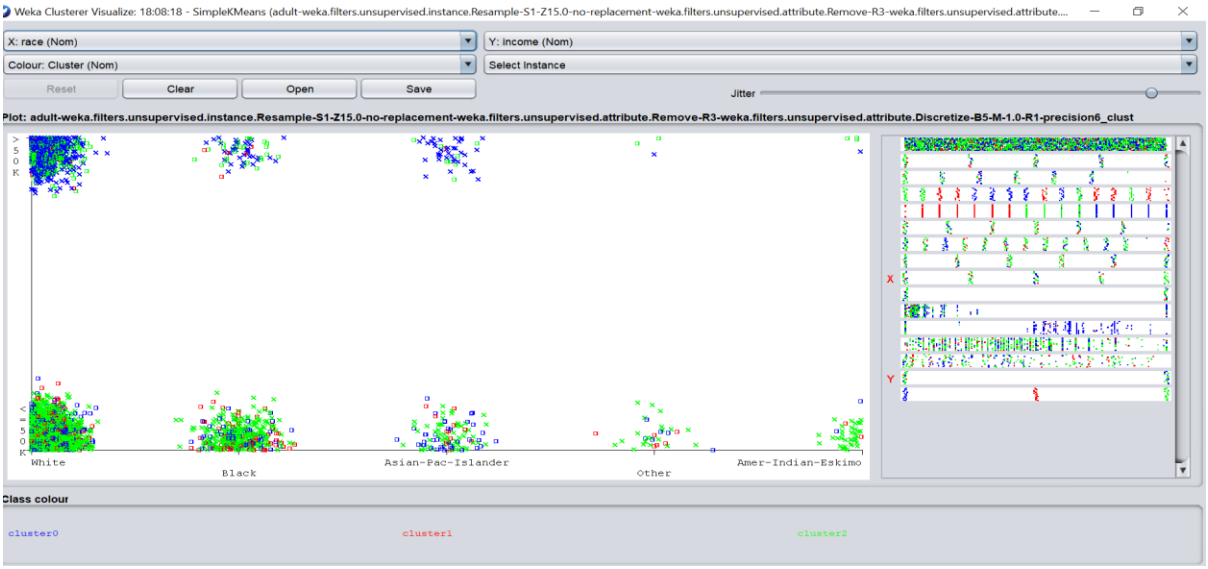


Figure 98Cluster : race Vs income

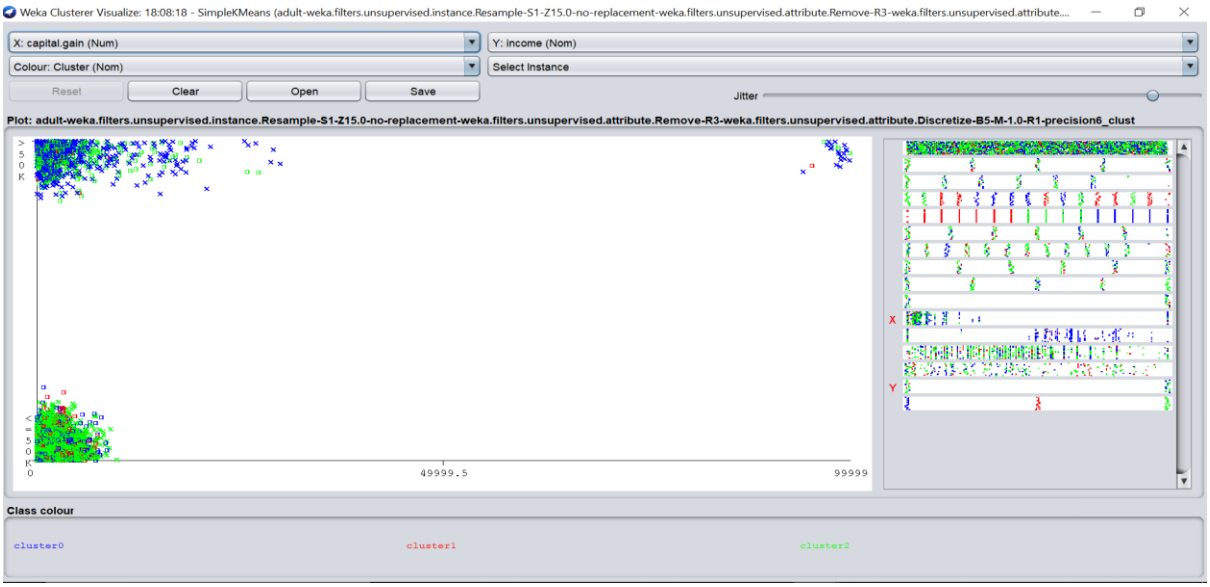


Figure 99capital.gain clusters

Findings

1. Using K-Means Algorithm with Euclidean distance measure, seed set to 20 and numClusters=3, 3 clusters are generated. Random initial points was used. The algorithm iterated 11 times.
   Cluster0:31.6_46.2,Private,Masters,14,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,0,46,United-States
   Cluster1:46.2_60.8,Private,7th-8th,4,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,0,24,United-States
   Cluster2:31.6_46.2,Private,Some-college,10,Married-civ-spouse,Tech-support,Husband,White,Male,0,0,40,United-StatesSum of Within cluster sum of squared errors reduced to 210.68 from the 249.55 in the previous experiment.
2. Classes To Cluster evaluation is used (cs.ccsu.edu, nd), and it is identified that most of the people with income exceeding 50k is assigned to cluster 2(2372

instances). 37.3055% of instances were incorrectly clustered. Thus this has to be improved.

3. Cluster centroids are the mean vectors for each cluster. In the final cluster centroids, some of the factors that are evident are:
    - Cluster 0- A male with Private workclass and Bachelors education, specialising in Prof-specialty with capital gain of 2361.5427.
    - Cluster 1- A male with Private workclass and 10th education, specialising in other service with capital gain of 305.3263.

    Cluster 2- A male with Private workclass and HS-Grad, specialising in craft-repair with capital gain of 536.6745.
4. From the cluster visualization, most of the people with high capital gain and income exceeding 50k belong to cluster 0.
5. All three of the cluster centroids have some attributes similar such as sex, workclass and marital status. However, there is significant difference in education, capital gain and loss etc.

Experiment 3

In this experiment, Manhattan distance which is another similarity function that can be used with K-Means in Weka. Manhattan distance is calculated using:

d(i, j)=|x i1 -x j1 |+|x i2 -x j2 |+...+|x ip -x jp | where i = (xi1, xi2, …, xip) and j = (xj1, xj2, …, xjp) are two p-dimensional data objects, and q is a positive integer. (Dr.AbubakrSiddig- Clustering lecture , 2020). In this experiment, Manhattan distance function with seed set to 30 is used.

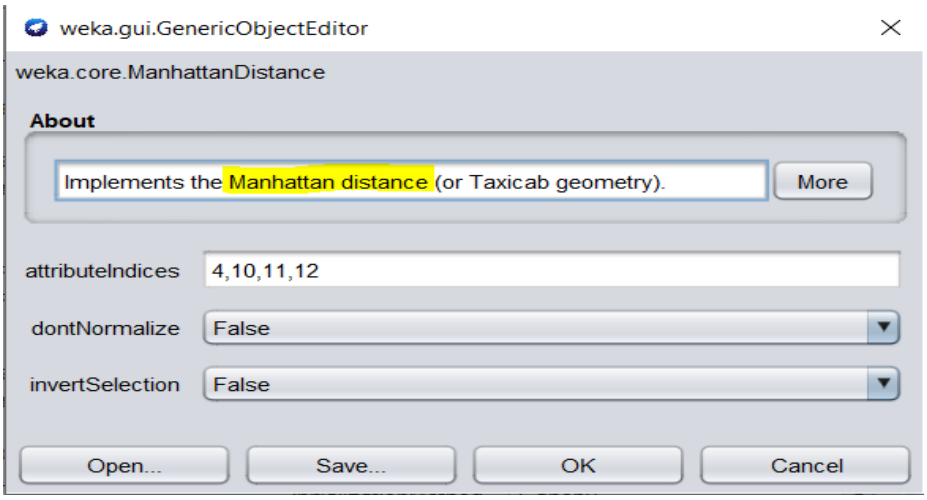


Figure 100 Manhattan distance

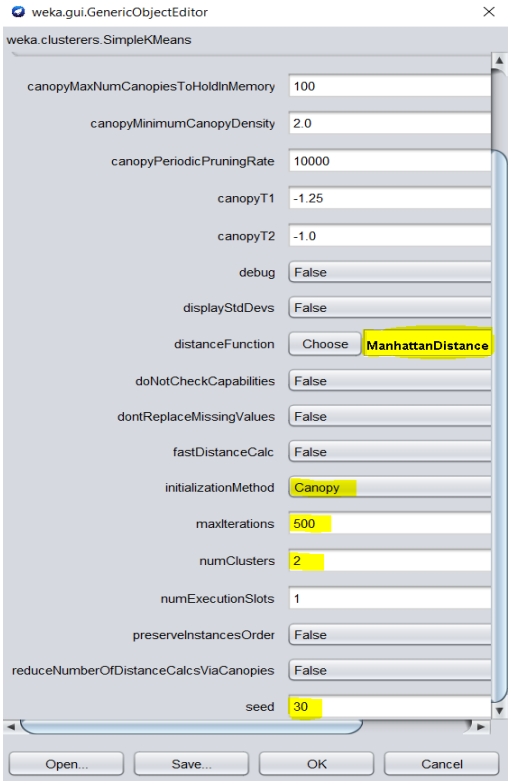


Figure 101 K-Means-Experiment 3-parameters

Result

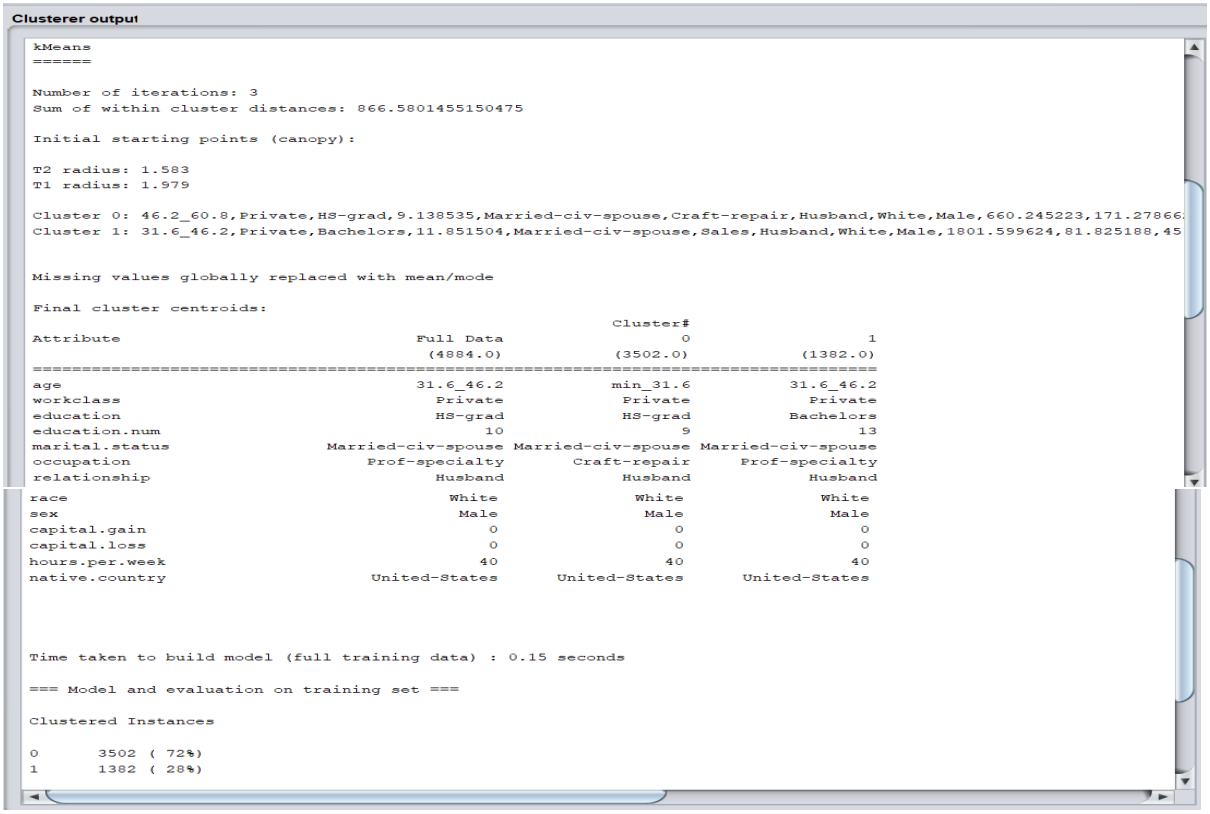


Figure 102 K-Means-Experiment 3-Resul

```
Class attribute: income
Classes to Clusters:

    0    1  <-- assigned to cluster
 2968  751 | <=50K
  534  631 | >50K

Cluster 0 <-- <=50K
Cluster 1 <-- >50K

Incorrectly clustered instances :       1285.0    26.3104 %
```

Figure 103K-Means Experiment 3-Classes to Cluster Evaluation

Figure 104Clusters :marital.status Vs income

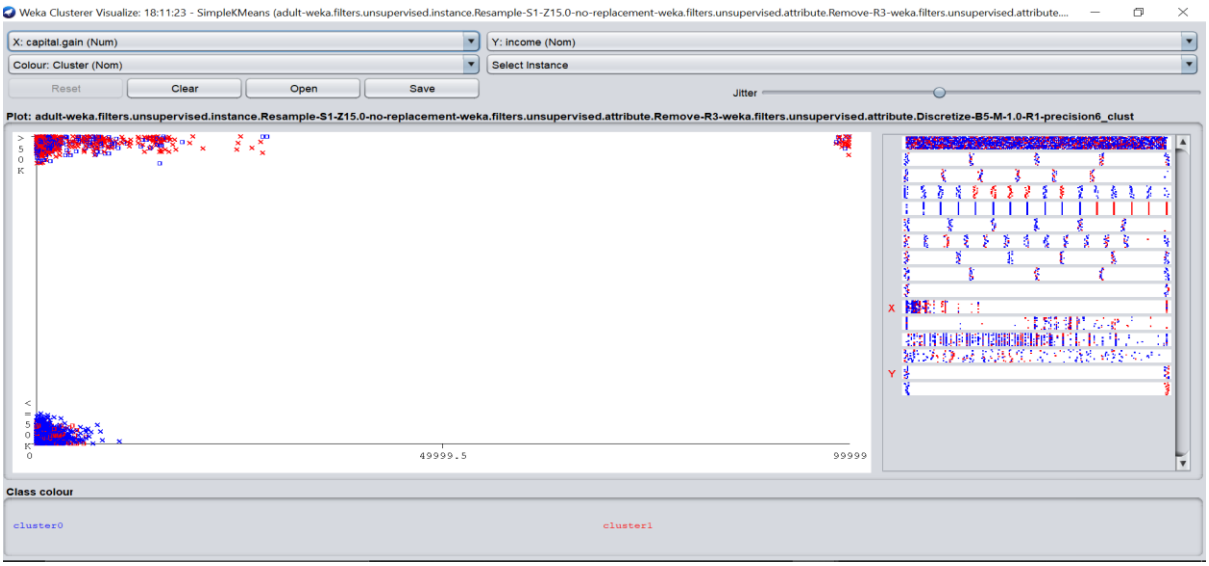
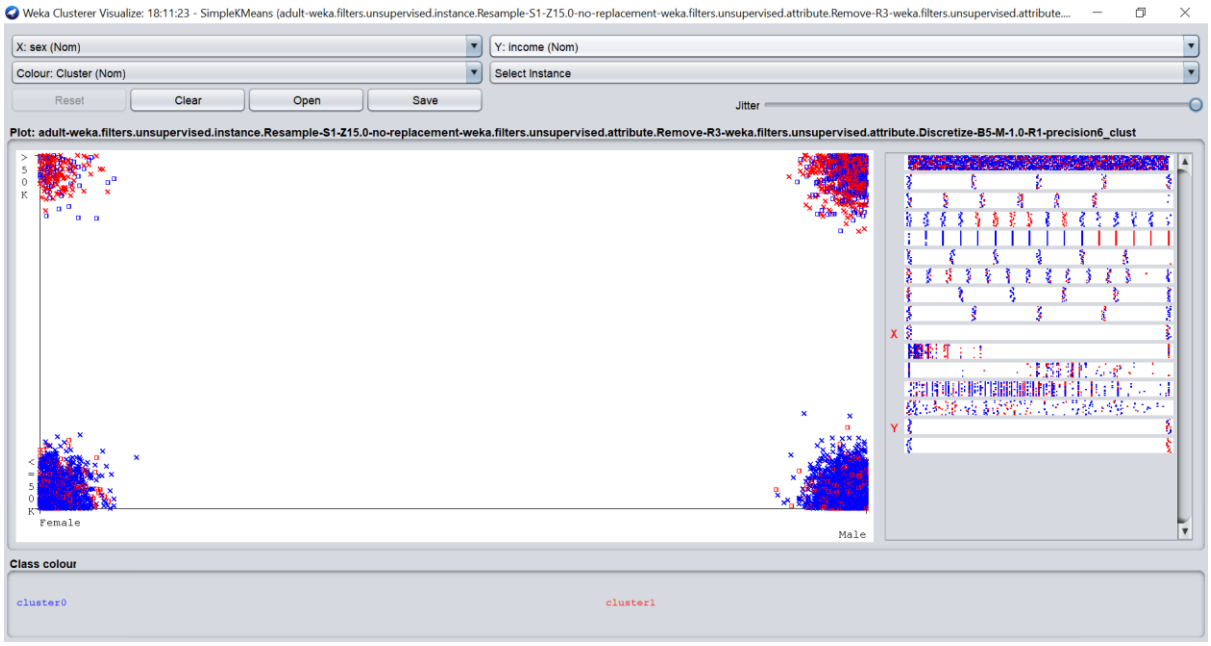


Figure 105capital.gain clusters



Findings

1. Manhattan distance function is used as a similarity function with seed set to 10.
   Cluster 0: 46.2_60.8,Private,HS-grad,9.138535,Married-civ-spouse,Craft-repair,Husband,White,Male,660.245223,171.278662,42.664013,United-States,{628}
   Cluster 1: 31.6_46.2,Private,Bachelors,11.851504,Married-civ-spouse,Sales,Husband,White,Male,1801.599624,81.825188,45.953008,United-States,{532}
2. In this experiment, canopy initialization method was used instead of K-Means clustering which slightly improved the performance with radius T2 radius: 1.583 and T1 radius: 1.979.
3. Sum of Within cluster sum of squared errors is increased to 866.58 even though there is a decrease in percentage of incorrectly clustered instances with classes to cluster evaluation.
4. With canopy preclustering, final centroid of cluster 0 is HS- Grad educated person and in cluster 1 centroid education is Bachelor's.
5. From cluster visualization, we can understand that cluster 0 is dominated by people with income less than or equal to 50K US Dollars.

7. Time Series Forecasting – 15%

7.1. Dataset description

Dataset for Time Series – Madrid Weather Dataset

- **Title**: Time Series Forecasting on the Weather Madrid 1997-2015 data.
- **Data description:**

1. The problem domain

Climate studies and weather forecasting is a very important study, which is increasingly becoming the need of the hour. Climate data analytics is gaining more popularity due to the catastrophic affects of climate change. Millions around the world are vulnerable to climate change related disasters, especially the poor and people residing in ecologically

sensitive areas. The Intergovernmental Panel on Climate Change, a body of 1,300 independent scientific experts from countries around the world under the auspices of the United Nations, stated in its Fifth Assessment Report that there is more than 95% chance that human activities and intrusion have heated the climate over the past 50 years leading to the melting up of glaciers(Gender economic inequality,2020).In this century, the global temperatures can be expected to rise by 3.4 to 3.9°C and the world is already 1.1°C warmer than it was during the beginning of the Industrial Revolution (Breast Cancer Wisconsin,2020). Dr. Waleed Abdalati, director of the Cooperative Institute for Research in Environmental Sciences (CIRES), former chief scientist of NASA and world's leading climate scientist has talked about "environmental intelligence" and the importance of continuous data. He warns about the consequences of further delay by stating "The way I look at it is that the longer we as a society wait, the greater the sacrifices we have to make to address the situation." (Pima Indian diabetes database,2020).

Thus, weather forecasting and climate studies as a domain is extremely crucial for this century.

2. **The source of the data**

The data is obtained from Kaggle from the link :
https://www.kaggle.com/juliansimon/weather_madrid_lemd_1997_2015.csv#weather_madrid_LEMD_1997_2015.csv
However, the original data comes from the Weather Underground which is owned by the Weather Company, LLC. (Local Weather Forecast, News and Conditions | Weather Underground, 2020) which is a weather forecasting and information technology company founded on 1982 with head quarters in the US (The Weather Company, 2020). Weather Underground is a commercial weather service that offers Weather information on the Internet in real time. Weather Underground publishes weather forecasts on its website for several significant cities around the world, as well as local weather updates for newspapers and third-party websites. Its information comes from the NWS and over 250,000 personal weather stations (PWS) and observatories. (Weather Underground (weather service), 2020).The data available on Kaggle, is obtained from the Weather Underground website and is made publicly available for use.

3. **Agencies working with the data**

The dataset used in this assignment is obtained from Weather Underground and is collected by an individual. However, the data from Weather Underground is utilised by several agencies working with climate data. Weather Underground's mission is to "make quality weather information available to every person on this planet." The data collected is combined with the vast knowledge of the meteorologists to produce meaningful information, hence all data is backed by science. Many organizations, researchers and agencies are working with the daily weather data from Weather Underground such as:

1. **Associated Press**

The Associated Press (AP) is an American not-for-profit news agency that utilises weather data from Weather Underground, to provide weather summaries. (Weather Underground (weather service), 2020) The AP operates 263 news bureaus in 106 countries worldwide.

2. Weather Underground Braille page

The company formerly broadcast NOAA Weather Radio stations' Internet radio broadcasts from around the country, as received by consumers, and has a Weather Underground Braille Site. (Associated Press, 2020)

4. The intended use of the data

Understanding weather patterns of a city like Madrid can be very useful in understanding the pattern of climate change from 1997, this decade is extremely important for the world as we have less time left for reversal of climate change affects. The natural calamities are alarming and raise several questions about the environmental conservation and sustainability policies of governments worldwide. The technology and the huge computational power at our disposal along with the data analytics and mining solutions can save the planet to a great extent. The problem of climate change is not an isolated issue, the repercussions of inactions are faced by millions of people around the globe, and the next generations will face the adversities of the activities of the mankind today.Big data and data miningcan save the planet by climate study and weather forecasting. Madrid is a significant city of the world, and the data from Madrid can form basis for research on this field. This data can be communicated as maps, charts, and other visual formats.

Weather forecasting helps us to know what to expect in the future based on the past data. This can be visualized alongside the effects of urbanizations and industrialization to get a clear picture of the cause of the change in pattern. Time Series Forecasting is a very important data mining technique that can be performed on this data, using Weka or another software tools like Tableau etc. The scope of data mining on this data is enormous such as Linear Regression, SMO Regression etc.

5. Attribute types of the data

The dataset is multivariate and consists of 23 attributes

Attribute Name	Description	Type	Distinct
CET	Date of sample of weather Madrid	Date	6812
Max TemperatureC	Maximum temperature recorded during a day	Numeric	42
Mean TemperatureC	Mean temperature recorded during a day	Numeric	36
Min TemperatureC	Minimum temperature recorded during a day	Numeric	39
Dew PointC	Dew point recorded in a day	Numeric	32
MeanDew PointC	Mean dew point recorded during a day	Numeric	32
Min DewPointC	Minimum dew point recorded during a day	Numeric	37
Max Humidity	Maximum dew point recorded during a day	Numeric	66
Mean Humidity	Mean humidity recorded during a day	Numeric	86
Min Humidity	Minimum humidity	Numeric	86

	recorded during a day		
Max Sea Level PressurehPa	Maximum sea level pressure recorded during a day	Numeric	51
Mean Sea Level PressurehPa	Mean sea level pressure recorded during a day	Numeric	54
Min Sea Level PressurehPa	Minimum sea level pressure recorded a day	Numeric	57
Max VisibilityKm	Maximum visibility kilometre recorded during a day	Numeric	21
Mean VisibilityKm	Mean visibility kilometre recorded during a day	Numeric	32
Min VisibilityKm	Minimum visibility kilometre recorded during a day	Numeric	24
Max Wind SpeedKm/h	Maximum wind speed kilometre per hour recorded during a day	Numeric	40
Mean Wind SpeedKm/h	Mean wind speed kilometre per hour recorded during a day	Numeric	24
Max Gust SpeedKm/h	Maximum gust speed kilometre per hour recorded during a day	Numeric	42
Precipitationmm	Precipitation recorded during a day	Numeric	21
CloudCover	Cloud cover in the region	Numeric	9
Events	Events different types that occurred are Rain, Rain-Snow, Snow, Fog, Fog-Rain, Rain-Thunderstorm, Thunderstorm, Rain-Hail-Thunderstorm, Fog-Thunderstorm, Tornado, Fog-Rain-Thunderstorm, Fog-Rain-snow, Fog-Snow, Rain-Snow-Thunderstorm, Rain-Hail recorded during a day	Nominal	15
WindDirDegrees	Wind dir degrees recorded during a day	Numeric	361

Screenshots of summary and graphs of attributes

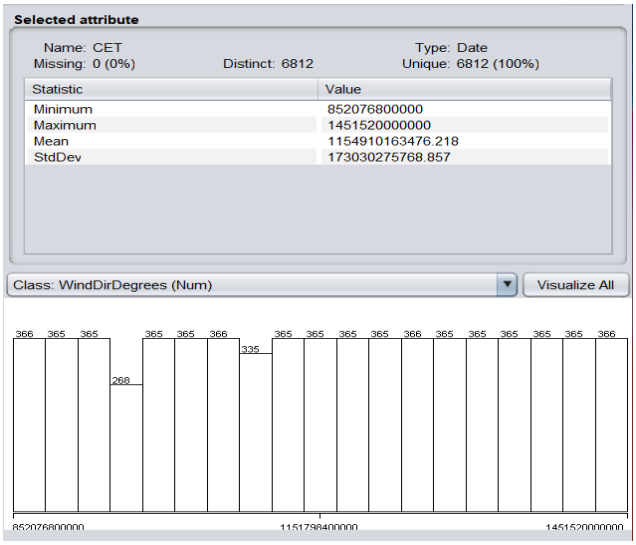


Figure 106 CET

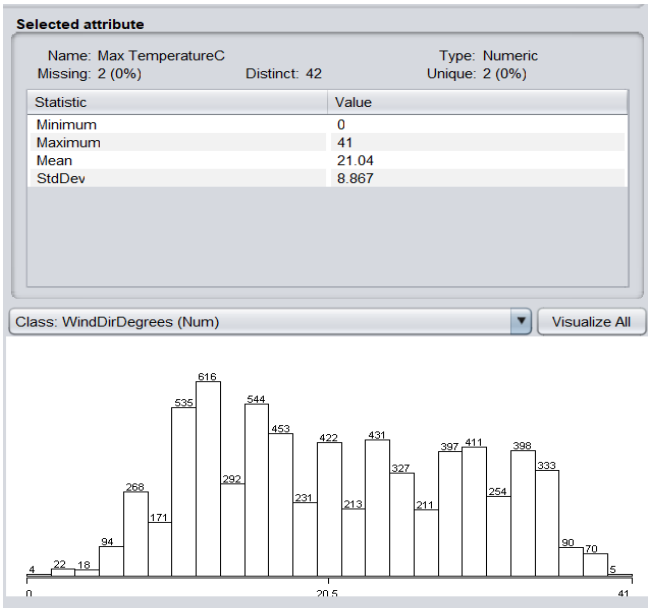


Figure 107 Max TemperatureC

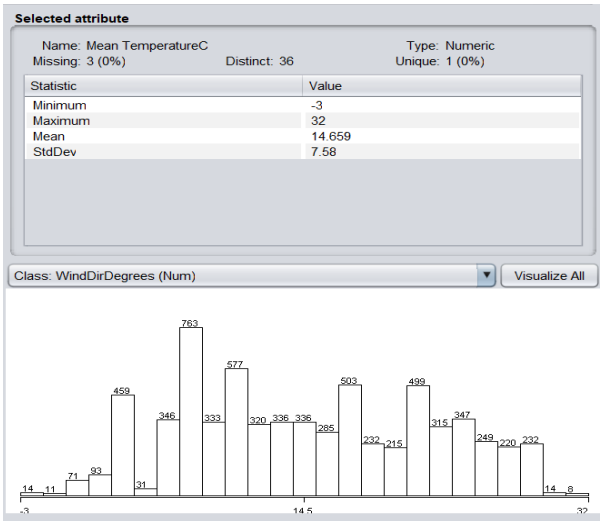


Figure 108 Mean TemperatureC

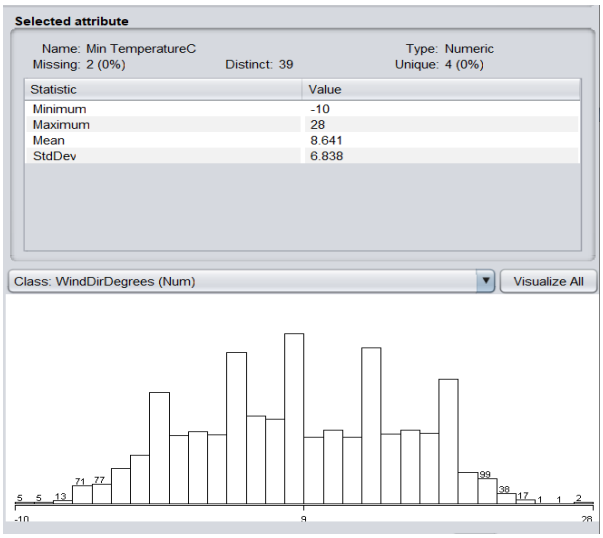


Figure 109 Min TemperatureC

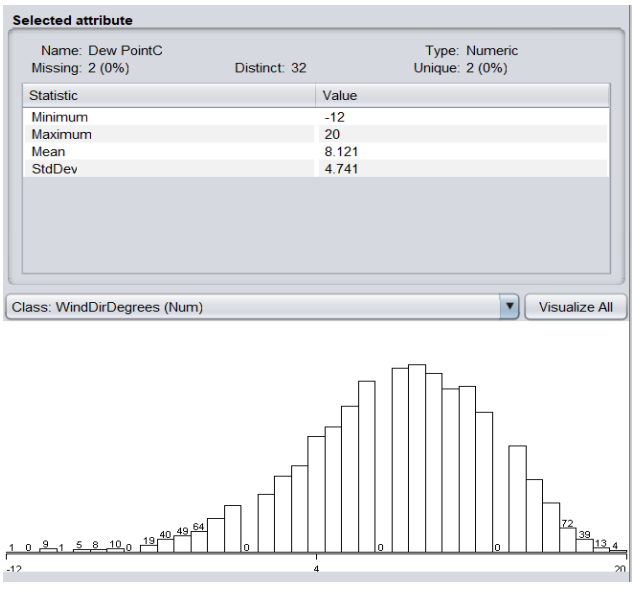


Figure 110 Dew PointC

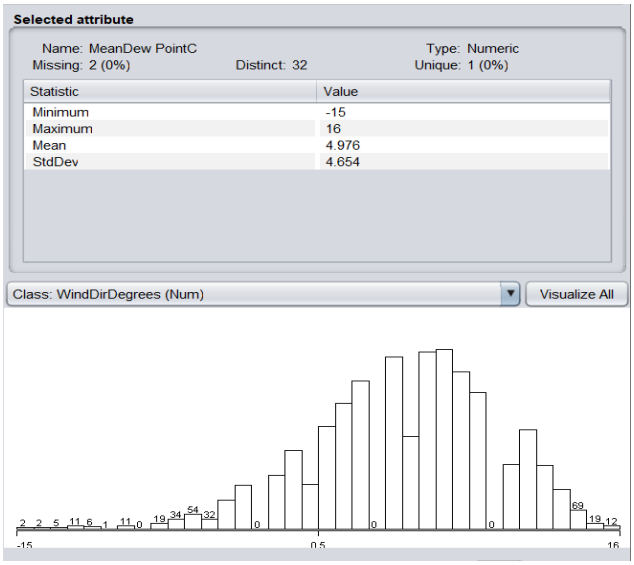


Figure 111 MeanDew PointC

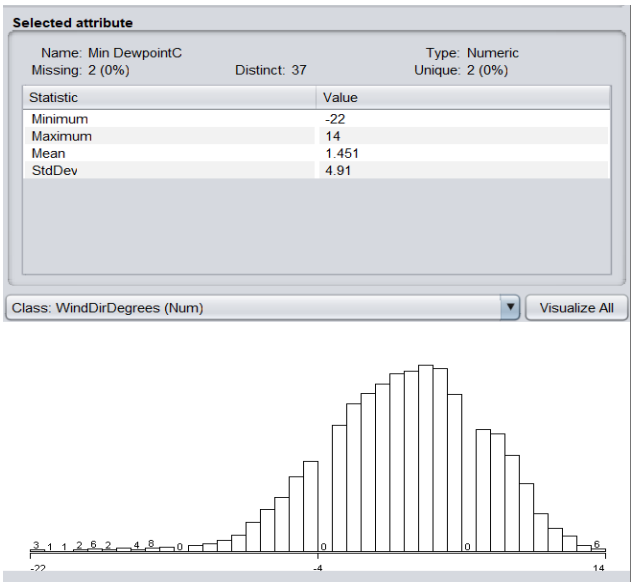


Figure 112 Min DewpointC

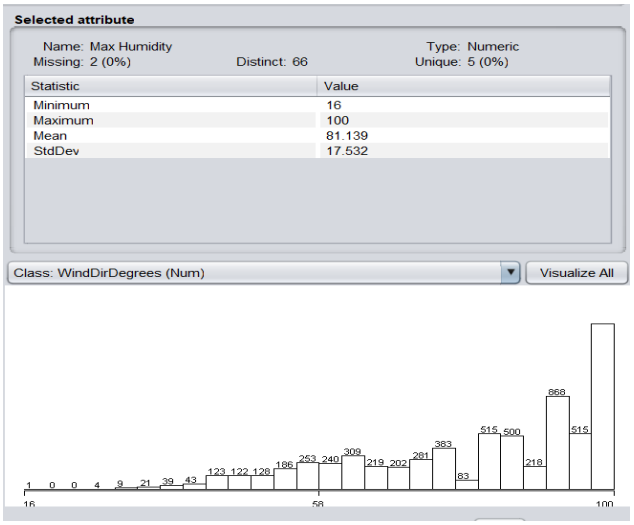


Figure 113 Max Humidity

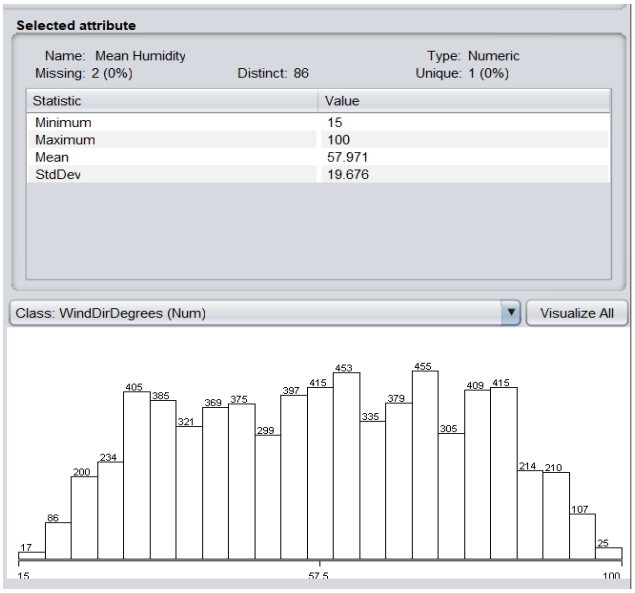


Figure 114 Mean Humidity

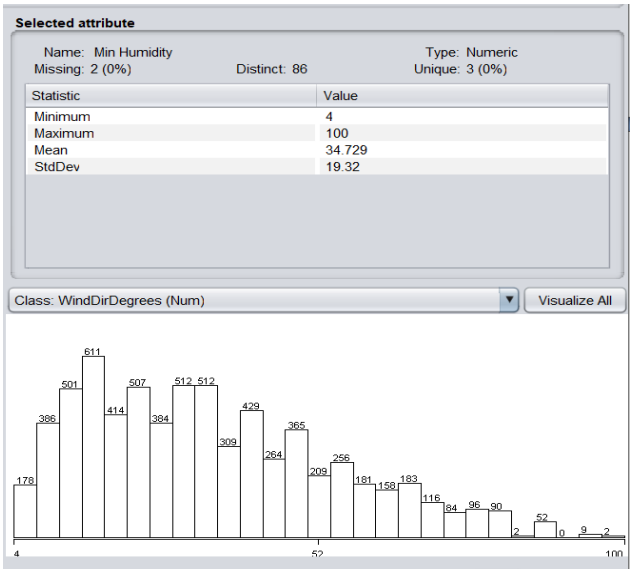


Figure 115 Min Humidity

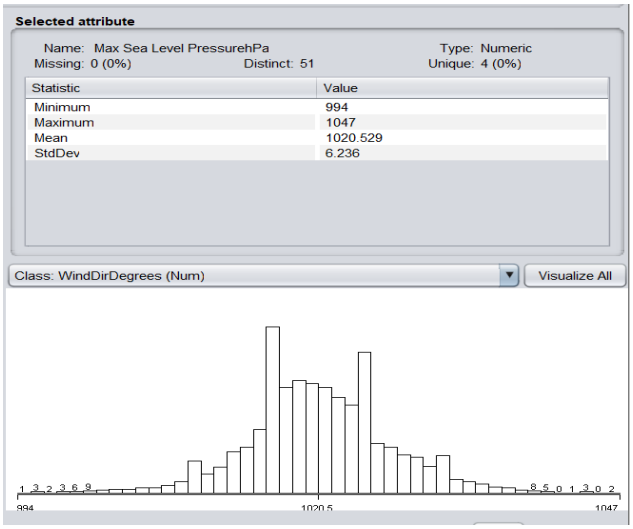


Figure 116 Max Sea level PressurehPa

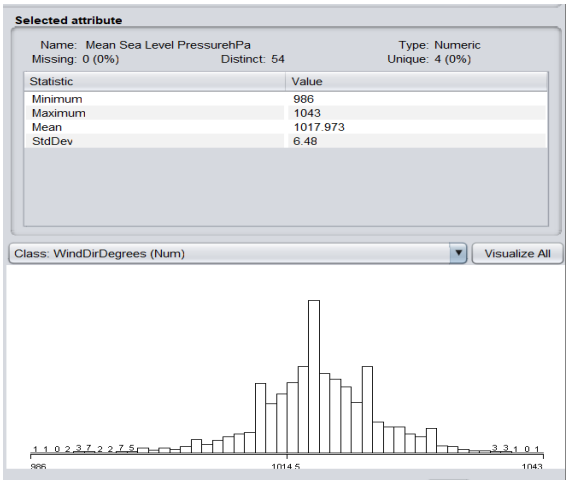


Figure 117 Mean Sea Level PressurehPa

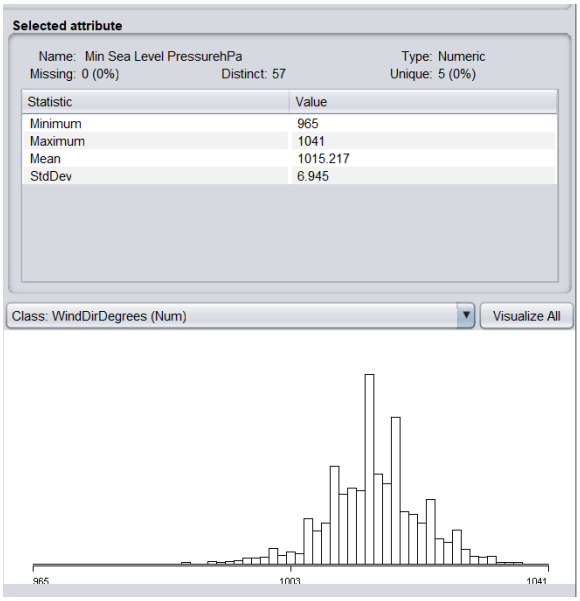


Figure 118 Min Sea Level PressurehPa

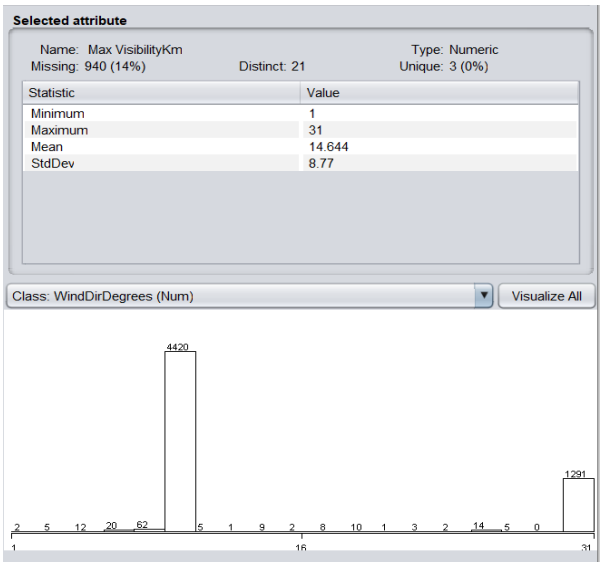


Figure 119 Max VisibilityKm

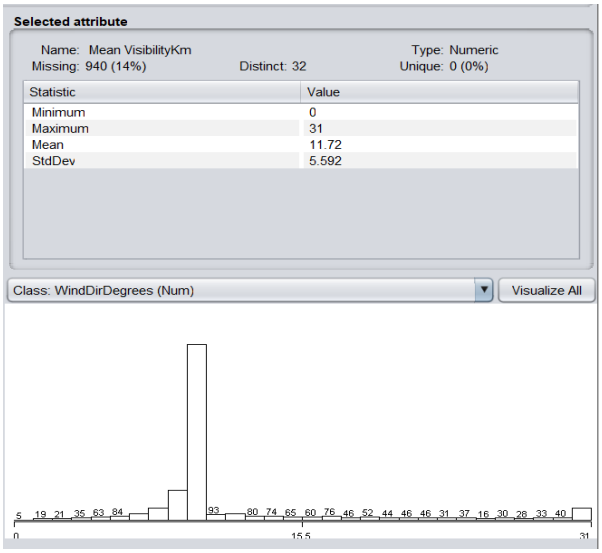


Figure 120 Mean VisibilityKm

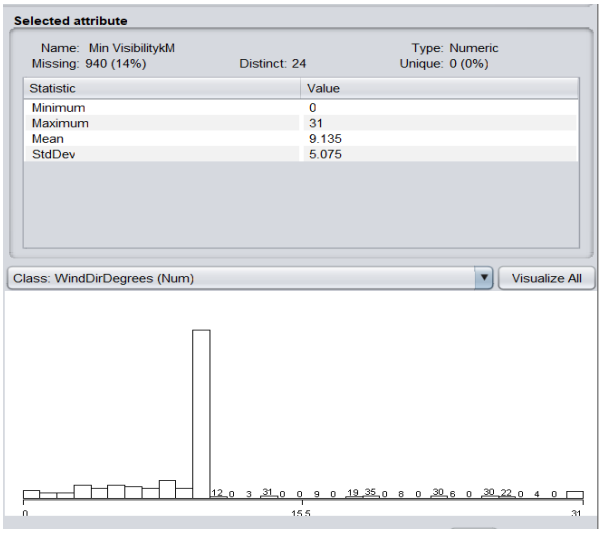


Figure 121 Min VisibilitykM

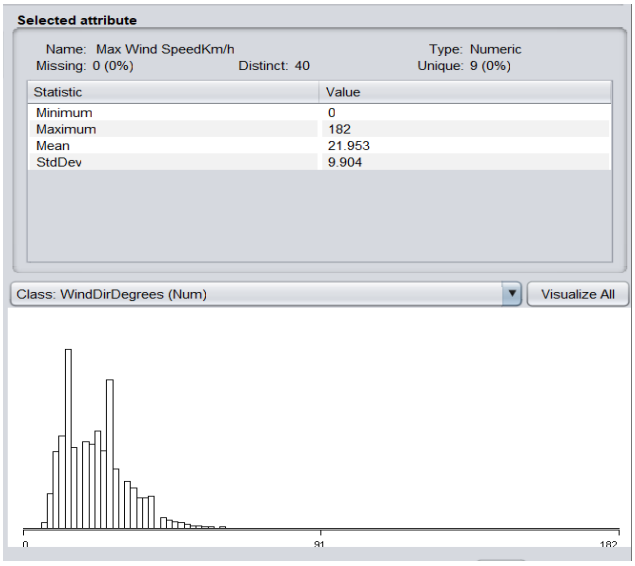


Figure 122 Max Wind SpeedKm/h

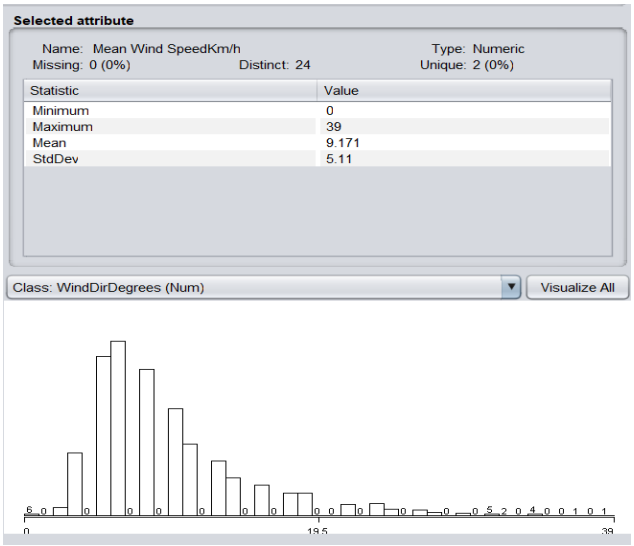


Figure 123 Mean Wind SpeedKm/h

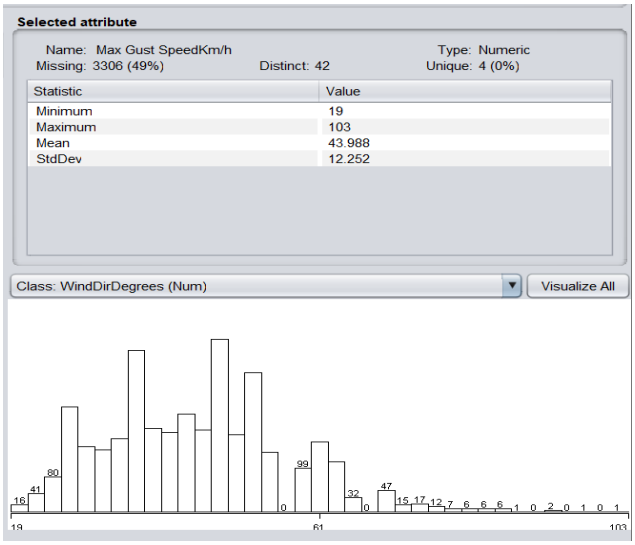


Figure 124 Max Gust SpeedKm/h

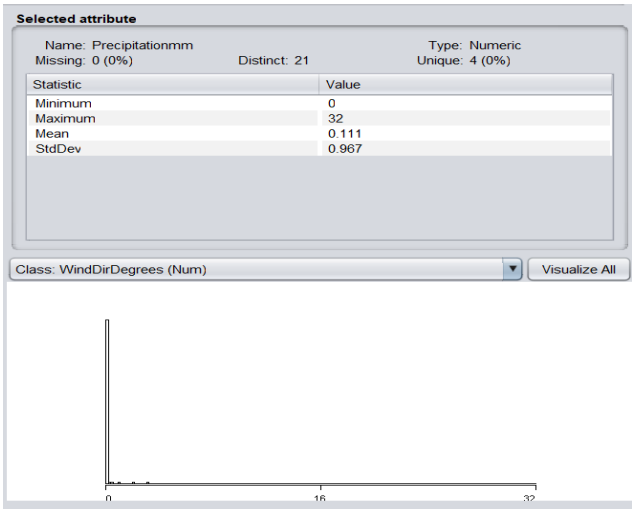


Figure 125 Precipitationmm

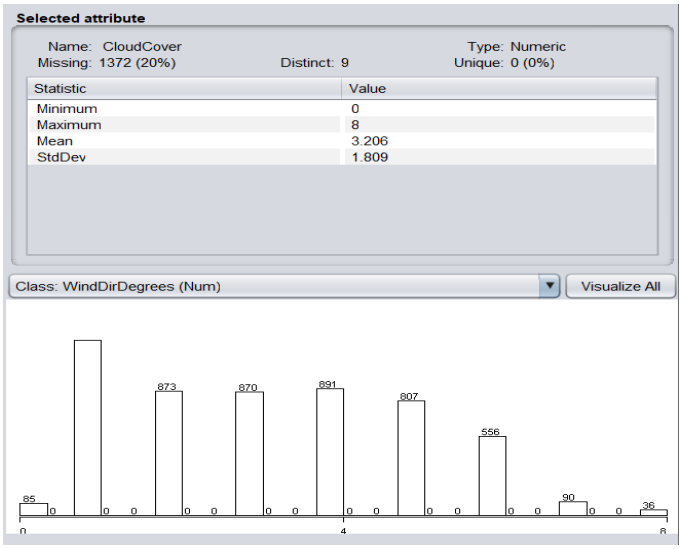


Figure 126 CloudCover

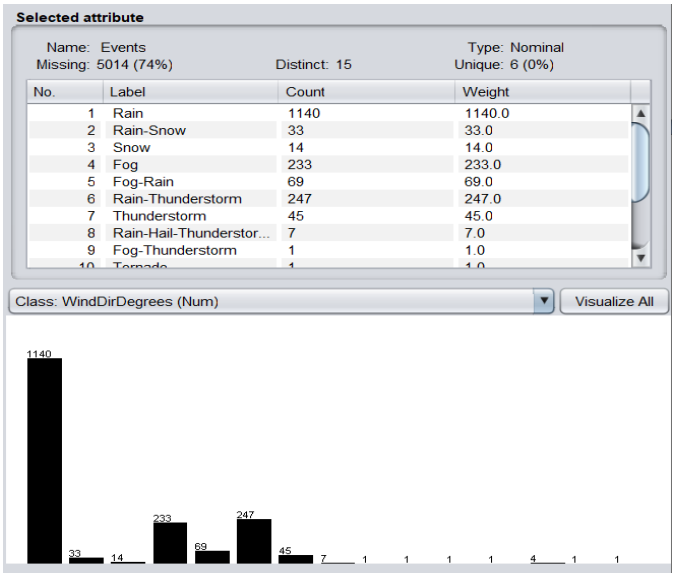


Figure 127 Events

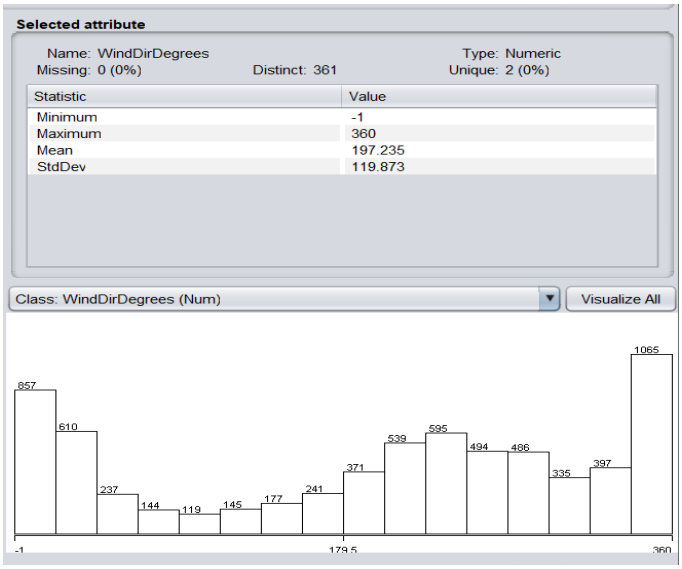


Figure 128 WindDirDegrees

Observations from visualizations

1. There are 23 attributes and 6812 instances.
2. There are 21 numeric attributes, 1 Date Attribute and 1 Nominal attribute.
3. Two attributes have huge percentage of missing values. 'Events' attribute has 74% missing values, 'MaxGust SpeedKm/h' has 49% missing values.

Objective

1. To forecast Mean Temperature in Celsius and Precipitation in millimetre for the next 20 consecutive days in the Madrid Weather 1997-2015 dataset.
2. To understand the affect of different base learners such as SMO Regression, Linear Regression and MultiLayer Perceptron and to compare their affects on time series forecasting.
3. To create graphs to visualize historical values and future predictions.
4. To perform evaluation on the forecasting models, with different parameters such as Root Mean Squared Error, Mean Absolute Error etc.

Summary of Findings

Preprocessing

- Time series preprocessing technique is different from classification preprocessing and clustering preprocessing technique.
- Applying SMOTE filter increased the performance of classification algorithm J48inthisdatasetdramatically.ThisisbecausetheSMOTEfilterreducescla ss imbalance.
- Preprocessing steps used in clustering is not the same as the steps used in the classification problem. Hence preprocessing steps are performed again on the original resampled dataset adult_resampled.Arff
- All missing values are replaced with modes and means accordingly using ReplaceMissingValues Filter.
- Outliers and extreme values in the dataset are detected using Weka's unsupervised attribute InterQuartileRange Filter.
- Data Type Conversion is performed using NumericToNominal

Time Series Experiments

1. Three experiments were performed for forecasting mean temperature in Celsius and precipitation in mm for the next consecutive 20 days using three different base learners such as Linear Regression, SMO Regression and MultiLayerPerceptron.
2. In all three experiments, evaluation was performed on held out training data of 0.1 and 95% confidence interval.
3. One step ahead evaluation is used as errors propagate in a time series, with each unit, so it can be seen using this evaluation.

4. Simple regression models worked well with the Time Series forecasting much better than the complicated Neural Network.
5. Linear Regression offered the best performance with mean absolute error for temperature after 20 steps is only 3. 5279, for precipitation is 1.0673 which is really good, this means that forecast is close to accurate.Root mean squared error for temperature is 4.5789 and precipitation is 2.2242.
6. Predictions using regression models were easy to interpret. SMO Regression also worked good enough with Mean Absolute Error of 4.0203 for temperature, and increased and 1.2431 for precipitation.

7.2. Preprocessing

1. Data Type Conversion

The dataset, contained the 'CTE' which is the attribute to store recorded dates in the nominal form.

This was converted into the Date format which is suitable for Time Series forecasting by changing Nominal to Date using a Text editor as shown below.

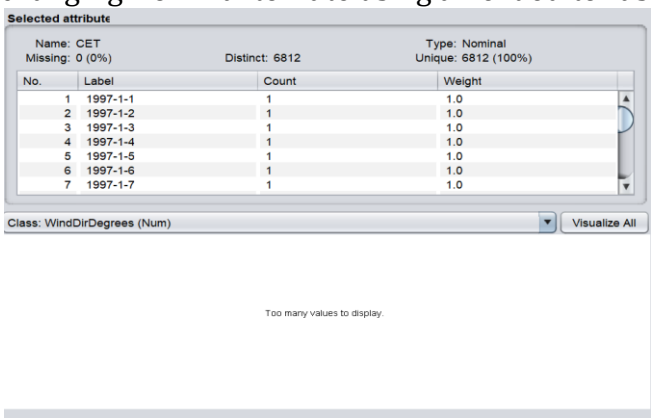


Figure 129 Before Conversion

Figure 130 CET as Nominal

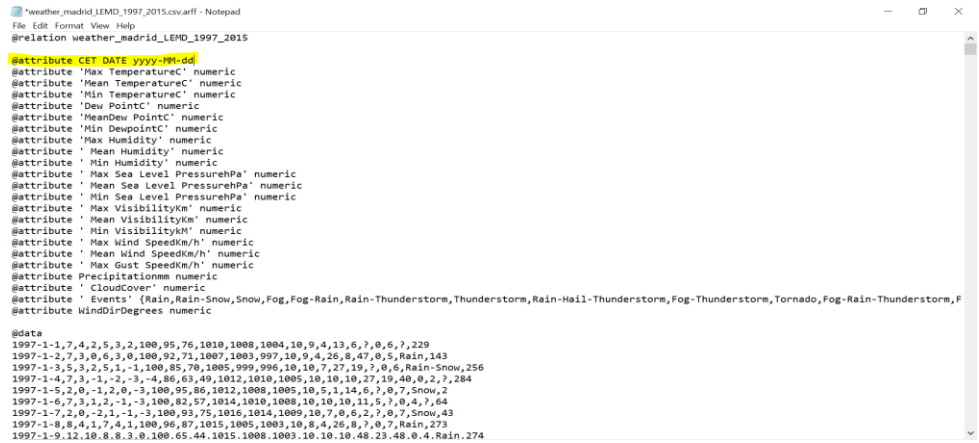


Figure 131 CET as Date

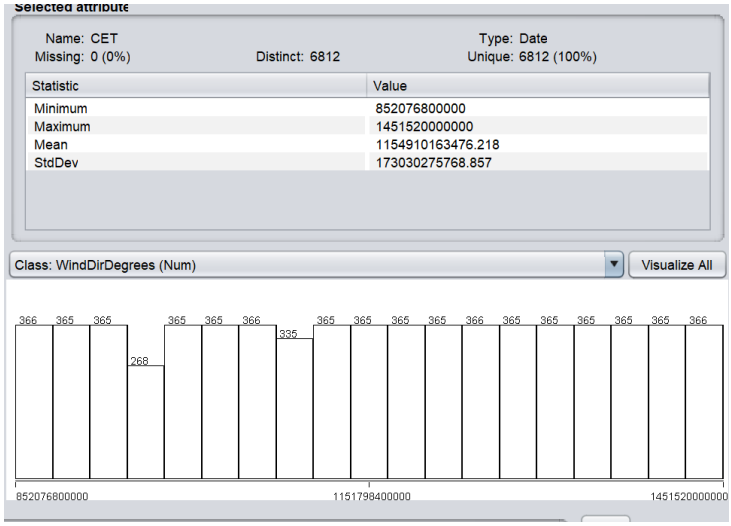
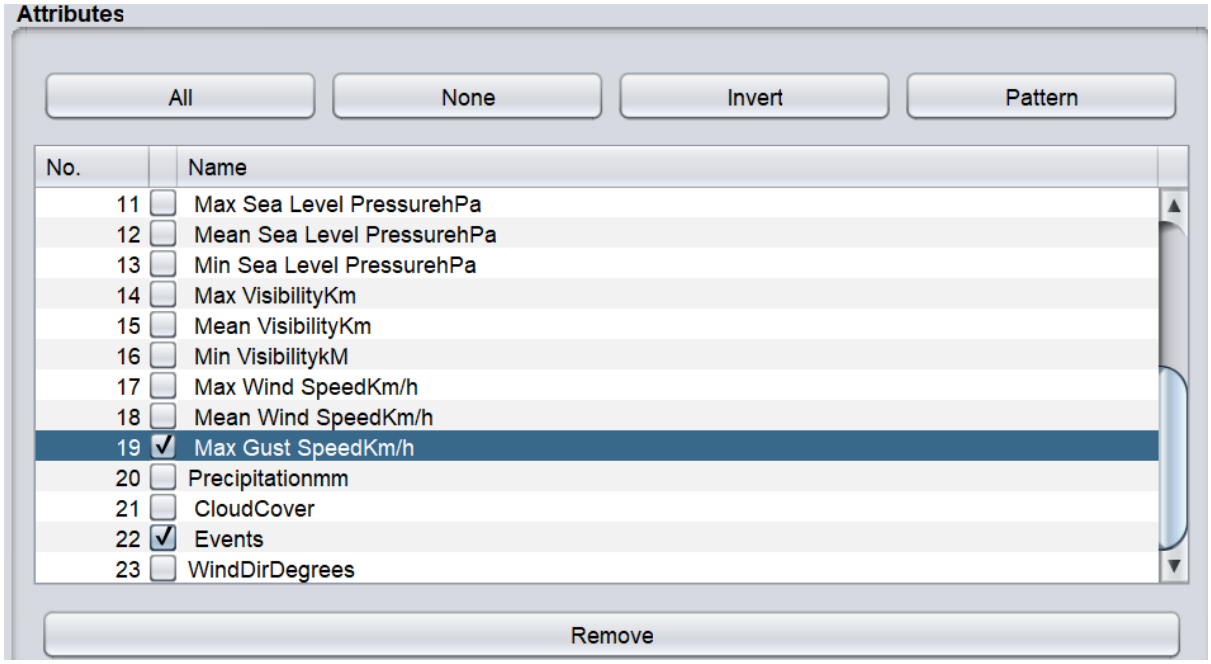


Figure 132 After Conversion

2. Missing Values

Data can be missing due to a variety of reasons- hesitance of the respondents to provide complete information, malfunctioning of equipments, errors when entering the data in to the database, sudden changes etc etc(Dr.AbubakrSiddig- Datasets,EDA and altering data structure lecture , 2020). A small amount of missing value is almost unavoidable in large datasets. However, a significant percentage of missing values can be problematic.

There are some attributes with missing values; 'MeanVisibilityKm'-14% missing values, 'MaxVisibilityKm'-14% missing values, 'MinVisibilityKm'-14% missing values'MaxGustSpeed Km/h' -49%, 'Events' -74% missing values, 'CloudCover' -20% missing values. As a rule of thumb, attributes with more than 40% missing values are not really useful observations.



Hence Max Gust SpeedKm/h and Events is removed.

Missing values of other attributes are removed by replacing with means using ReplaceMissingValues Filter.

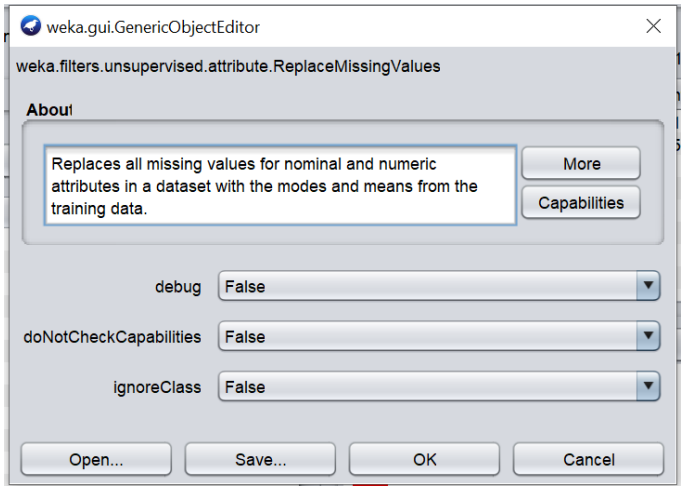


Figure 133 Replace Missing Values Filter

Figure 134 Before replacing missing values

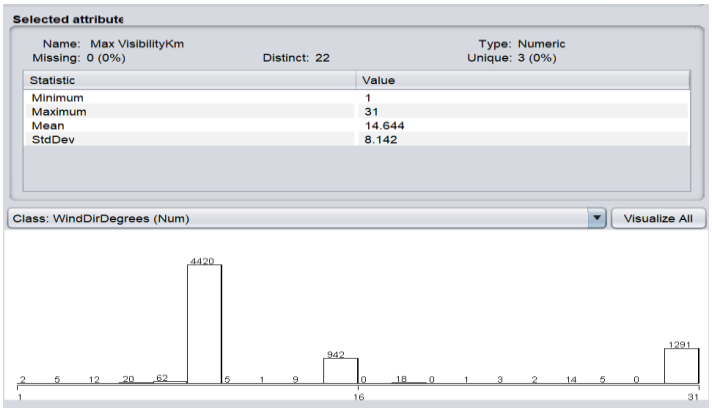


Figure 135 After replacing missing values

3. Outliers

Outlier is a data point that differs significantly from other observations which can be due to error in measurements or exceptional cases (Wikipedia Outlier, 2020). Outliers and extreme values in the dataset are detected using Weka's unsupervised attribute InterQuartileRange Filter. It gives us the middle spread of the data. This filter skips the class values. It creates two new features Outlier and ExtremeValue with two distinct values 'No' and 'Yes' for all instances.

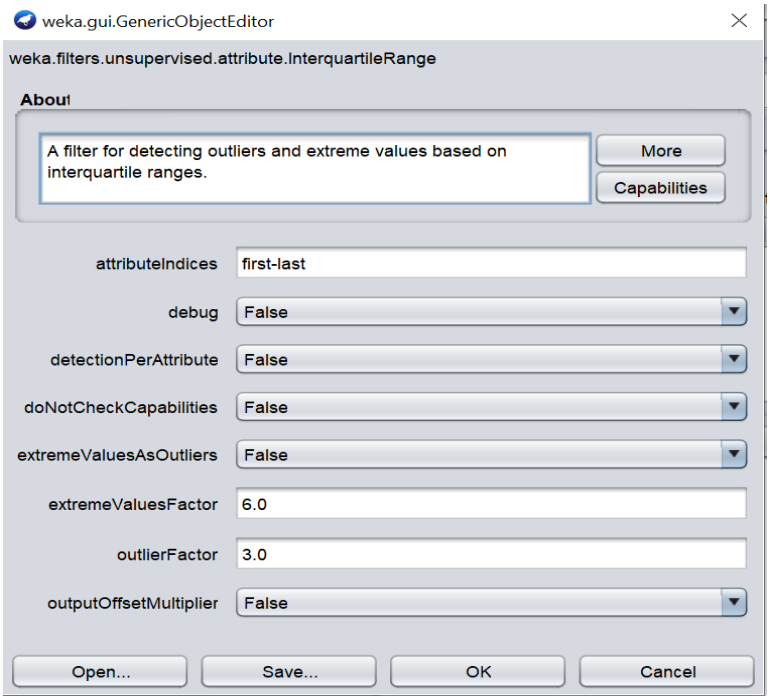


Figure 136InterQuartileRange Filter

This dataset does not contain any outliers or extreme values.

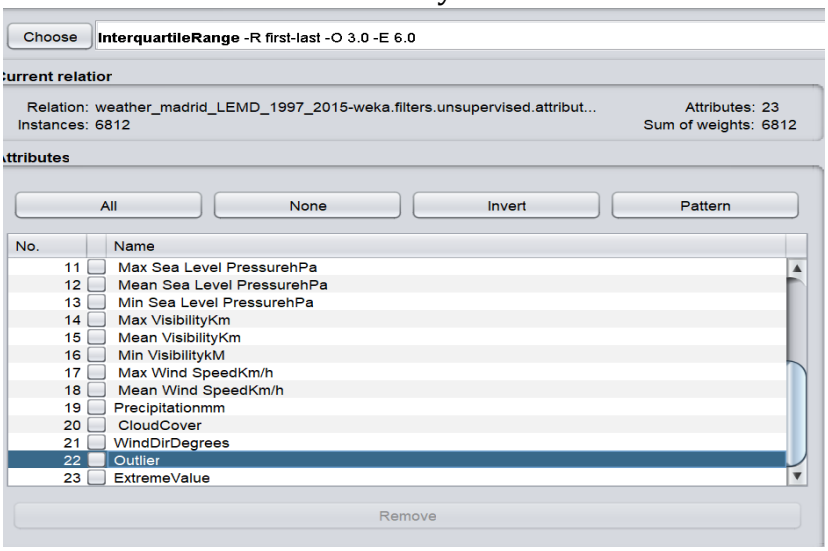


Figure 137 Attributes

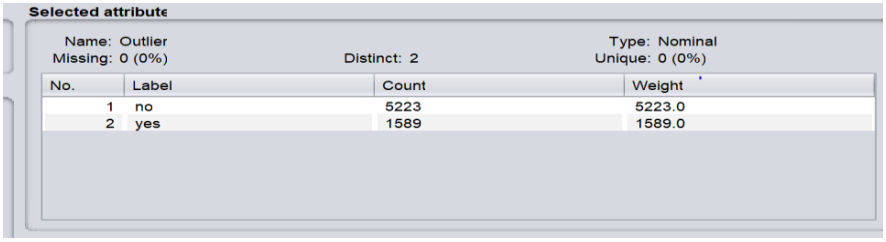


Figure 138 Outlier

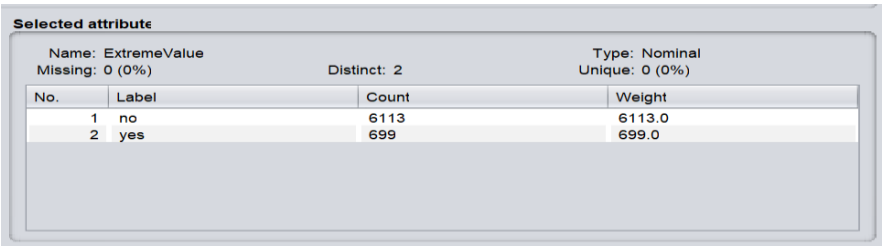


Figure 139ExtremeValue

There are very few outliers and extreme values, hence it is ignored as it will not adversely affect time series forecasting.

7.3. Experiments

Time Series forecasting is used to predict two attributes Mean TemperatureC and precipitationmm

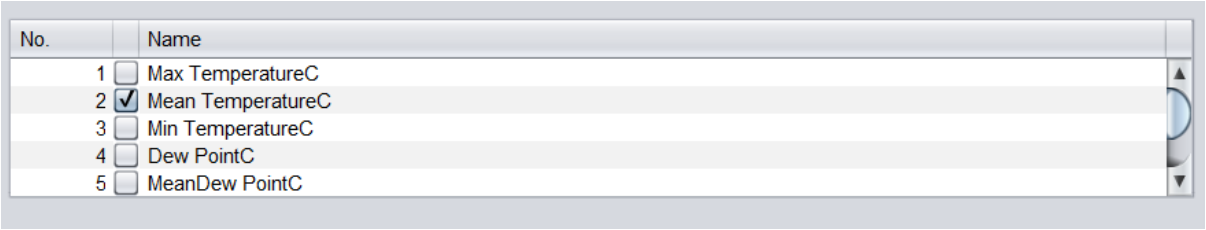


Figure 140 Basic configuration- Mean TemperatureC

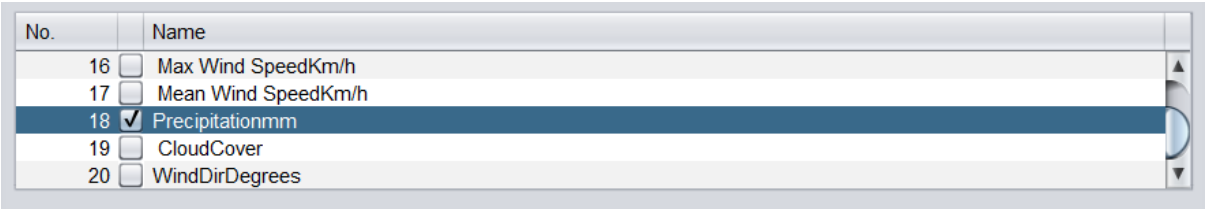


Figure 141Basic configuration Precipitationmm

20 units are forecast in every experiment, with 95% confidence interval and evaluation is performed using parameters such as Mean Absolute Error, Root Mean Squared Error etc.

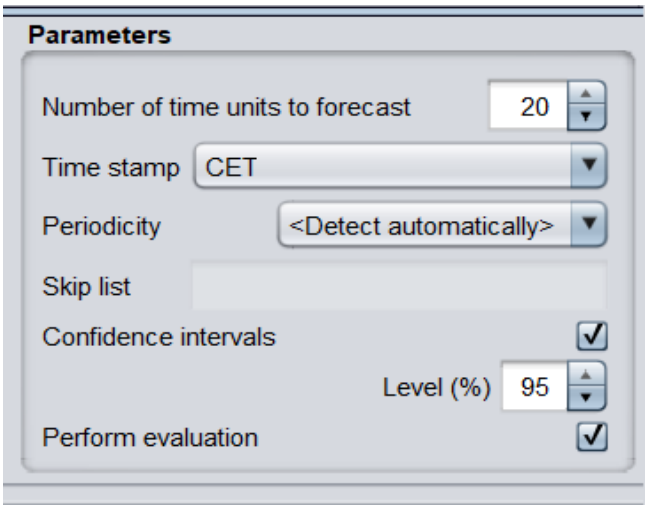
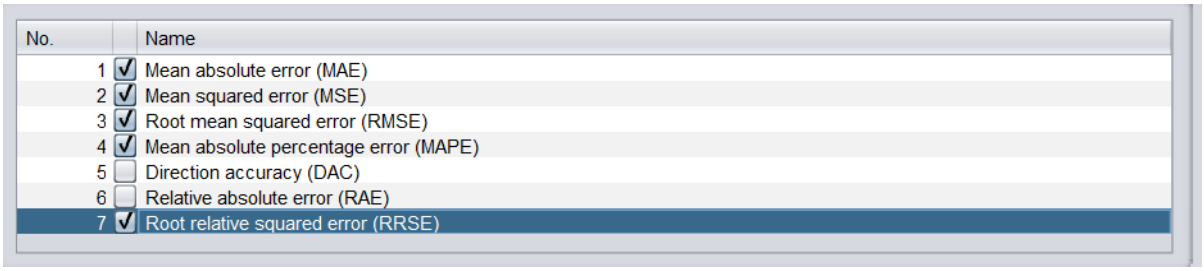


Figure 142 Time series-Parameters



Evaluation is performed on held out training as 0.1.

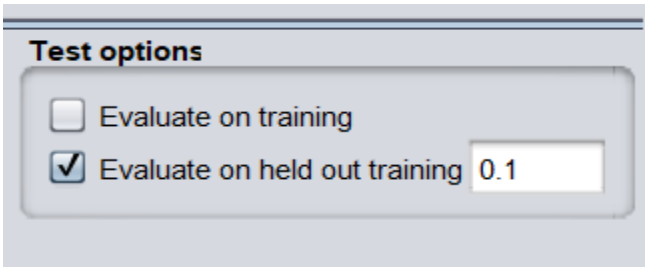


Figure 143 Test options

Experiment 1

In experiment 1, Linear Regression is used as a base learner

Result

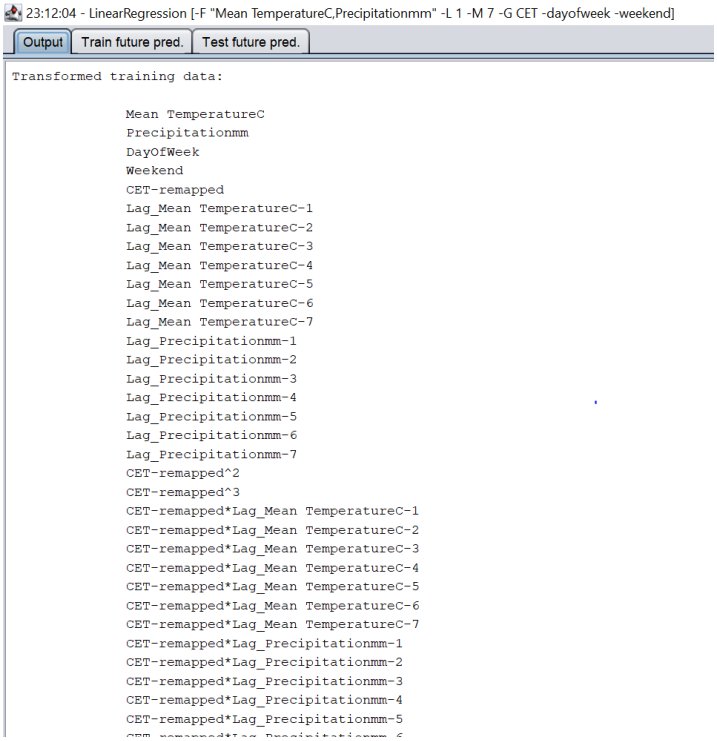


Figure 144 Transformed training data

This shows the transformed values of training data.

Linear regression equation

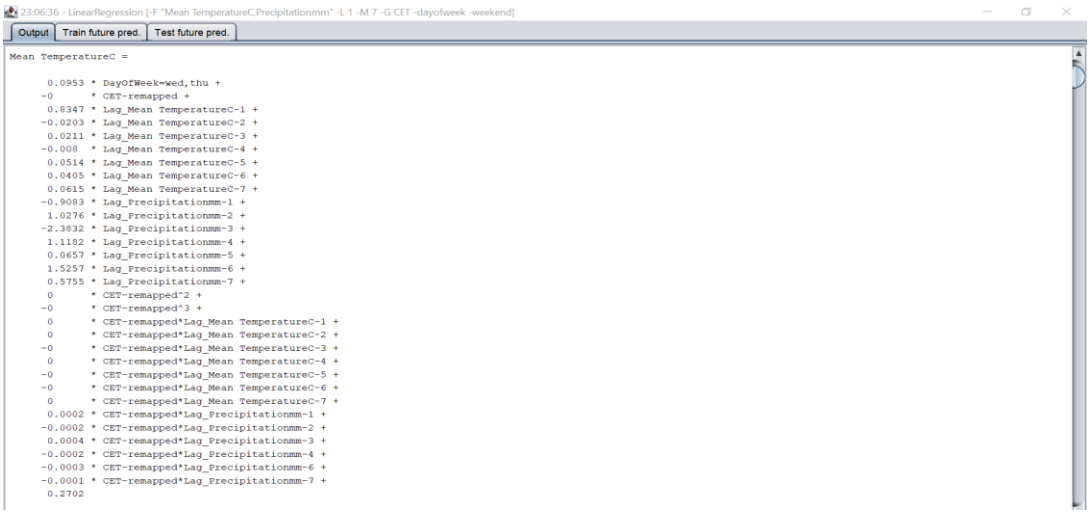


Figure 145 Regression model- Mean TemperatureC

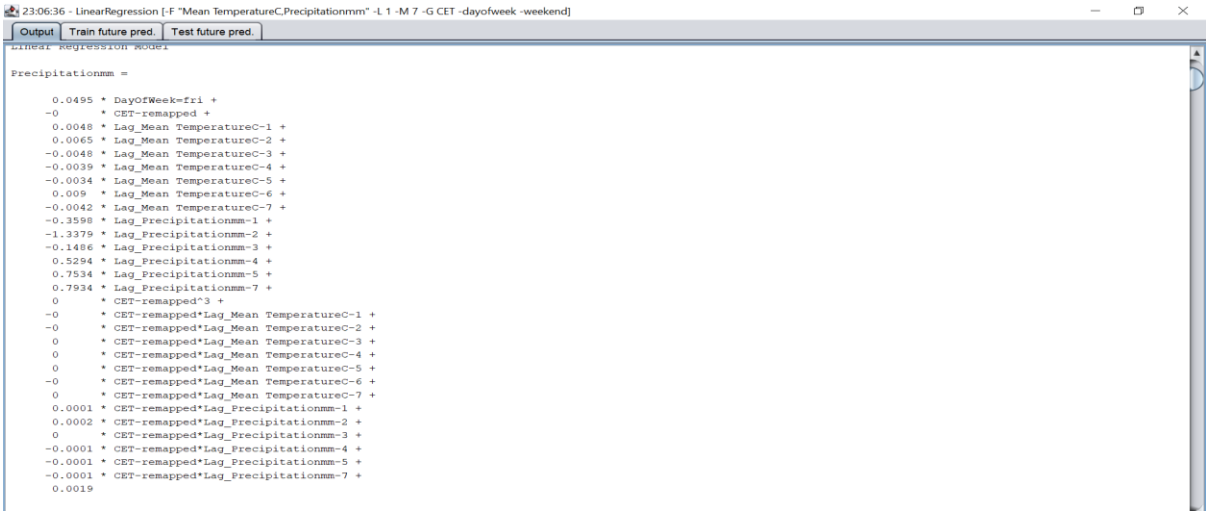
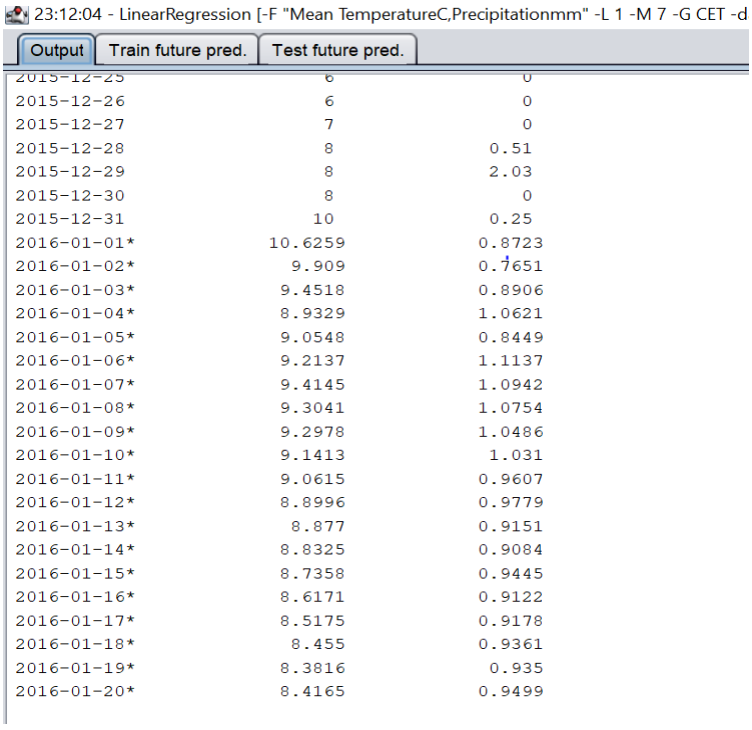


Figure 146 Regression model- Precipitationmm

The regression model generated can be seen above, it is quite large. Day of week in temperature regression model is Wednesday and Thursday and in precipitation it is Friday. There is a lag of 7 in the model.



2015-12-25	6	0
2015-12-26	6	0
2015-12-27	7	0
2015-12-28	8	0.51
2015-12-29	8	2.03
2015-12-30	8	0
2015-12-31	10	0.25
2016-01-01*	10.6259	0.8723
2016-01-02*	9.909	0.7651
2016-01-03*	9.4518	0.8906
2016-01-04*	8.9329	1.0621
2016-01-05*	9.0548	0.8449
2016-01-06*	9.2137	1.1137
2016-01-07*	9.4145	1.0942
2016-01-08*	9.3041	1.0754
2016-01-09*	9.2978	1.0486
2016-01-10*	9.1413	1.031
2016-01-11*	9.0615	0.9607
2016-01-12*	8.8996	0.9779
2016-01-13*	8.877	0.9151
2016-01-14*	8.8325	0.9084
2016-01-15*	8.7358	0.9445
2016-01-16*	8.6171	0.9122
2016-01-17*	8.5175	0.9178
2016-01-18*	8.455	0.9361
2016-01-19*	8.3816	0.935
2016-01-20*	8.4165	0.9499

Figure 147 Predictions

```
=== Evaluation on test data ===
Target                                1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead  7-steps-ahead  8-steps-ahead  9-steps-ahead 10-
=============================================================================================================================================================================
Mean TemperatureC
  N                                            681            680            679            678            677            676            675            674            673
  Mean absolute error                       1.6914         2.1253         2.4515         2.5384         2.6504         2.6784         2.7904         2.8741         2.932
  Root relative squared error             113.9742       137.9107       117.5212       107.1477       105.768        102.5626       103.4788       104.0299       104.4902
  Mean absolute percentage error           14.617         18.9146        21.7873        22.4363        23.914         24.063         25.0021        25.9037        25.8433
  Root mean squared error                   2.2481         2.722          3.0998         3.211          3.4017         3.4758         3.6113         3.7104         3.7906
  Mean squared error                        5.0542         7.4092         9.6089        10.3106        11.5719        12.0814        13.0414        13.767         14.3689
Precipitationmm
  N                                            681            680            679            678            677            676            675            674            673
  Mean absolute error                       1.0244         1.0575         1.0857         1.1054         1.1128         1.1126         1.1133         1.0748         1.0739
  Root relative squared error              76.7559        76.5289        73.2511        74.0274        72.023         72.2434        70.6897        71.3011        69.6521
  Mean absolute percentage error          111.9926       102.5697        90.7714        87.8032        83.9193        83.9817        82.5828        83.0276        83.0033
  Root mean squared error                   2.2941         2.2889         2.2499         2.2361         2.2276         2.2286         2.2272         2.2072         2.2084
  Mean squared error                        5.263          5.2393         5.0621         5.0003         4.9622         4.9668         4.9602         4.8716         4.8772
```

Figure 148 Evaluation

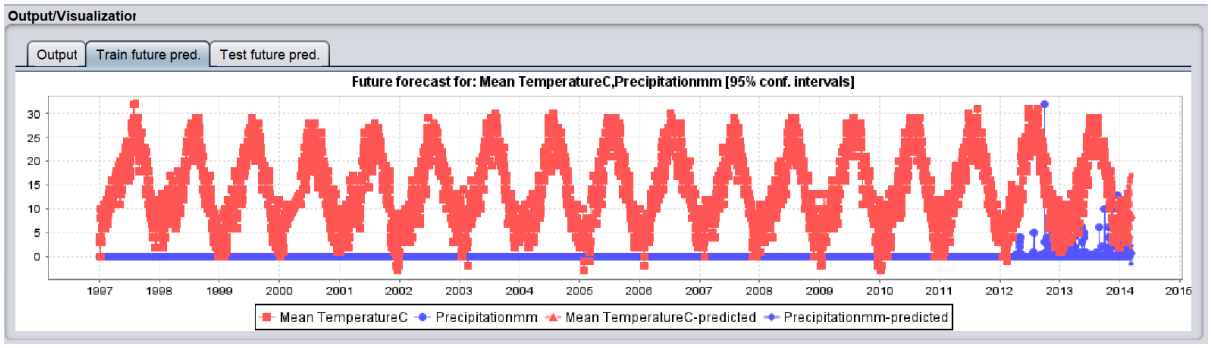


Figure 149 Train future prediction

Graphs of prediction on test data

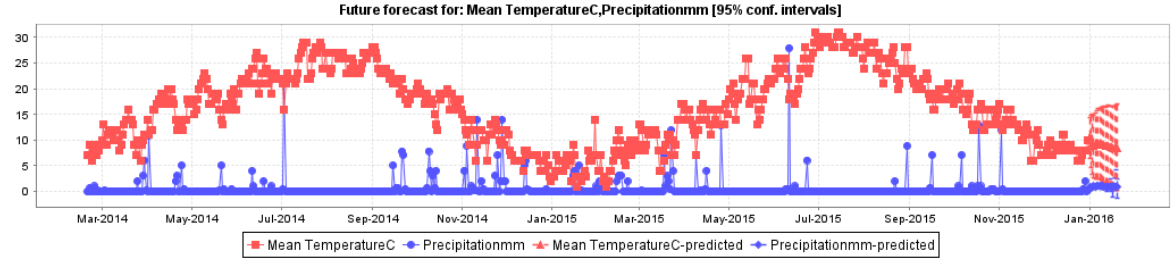


Figure 150 Test future prediction

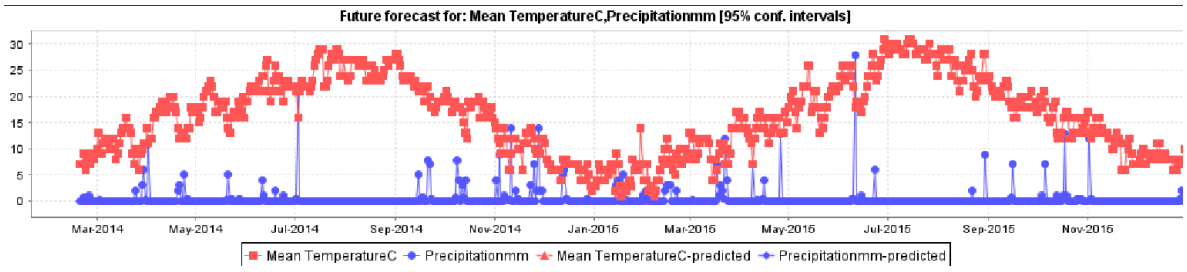


Figure 151 Diagram of historical values

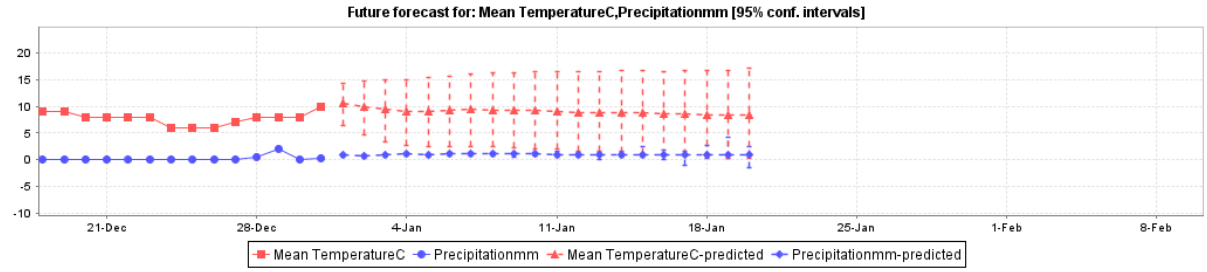


Figure 152 Diagram of future predictions

Findings

1. The weather is forecasted for next 20 days from 01-01-2016 to 20-01-2016.
2. Error for each step is observed, and mean absolute error for temperature after 20 steps is only 3.5279, for precipitation is 1.0673which is really good, this means that forecast is close to accurate.Root mean squared error for temperature is 4.5789 and precipitation is 2.2242
3. It can be seen that prediction graph on training data is more smooth
4. The diagram for historical and future predictions show that, the temperature is increasing slightly the next day and so is precipitation. Overall, the trend for next 20 days is forecast to be stable with little fluctuations.
5. There is a lag of 7 instances which has to be removed in the following experiments. However the error is very low, in the range of 1-3. Propagation of errors can be seen using one step ahead evaluation used.

Experiment 2

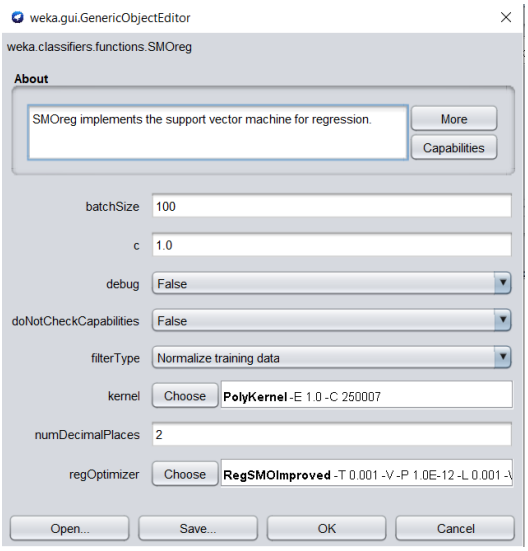


Figure 153 Time series-Experiment 2-parameters

SMO Regression is used here which is Sequential minimal optimization (SMO), an algorithm for solving the quadratic programming problem. It implements Support

Vector Machine. (Sequential minimal optimization, 2020).The default parameters are used with PolyKernel kernel and RegSMOImproved regression optimizer.

Result

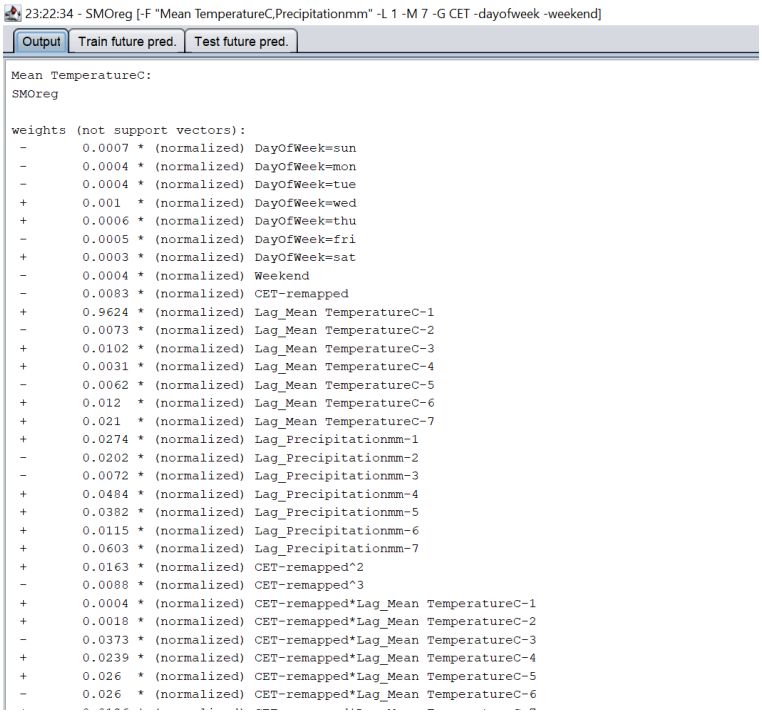


Figure 154Regression Model-Mean temperature

We can see that model generated by regression using normalized values.

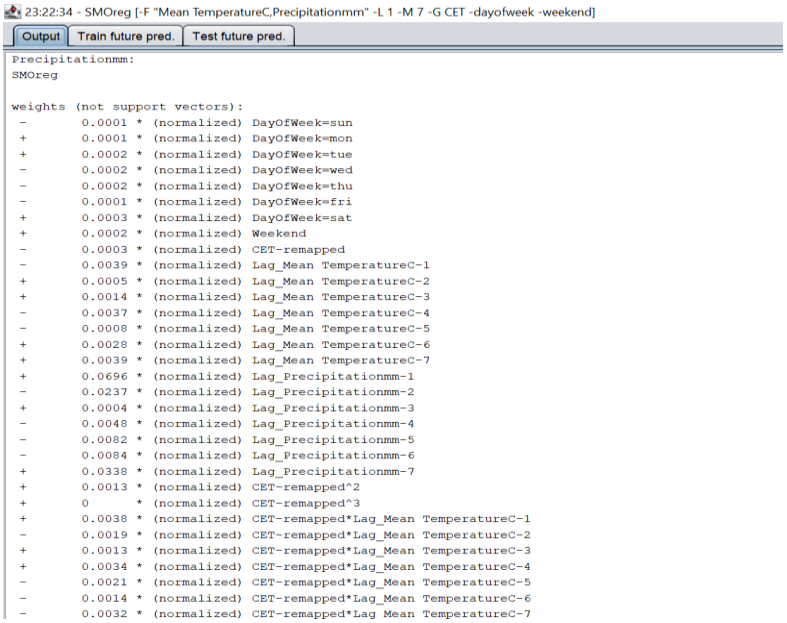


Figure 155 Regression Model - precipitationmm

23:22:34 - SMOreg [-F "Mean TemperatureC,Precipitationmm" -L 1 -M 7 -G CET -dayofweek -weekend]

Output | Train future pred. | Test future pred.

2015-12-18	9	0
2015-12-19	9	0
2015-12-20	8	0
2015-12-21	8	0
2015-12-22	8	0
2015-12-23	8	0
2015-12-24	6	0
2015-12-25	6	0
2015-12-26	6	0
2015-12-27	7	0
2015-12-28	8	0.51
2015-12-29	8	2.03
2015-12-30	8	0
2015-12-31	10	0.25
2016-01-01*	10.0127	0.0607
2016-01-02*	9.9504	0.0522
2016-01-03*	9.7898	0.0515
2016-01-04*	9.6015	0.0495
2016-01-05*	9.7062	0.0292
2016-01-06*	9.716	0.0339
2016-01-07*	9.8128	0.0299
2016-01-08*	9.8462	0.0367
2016-01-09*	9.8929	0.0532
2016-01-10*	9.8994	0.0416
2016-01-11*	9.922	0.0435
2016-01-12*	9.9505	0.0472
2016-01-13*	10.0265	0.0343
2016-01-14*	10.089	0.0326
2016-01-15*	10.1097	0.0375
2016-01-16*	10.1466	0.0532
2016-01-17*	10.1467	0.0421
2016-01-18*	10.1715	0.0437
2016-01-19*	10.1992	0.0472
2016-01-20*	10.2764	0.0344

Figure 156 Predictions

=== Evaluation on test data ===

Target	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead	6-steps-ahead	7-steps-ahead	8-steps-ahead	9-steps-ahead	10-
Mean TemperatureC										
N	681	680	679	678	677	676	675	674	673	
Mean absolute error	1.4938	1.9772	2.3058	2.4344	2.5238	2.6005	2.6954	2.7402	2.8203	
Root relative squared error	100.2468	130.3421	108.9216	101.9244	99.9033	97.6355	97.3487	97.2347	98.1079	
Mean absolute percentage error	13.2982	17.9056	20.7685	21.7464	23.3107	23.7092	24.762	25.8487	26.1972	
Root mean squared error	1.9777	2.5726	2.873	3.0545	3.2131	3.3089	3.3974	3.468	3.5591	
Mean squared error	3.9113	6.6182	8.2541	9.3298	10.3242	10.9485	11.542	12.0272	12.6672	
Precipitationmm										
N	681	680	679	678	677	676	675	674	673	
Mean absolute error	0.5748	0.5692	0.5696	0.5688	0.5695	0.5698	0.5685	0.5696	0.5703	
Root relative squared error	74.6779	75.0522	73.1486	74.4444	72.7654	73.0147	71.5371	72.8381	71.1643	
Mean absolute percentage error	89.4038	95.3045	95.5314	95.8474	95.7115	95.8176	95.9616	95.7403	95.7503	
Root mean squared error	2.2319	2.2448	2.2468	2.2487	2.2506	2.2524	2.2538	2.2547	2.2564	
Mean squared error	4.9816	5.039	5.048	5.0568	5.065	5.0734	5.0798	5.0838	5.0912	

Total number of instances: 681

Figure 157 Evaluation

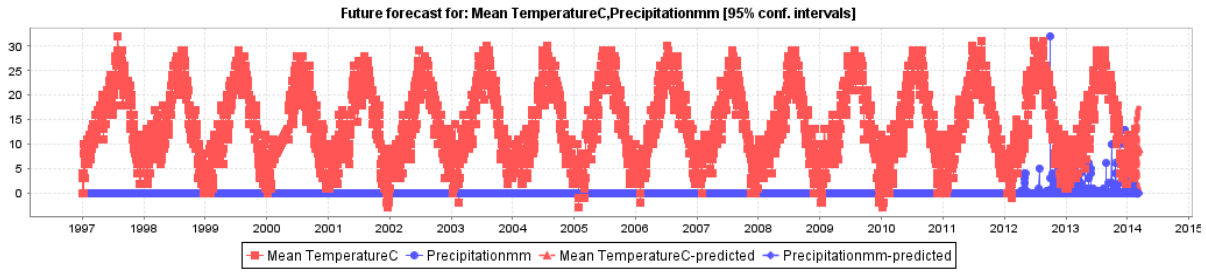


Figure 158 Train future prediction

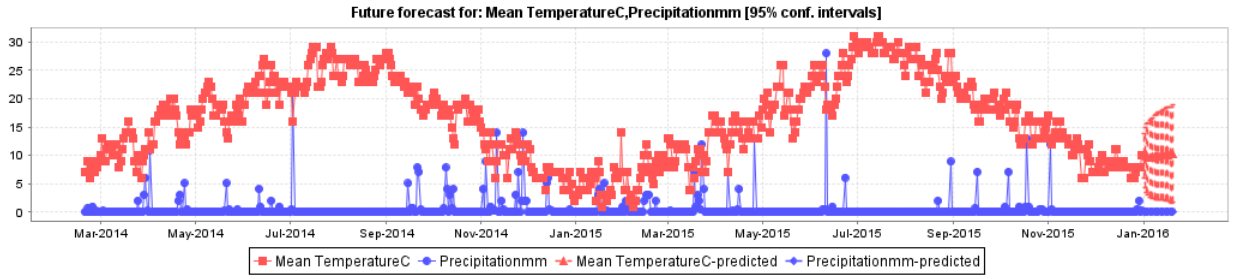


Figure 159 Test future prediction

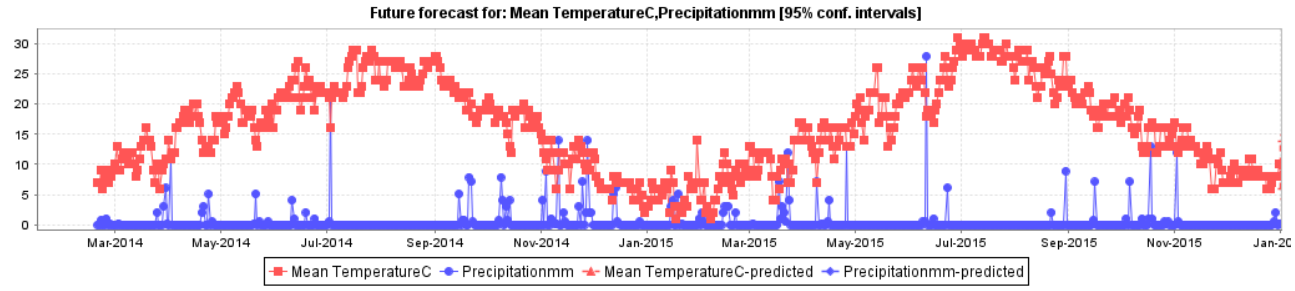


Figure 160 Diagram of historical values

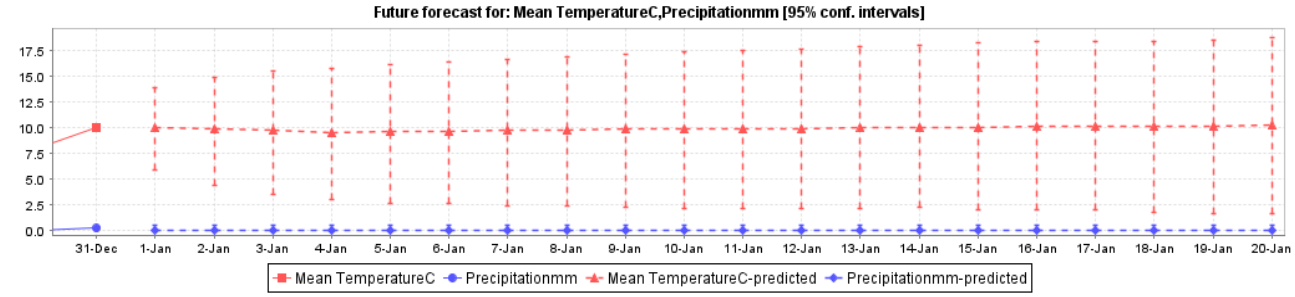


Figure 161 Diagram of future predictions

Findings

1. The mean temperature in Celsius and precipitation in millimetre is forecast for 20 days, and the trend remains to be stable. There is no massive increase or decrease in weather and precipitation values for the next 20 days.
2. However, with SMO Regression, Mean Absolute Error increased by 0.5 to 4.0203 for temperature, and increased to 1.2431 for precipitation.
3. Similarly, Root Mean Squared Error increased to 5.033 for mean temperature and 2.260 for precipitation
4. One step ahead evaluation is used as errors propagate in time series, and in this case mean absolute error propagated from 1.4938 to 4.0203 for temperature.

Experiment 3

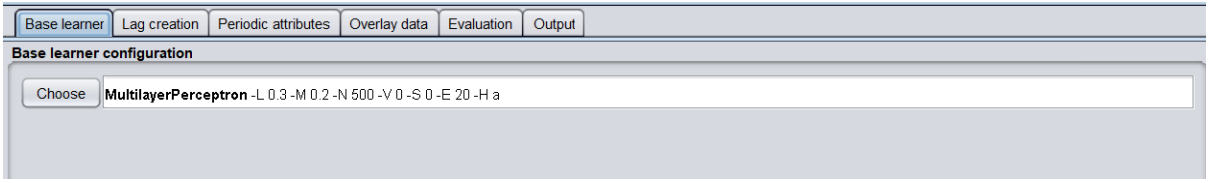


Figure 162 MultilayetPerceptron

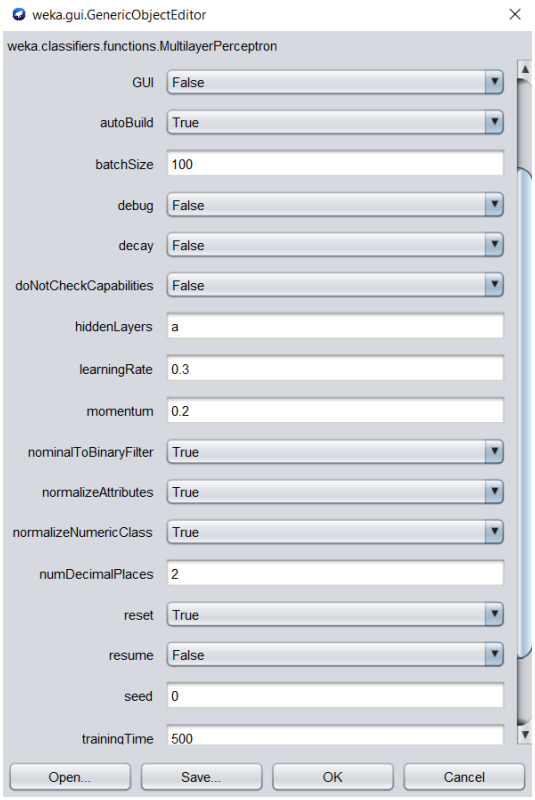


Figure 163 parameters

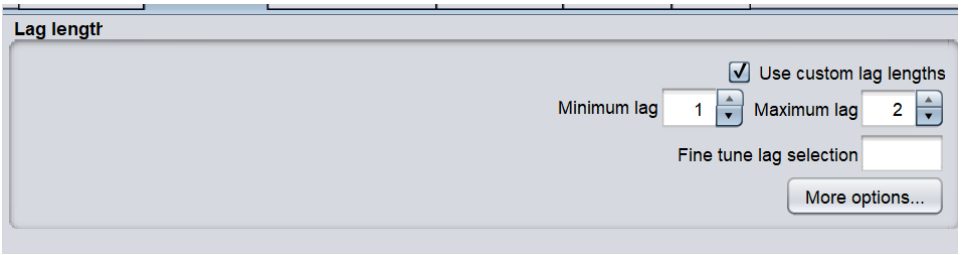


Figure 164 Lag creation

Result

```
Transformed training data:

               Mean TemperatureC
               Precipitationmm
               DayOfWeek
               Weekend
               CET-remapped
               Lag_Mean TemperatureC-1
               Lag_Mean TemperatureC-2
               Lag_Precipitationmm-1
               Lag_Precipitationmm-2
               CET-remapped^2
               CET-remapped^3
               CET-remapped*Lag_Mean TemperatureC-1
               CET-remapped*Lag_Mean TemperatureC-2
               CET-remapped*Lag_Precipitationmm-1
               CET-remapped*Lag_Precipitationmm-2
```

Figure 165 Transformed Training data

2 lags are only produced now as it is given as the maximum lag allowed.

```
Sigmoid Node 1
    Inputs    Weights
    Threshold    -0.9617202767654485
    Attrib DayOfWeek=sun    0.6318318526578136
    Attrib DayOfWeek=mon    0.7073804268669518
    Attrib DayOfWeek=tue    0.6756091482393236
    Attrib DayOfWeek=wed    0.749454248027222
    Attrib DayOfWeek=thu    0.7043407002555492
    Attrib DayOfWeek=fri    0.7247816767824226
    Attrib DayOfWeek=sat    0.5619657209589448
    Attrib Weekend    0.18271650325830746
    Attrib CET-remapped    -0.3380490850562762
    Attrib Lag_Mean TemperatureC-1    5.03839822064865
    Attrib Lag_Mean TemperatureC-2    0.39564134343633967
    Attrib Lag_Precipitationmm-1    -1.0963113430292268
    Attrib Lag_Precipitationmm-2    -0.5092948203649103
    Attrib CET-remapped^2    1.0144962055126316
    Attrib CET-remapped^3    -0.5047525354442166
    Attrib CET-remapped*Lag_Mean TemperatureC-1    0.2148366933330315
    Attrib CET-remapped*Lag_Mean TemperatureC-2    -0.013447094348124577
    Attrib CET-remapped*Lag_Precipitationmm-1    -0.9754035832782126
    Attrib CET-remapped*Lag_Precipitationmm-2    -0.7583524618904578
Sigmoid Node 2
    Inputs    Weights
    Threshold    -0.6052272073938127
    Attrib DayOfWeek=sun    0.5870410641894656
    Attrib DayOfWeek=mon    0.41949271333777705
    Attrib DayOfWeek=tue    0.7695273598525495
    Attrib DayOfWeek=wed    0.5566236988085238
    Attrib DayOfWeek=thu    0.3592873506251209
    Attrib DayOfWeek=fri    0.5091458192870777
    Attrib DayOfWeek=sat    0.11174228788717627
    Attrib Weekend    0.0040639528676743 93
    Attrib CET-remapped    -1.5650448910257901
    Attrib Lag_Mean TemperatureC-1    2.379476339638157
    Attrib Lag_Mean TemperatureC-2    -0.6042422994395241
```

Figure 166 Sigmoid nodes and weights

10:04:19 - MultilayerPerceptron [-F "Mean TemperatureC,Precipitationmm" -trim-leading -L 1 -M 2 -G CET -dayofweel

Output | Train future pred. | Test future pred.

2014-02-10	7	4.06
2014-02-11	3	8.89
2014-02-12	6	0
2014-02-13	11	0
2014-02-14	13	0.25
2014-02-15	8	4.06
2014-02-16	7	0
2014-02-17	4	0
2014-02-18	8	0
2014-02-19*	8.3423	0.6231
2014-02-20*	9.2516	0.4843
2014-02-21*	9.9004	0.5255
2014-02-22*	9.7496	0.854
2014-02-23*	9.0176	0.3639
2014-02-24*	9.4528	0.1728
2014-02-25*	10.0427	0.8377
2014-02-26*	10.6347	0.5403
2014-02-27*	11.0843	0.2927
2014-02-28*	11.3083	0.4227
2014-03-01*	10.9266	0.8068
2014-03-02*	10.0815	0.2525
2014-03-03*	10.256	0.1394
2014-03-04*	10.6743	0.8008
2014-03-05*	11.1343	0.5189
2014-03-06*	11.463	0.2656
2014-03-07*	11.6089	0.405
2014-03-08*	11.1753	0.8093
2014-03-09*	10.297	0.223
2014-03-10*	10.4157	0.132

Figure 167 Future predictions from 2014-02-1

```
=== Evaluation on test data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead  7-steps-ahead  8-steps-ahead  9-steps-ahead 1
==========================================================================================================================================================================
Mean TemperatureC
  N                                        681            680            679            678            677            676            675            674            673
  Mean absolute error                   1.7973         2.9518         4.5155         7.0747         9.4555        11.9784        14.3638        17.1752        20.7439
  Root relative squared error         121.496        217.7837       369.3614       616.4688       764.8944       866.4097       945.5388      1039.6766         1163
  Mean absolute percentage error       15.9924        24.2633        32.7906        43.8954        54.6081        65.5855        76.4668        89.4787       106.5948
  Root mean squared error               2.3968         4.2984         9.7425        18.4744        24.6008        29.3625        32.9982        37.0817        42.1906
  Mean squared error                    5.7446        18.4767        94.9168       341.3017       605.1994       862.1551      1088.8789      1375.0552      1780.045
Precipitationmm
  N                                        681            680            679            678            677            676            675            674            673
  Mean absolute error                   1.847          1.6343         1.5308         1.6154         1.6677         1.8669         1.9355         2.0453         2.1726
  Root relative squared error         106.3186        99.558         91.9355        98.4536       100.9432       123.8265       124.0664       130.5582       133.4612
  Mean absolute percentage error      177.0073       161.1747       136.0515       143.2435       147.1316       146.6705       146.924        147.6684       149.0989
  Root mean squared error               3.1777         2.9777         2.8238         2.974          3.1221         3.8199         3.9088         4.0415         4.2316
  Mean squared error                   10.0978         8.8669         7.9739         8.8445         9.7473        14.5917        15.279         16.3336        17.9064

Total number of instances: 681
```

Figure 168 Evaluation from 1 steps ahead

10-steps-ahead	11-steps-ahead	12-steps-ahead	13-steps-ahead	14-steps-ahead	15-steps-ahead	16-steps-ahead	17-steps-ahead	18-steps-ahead	19-steps-ahead	20-steps-ahead
672	671	670	669	668	667	666	665	664	663	662
24.3321	27.5832	30.4603	32.7313	34.6814	36.9984	38.8481	40.9711	43.2411	46.0178	48.5611
1266.2198	1324.0771	1372.078	1410.5116	1434.323	1474.073	1481.4777	1493.2385	1538.32	1591.404	1630.6771
124.2137	141.0039	155.3404	166.921	178.1843	191.8782	203.4715	214.6046	226.3332	243.6383	259.6952
47.1024	51.2227	54.6985	57.4313	59.2089	61.6926	63.4357	65.4893	67.5067	70.1997	72.581
2218.6317	2623.7608	2991.9261	3298.3587	3505.6954	3805.9801	4024.086	4288.8533	4557.1565	4928.0002	5268.0056
672	671	670	669	668	667	666	665	664	663	662
2.3482	2.5897	2.7405	2.9878	3.011	3.0987	3.1525	3.4237	3.3613	3.5754	3.5792
141.8342	158.2037	163.391	188.6809	180.6909	184.652	184.0909	199.9834	209.6671	205.3576	204.9398
156.0007	165.9255	191.9462	228.6112	224.9507	278.0523	266.723	207.0378	243.8551	236.0648	258.964
4.486	5.0132	5.218	5.7662	5.7662	5.7179	5.6882	6.2479	6.212	6.4156	6.409
20.1244	25.1323	27.2274	33.249	33.2487	32.6948	32.3558	39.0365	38.5886	41.1602	41.0753

Figure 169 Evaluation from 10 steps ahead

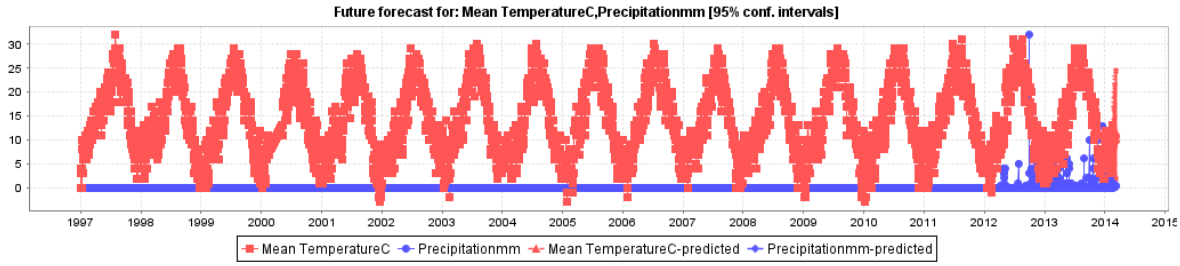


Figure 170 Training future predictions

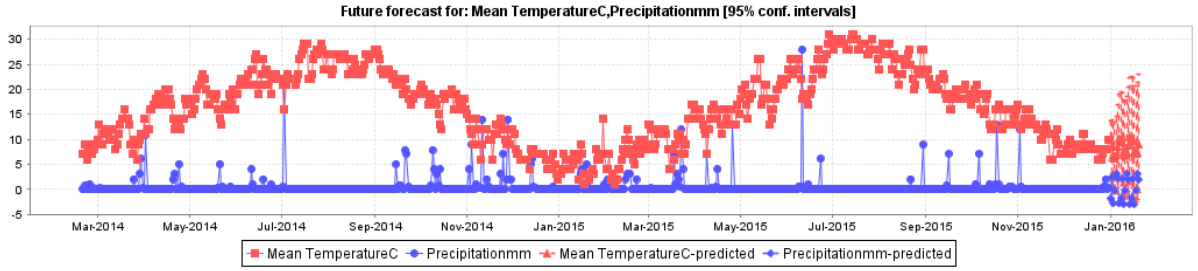


Figure 171 Testing future predictions

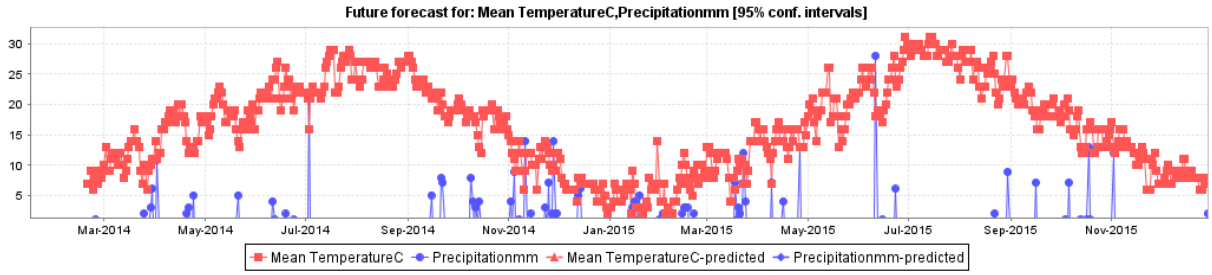


Figure 172 Diagram of historical values

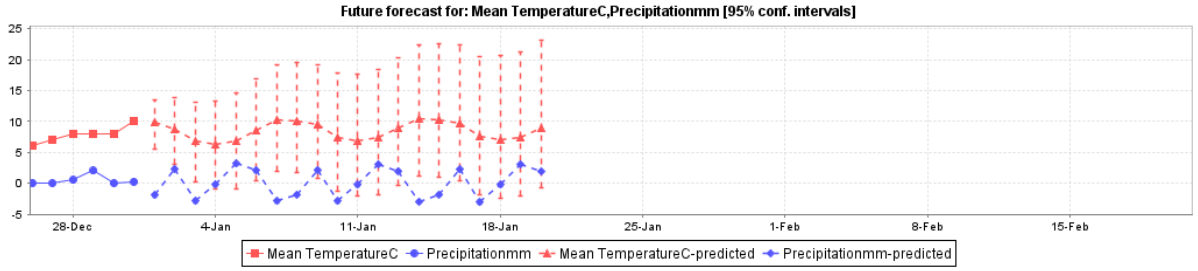


Figure 171 Diagram of future predictions

Findings

1. This experiment, performed the worst with base learner as MultiLayerPerceptron. The neural network didn't work really well with this prediction problem.
2. Mean absolute error increased drastically to 48 from 3 in the previous experiments for mean temperature. However, for precipitation error increased to only 3.579 from 1. 2431.
3. There is a cyclical fluctuating component in predictions which was absent in the previous experiments.
4. MLP model didn't suit the Time Series problem mostly because of overfitting.

Research publication

8. Research Publication Summary and relevance / potential relevance to your work – 20% (2-4 pages)

Please discuss under the following headings

a. Publication and Researchers

1. Name of the publication:

Scaling Up the Accuracy of Naive-Bayes Classifiers- A Decision-Tree Hybrid

2. Conference and Year of Publication

The paper was presented on the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining in the year 1996.

3. Researchers

Ron Kahavi, the researcher of this paper is a Vice President at Airbnb. He conducted this research and published this paper while he was working in Silicon Graphics in Data Mining and Visualization. He previously led the Analysis and Experimentation at Microsoft's Cloud and AI group as Technical Fellow and Corporate Vice President. Prior to Microsoft, he was the director of data mining and personalization at Amazon.com, and the Vice President of Business Intelligence at Blue Martini Software. He joined Silicon Graphics after getting a Ph.D. in Machine Learning from Stanford University before joining Silicon Graphics.

Three of his papers are in the top 1,000 most cited articles in computer science, including the article Wrappers for Feature Subset Selection, which is in the top 300.His papers have over 45,000 citations andin 2016, he was named the 5th most influential scholar in AI and the 26th most influential scholar in Machine Learning.He started the MLC++ project, Machine Learning library in C++, which formed the basis for SGI's MineSet. Ron Kohavi is also the member of General Chair, KDD in the year 2000. He is also the co-author of the book "Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing". (Ron Kohavi, 2020)

b. Dataset

Scaling Up the Accuracy of Naive-Bayes Classifiers- A Decision-Tree Hybrid research paper utilised several datasets. Adult Census Dataset is one of them. The datasets that are utilised for this research are:

Dataset	No attrs	Train size	Test size	Dataset	No attrs	Train size	Test size	Dataset	No attrs	Train size	Test size
adult	14	30,162	15,060	breast (L)	9	277	CV-10	breast (W)	10	683	CV-10
chess	36	2,130	1,066	cleve	13	296	CV-10	crx	15	653	CV-10
DNA	180	2,000	1,186	flare	10	1,066	CV-10	german	20	1,000	CV-10
glass	9	214	CV-10	glass2	9	163	CV-10	heart	13	270	CV-10
ionosphere	34	351	CV-10	iris	4	150	CV-10	led24	24	200	3000
letter	16	15,000	5,000	monk1	6	124	432	mushroom	22	5,644	3,803
pima	8	768	CV-10	primary-tumor	17	132	CV-10	satimage	36	4,435	2,000
segment	19	2,310	CV-10	shuttle	9	43,500	14,500	soybean-large	35	562	CV-10
tic-tac-toe	9	958	CV-10	vehicle	18	846	CV-10	vote	16	435	CV-10
vote1	15	435	CV-10	waveform-40	40	300	4,700				

Figure 173 Datasets used and attributes in the research, Source: Scaling up the Accuracy of Naive-Bayes Classifiers- Research Paper

Most of these datasets are famous and are publicly available. Some of these datasets are available in UCI Machine Learning Repository. These datasets are always used in Machine Learning research and is cited very often. Some of the most popular datasets are the adult census data, pima diabetes data, mushroom, breast cancer data, chess, German credit data etc. Few of these datasets are described below.

1. **Adult Census Dataset**

This is the same dataset that is being discussed in this Assignment and is used for Classification and Clustering. Detailed description is available in the Data Description section of this assignment.

2. **Breast Cancer Wisconsin Dataset**

This is a very popular dataset in the field of machine learning for diagnostics. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.(Breast Cancer Wisconsin (Diagnostic) Data Set, 2020)

3. **Pima Indian Diabetes Dataset**

This dataset is originally from the National Diabetes and Digestive and Kidney Diseases Center. The purpose of the dataset is to predict diagnostically whether a patient has diabetes or not, based on certain diagnostic measurement values included in the dataset. Several constraints have been imposed on selecting such instances from a larger database. All of the patients here are female. (Pima Indians Diabetes Database, 2020)

4. **Mushroom Dataset**

The observations are obtained from The Audubon Society Field Guide to North American Mushrooms. This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Individual species is identified as definitely edible, definitely poisonous, or of unknown edibility and not suitable for consumption. This latter class was merged with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three,

let it be'' for Poisonous Oak and Ivy.(UCI Machine Learning Repository: Mushroom Data Set, 2020)

5. German Credit Dataset

German Credit Dataset classifies people described by a set of attributes as good or bad credit risks. Comes in two formats (one all numeric). This dataset also comes with a cost matrix. This dataset is widely cited and forms part of many notable researches such as "Genetic Programming for data classification: partitioning the search space" and "Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection" etc.

c. Technique (mention any adaptions)

This research introduced the famous NBTree algorithm which is very popular in the Knowledge Discovery process.

NBTree is a hybrid of decision tree classifiers and Naïve Bayes Classifiers. This algorithm improves the performance of Naïve Bayes on large datasets because the early Naïve Bayes Algorithms were attaining desired performance only with small datasets. This technique utilises the merits of trees such as segmentation and, thespeciality of Naïve Bayes algorithms which is gaining information from many attributes. The research paper reviews several Naïve Bayes induction algorithms.

The NBTree algorithm is very similar to recursive partitioning algorithms with few exceptions.

The major points, as presented in the research paper are:

1. A threshold for continuous attributes is chosen using the standard entropy minimization technique , similar tothe one used in decision-trees.
2. The use of a node is calculated by binning the data and calculating the 5-fold cross-validation accuracy estimate of using Naive- Bayes at the node.
3. The utility of a split is the weighted sum of the utility of the nodes, where the weight given to a node is proportional to the number of instances that go down to that node.
4. Direct Cross-validation is not supported. However, for discretized data, cross-validation can be performed linearly.

Steps in the NBTreealgorithm as presented in the paper

Input: a set 2 'of labelled instances. Output: a decision-tree with naive-bayes categorizers at the leaves.

1. For each attribute Xi, evaluate the utility, u(Xi), of a split on attribute Xi. For continuous attributes, a threshold is also found at this stage.
2. Let j = argmaxi(ui), i.e., the attribute with the highest utility.
3. If uj is not significantly better than the utility of the cur- rent node, create a Naive-Bayes classifier for the current node and return.
4. Partition T according to the test on Xj. If Xj is con- tinuous, a threshold split is used; if Xj is discrete, a multi-way split is made for ail possible values.
5. For each child, call the algorithm recursively on the por- tion of T that matches the test leading to the child.

d. Major Findings

Experiments were conducted on all of the datasets shown in the table above which are files from UC Irvine Repositoy and there are some interesting findings of these experiments.

The average accuracy for C4.5 was 81.91%, for Naive-Bayes 81.69%, and for NBTree it is 84.47% which is really good. The findings are not only about increasing the prediction rate but also decreasing the errors. For example, the largest dataset used was the shuttle dataset. Absolute difference was only 0.04% between NBTree and C4.5, however error drastically reduced from 0.05% to 0.01%.

Other experiments found that, NBTree produced much less nodes. the number of nodes induced by NBTree was in many cases significantly smaller than that of C4.5.For example, for the letter dataset, C4.5 induced 2109 nodes while NBTree induced only 251; in the adult dataset, C4.5 induced 2213 nodes while NBTree induced only 137; for DNA, C4.5 induced 131 nodes and NBTree induced 3; for led24, C4.5 induced 49 nodes, while NBTree used a single node.

Thus, NBTrees are much easier to interpret because of less nodes even though NBTree is much complex.

e. Relevance / potential relevance to your work

The findings from NBTree are relevant for the classification problem in this assignment, i.e to predict people with income exceeding 50k US Dollars or less than 50K US Dollars. For this assignment, we have used a small sample of the Adult Census Dataset so the accuracy results are not directly comparable. But performance of J48 and NBTree for the classification can be understood. NBTree scales up accuracy in some cases as it follows a hybrid approach. The combination of Naïve Bayes and decision trees for the classification problem is expected to produces better results with lesser nodes than J48 algorithm. Weka documentation provides details about the NBTree class and also cites the research paper discussed here.

public class **NBTree**
extends Classifier
implements WeightedInstancesHandler, Drawable, Summarizable, AdditionalMeasureProducer, TechnicalInformationHandler

This is the NBTree class in Weka Class for generating a decision tree with Naive Bayes classifiers at the leaves(Generated Documentation (Untitled), 2020).An experiment was performed on Weka to see how NBTree algorithm works with the resampled Adult Dataset used in the assignment.

NBTree Experiment

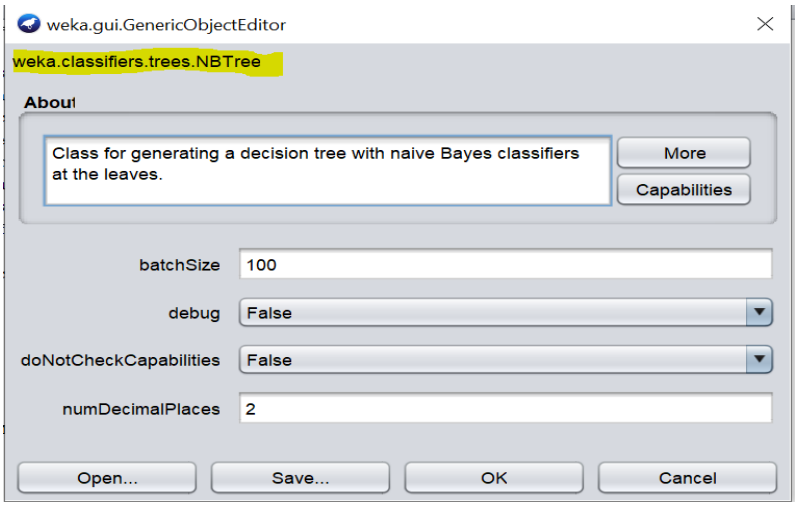


Figure 174NBTree Experiment

Result

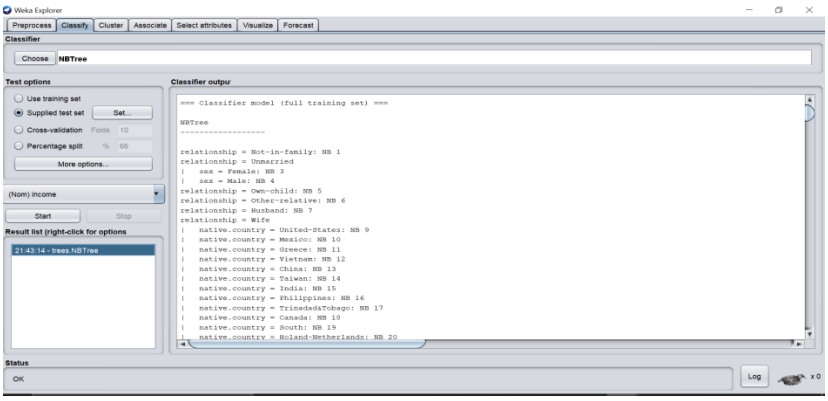


Figure 175NBTree model

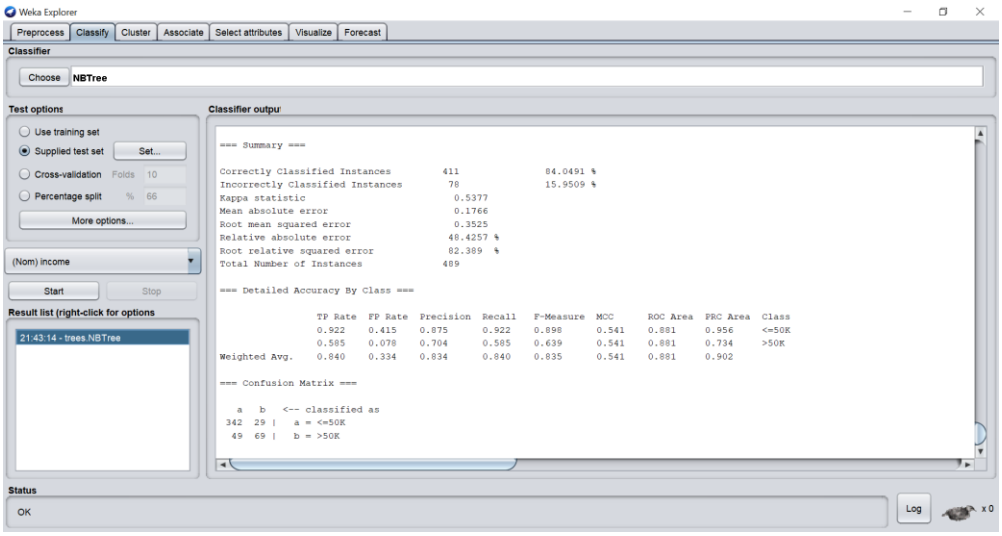


Figure 176 Output summary

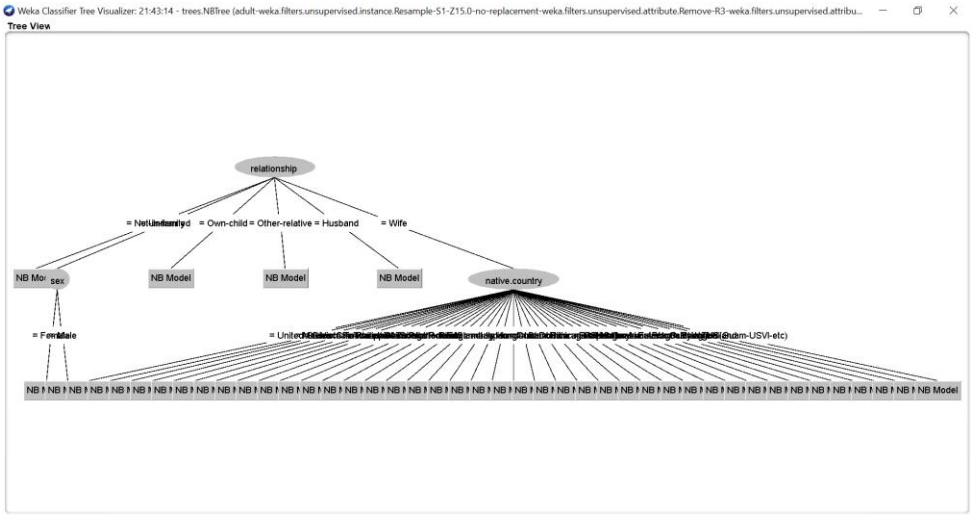


Figure 177 Tree visualization

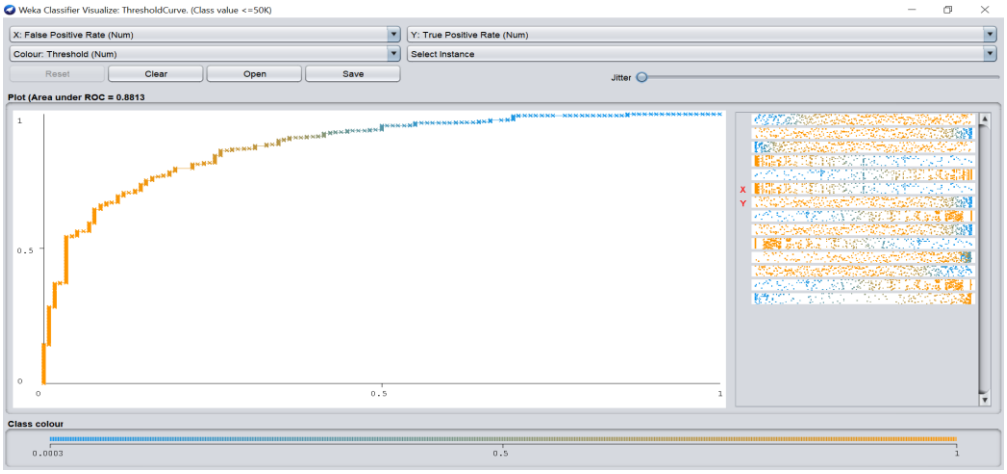


Figure 178 Threshold curve for <=50K

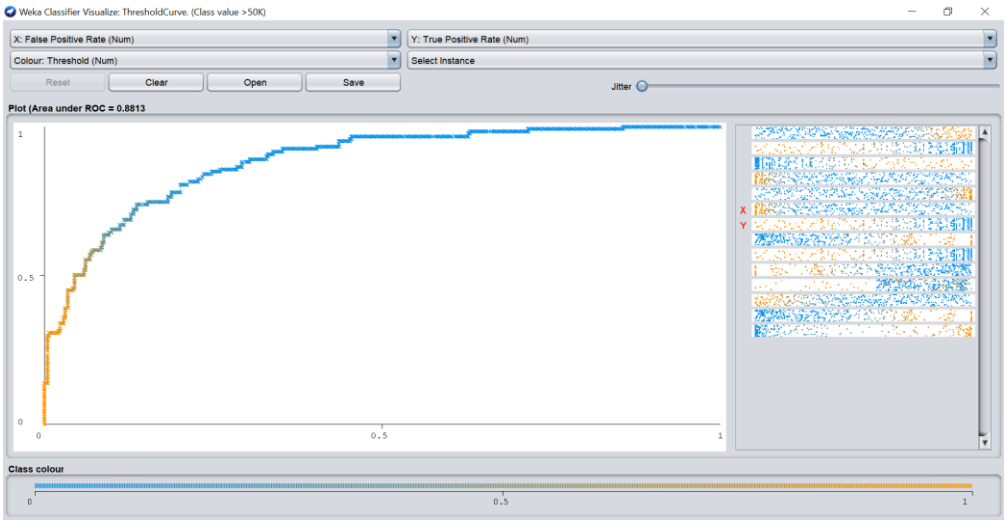


Figure 179Threshold curve for >50K

Findings from this experiment.

1. Accuracy obtained with NBTree with the test dataset is 84. 0491 which is really impressive as the average of all three J48 experiments conducted in this dataset is
2. The tree produced is really small and is easy to interpret in comparison to the J48 tree.
3. Mean absolute error has reduced in comparison to J48.

After the experiment and the review of the dataset, it is understood that NBTree algorithm has desirable performance with large datasets. It is much quicker to execute and easier to interpret. The research paper is thus very much useful for the classification problem that is discussed in this assignment.

Division of Labor

9. Division of Labor

Please complete the sections below with regard to the estimate of the division of work between the two partners

Summary of division of work

Work was evenly divided.

Percentage of work completed by each partner on each class / task

Some area requires more work than others so this is only for reference. An average of these values will not be calculated.

Please ensure that both students contribute to the research. The remaining topics can be divided among the students, but each student must apply at least 1 techniques.

Filename / Task	Subhasree Vadukoot	Minna GeorgeKaiprambadan
Selection of dataset	50%	50%
Cleaning of dataset	50%	50%
Classification	50%	50%
Clustering	50%	50%
Time Series	50%	50%
Paper selection	50%	50%
Paper Review	50%	50%

References

1. A. Siddig, Data Mining Algorithms and Techniques Clustering, D., 2020. Data Mining Algorithms And Techniques -Clustering.
2. A. Siddig, Data Mining Algorithms and Techniques Clustering, D., 2020. Data Mining Algorithms And Techniques -Decision Tree 2
3. A. Siddig, Datasets, Datasets,EDA and altering data structure D., 2020. Data Mining Algorithms And Techniques - Datasets,EDA and altering data structure
4. A. Siddig, Datasets, Datasets,EDA and altering data structure D., 2020. Data Mining Algorithms And Techniques Time Series Lecture
5. Archive.ics.uci.edu. 2020. *UCI Machine Learning Repository: Adult Data Set*. [online] Available at: <http://archive.ics.uci.edu/ml/datasets/adult> [Accessed 26 April 2020].
6. Bureau, U., 2020. *Census.Gov*. [online] Census.gov. Available at: <https://www.census.gov/> [Accessed 26 April 2020].
7. Cs.ccsu.edu. 2020. *Computer Science*. [online] Available at: <http://www.cs.ccsu.edu/~markov/ccsu_courses/dataminingex3.html> [Accessed 3 May 2020].
8. David, L. (2019). If Climate Change Figures Big in NASA's Next Earth Science Roadmap, Thank This Guy. [online] SpaceNews.com. Available at: https://spacenews.com/if-climate-change-figuresbig-in-nasas-next-earth-science-roadmap-you-can-thank-this-guy/ [Accessed 29 Dec. 2019].
9. Dictionary.cambridge.org. 2020. *INCOME | Meaning In The Cambridge English Dictionary*. [online] Available at: <https://dictionary.cambridge.org/dictionary/english/income> [Accessed 24 April 2020].
10. Dl.acm.org. 2020. *Privacy Preserving OLAP | Proceedings Of The 2005 ACM SIGMOD International Conference On Management Of Data*. [online] Available at: <https://dl.acm.org/doi/10.1145/1066157.1066187> [Accessed 26 April 2020].
11. En.wikipedia.org. 2020. *Associated Press*. [online] Available at: <https://en.wikipedia.org/wiki/Associated_Press> [Accessed 6 May 2020].
12. En.wikipedia.org. 2020. *Sequential Minimal Optimization*. [online] Available at: <https://en.wikipedia.org/wiki/Sequential_minimal_optimization> [Accessed 6 May 2020].
13. En.wikipedia.org. 2020. *The Weather Company*. [online] Available at: <https://en.wikipedia.org/wiki/The_Weather_Company> [Accessed 6 May 2020].
14. En.wikipedia.org. 2020. *United States Census Bureau*. [online] Available at: <https://en.wikipedia.org/wiki/United_States_Census_Bureau> [Accessed 23 April 2020].
15. En.wikipedia.org. 2020. *Weather Underground (Weather Service)*. [online] Available at: <https://en.wikipedia.org/wiki/Weather_Underground_(weather_service)> [Accessed 6 May 2020].
16. En.wikipedia.org. 2020. *Weather Underground (Weather Service)*. [online] Available at:

<https://en.wikipedia.org/wiki/Weather_Underground_(weather_service)> [Accessed 6 May 2020].
17. Hannon, P., 2020. *How The Coronavirus Might Reduce Income Inequality*. [online] WSJ. Available at: <https://www.wsj.com/articles/how-the-coronavirus-might-reduce-income-inequality-11587304801> [Accessed 21 April 2020].
18. Inequality.org. 2020. *Gender Economic Inequality - Inequality.Org*. [online] Available at: <https://inequality.org/gender-inequality/> [Accessed 25 April 2020].
19. Kaggle.com. 2020. *Breast Cancer Wisconsin (Diagnostic) Data Set*. [online] Available at: <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data> [Accessed 26 April 2020].
20. Kaggle.com. 2020. *Pima Indians Diabetes Database*. [online] Available at: <https://www.kaggle.com/uciml/pima-indians-diabetes-database> [Accessed 26 April 2020].
21. Kaggle.com. 2020. *Weather Madrid 1997 - 2015*. [online] Available at: <https://www.kaggle.com/juliansimon/weather_madrid_lemd_1997_2015.csv#weather_madrid_LEMD_1997_2015.csv> [Accessed 6 May 2020].
22. Kohavi, R., 1997. *Scaling Up The Accuracy Of Naive-Bayes Classifiers: A Decision-Tree Hybrid*. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/2669468_Scaling_Up_the_Accuracy_of_Naive-Bayes_Classifiers_a_Decision-Tree_Hybrid> [Accessed 6 May 2020].
23. NASA (2019). The Causes of Climate Change. [Blog] Global Climate Change-NASA. Available at: https://climate.nasa.gov/causes/ [Accessed 27 Dec. 2019].
24. ResearchGate. 2020. *(PDF) A Statistical Approach To Adult Census Income Level Prediction*. [online] Available at: <https://www.researchgate.net/publication/328494313_A_Statistical_Approach_to_Adult_Census_Income_Level_Prediction> [Accessed 24 April 2020].
25. Robotics.stanford.edu. 2020. *Ron Kohavi*. [online] Available at: <http://robotics.stanford.edu/~ronnyk/> [Accessed 26 April 2020].
26. UK Government, 2020. [online] Available at: <https://www.youtube.com/watch?v=mo2dqHbLpQo> [Accessed 28 May 2020].
27. UK Government. 2020. [online] Available at: <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/articles/analysisoffactorsaffectingearningsusingannualsurveyofhoursandearnings/01> [Accessed 20 April 2020].
28. UN News. (2019). COP25: UN climate change conference, 5 things you need to know. [online] Available at: https://news.un.org/en/story/2019/12/1052251 [Accessed 29 Dec. 2019].
29. Wunderground.com. 2020. *Local Weather Forecast, News And Conditions | Weather Underground*. [online] Available at: <https://www.wunderground.com/> [Accessed 6 May 2020].