International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2019.33565

Volume 5, Issue 10
October - 2019

Data Mining Approaches in the Study of the Nigerian Informal Sector

ERNEST Paul[1] and OWOLABI Olumide[2]

Research Scholar[1] and Lecturer[2]

Department of Computer Science,

University of Abuja, Abuja,

Nigeria

Abstract

The vast amounts of economic data currently being generated and collected calls for the application of new methodologies in order to make sense of them. It is for this reason that we have undertaken the analysis of some of data collected on the informal sector of the Nigerian economy using data mining and machine learning techniques. This is different from the traditional bare statistical methods hitherto being used in such analysis. In this work, we used data gathered by the National Salaries, Incomes and Wages Commission (NSIWC) on the informal sector of the Nigerian economy between the year 2014 and 2016. These data were subjected to analysis using WEKA data mining/machine learning tools and the Random Forest ensemble algorithm. The results show that the modal population age group of Nigerians working in the informal sector is in the age bracket of 30-44 years. Similarly, it was discovered that wholesale business is the dominant activity in the informal sector and women's participation in informal sector business is still low compared to their male counterparts. This study shows the relevance or implication of utilizing data mining methodologies over the conventional statistical analysis of data.

Key Words: *Data Mining, NSIWC, WEKA, Machine learning, Forest ensemble algorithm, Informal Sector*.

1. INTRODUCTION

Our capabilities of both generating and collecting data have been increasing rapidly in the last several decades. Contributing factors include the widespread use of bar codes for most commercial products, the computerization of many business, scientific and government transactions and managements, and advances in data collection tools ranging from scanned texture and image platforms, to on-line instrumentation in manufacturing and shopping, and to satellite remote sensing systems. In addition, popular use of the World Wide Web as a global information system has flooded us with a tremendous amount of data and information. This explosive growth in stored data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. The abundance of data, coupled with the need for powerful data analysis tools, has been described as a “data rich but information poor" situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. The steady and amazing progress of computer hardware technology in the past three decades has led to powerful, affordable, and large supplies of computers, data collection equipment, and storage media. This technology provides a great boost to the database and information industry, and makes a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis [1].

The data we collect these days in such huge amounts could be useful if we had the ability to organize and transform them in such a way that they provide useful insights into our everyday activities and also provide us with useful guides and options for our future actions. While technological advances have solved the problems of storage and organization of data on a large scale, gaps remain to be filled with respect to the transformation of the data and the extraction of useful information from them. These gaps are currently being filled by Data Mining and Machine Learning methodologies.

Licensed Under Creative Commons Attribution CC BY-NC

Data Mining (DM) is the process of discovering various models, derived values and summaries from a given collection of data. Data Mining deals with the problem of extracting features from data so as to solve many different predictive tasks. It is important to realize that the problem of discovering or estimating dependencies from data or discovering new data is only one part of the general experimental procedure used by researchers and others who apply standard steps to draw conclusions from data [2].

The economic growth and structural transformation of any nation is dependent on several factors such as access to finance, young population, literacy level, state of infrastructure and informal sector practices, among others.

The Nigerian economic growth faces various supply constraints including infrastructure decay, entry barriers and business unfriendly regulations. In addition, there is shortage of requisite skills, appropriate technology mix and access to finance/capital outlay to drive growth.

Applying Data Mining (DM) and Machine Learning (ML) techniques, this study seeks to reveal the modal population age group, enterprise sectors of activities, literacy level and gender distribution of Nigerians working in the informal employment. The outcome of this study will serve as statistical input for macro-economic policy framework for the Nigerian government to help create jobs, increase social inclusion, achieve sustainable diversification of production/ manufacturing and improve the citizens/human capital base of the economy.

2. STATEMENT OF THE PROBLEM

Over the years in developing countries like Nigeria, there has been a steady increase in the number of working population absorbed by the informal sector. This makes it a key sector in the economy. Informal sector operators engage in enterprises that come under the purview of rules characterized by free entry, free exit, perfect market knowledge and market force price determination without regulations, as is the case of perfect competitive market. Such businesses are also characterized by no formal organizational structures, low and irregular earnings, business insecurity and lack of social welfare schemes such as pensions and other palliatives, for both the employer and the employee. Understanding the operational dynamics of the informal sector has become a central issue for achieving inclusive growth and sustainable development in the new global economy. With half of working-age adults globally or 500 million households, mostly in informal economy, having no access to formal financial services, recent developments in the Nigerian financial system have particularly led to a renewed interest in desirable policies and programmes for enhancing financial accessibility of informal business operators [3].

Over the past decades, Nigeria has grown her economy yet many still live below the poverty line with large swaths of the poplation working in low-productivity jobs, often in the informal sector. Many citizens of the country are unable **to acquire the foundational skills they need** to thrive and prosper in an increasingly competitive global economy. The National Salaries, Incomes and Wages Commission (NSIWC), is an agency of federal government of Nigeria charged with responsibility, among others, of providing the Federal Government, on a continuous basis, with information on the movement of all forms of incomes, employment and other socio-economic data for national planning. The NSIWC, in recognition of the immense contributions of the informal sector to the Nigerian economy and to generate data for the informal sector, conducted a study of the informal sector to obtain data which will aid decision making processes in the areas of both Incomes policies and Compensation policies in Nigeria. The survey datasets obtained in that exercise is what is to be analyzed, interpreted and presented in this study using Data Mining (DM) tools and Machine Learning (ML) algorithms.

3.AIM AND OBJECTIVES OF THE STUDY

The aim of this project is to apply Data Mining (DM) methodologies to data collected by the National Salaries, Incomes and Wages Commission (NSIWC) on the informal sector of the economy in order to extract useful information that will assist the Nigerian government in its economic recovery and growth strategy.

The objectives are to:

(a) Apply the various data mining methodologies and machine learning (ML) algorithms in order to discover and extract patterns in the collected data of the Nigerian Informal Economic Sector.

(b) Develop data prediction model that correctly or accurately predicts future data by analyzing historical data. Finalize a model for making predictions on new data and predicting the estimated accuracy of the model on future or unseen data

(c) Analyze, interpret and present results.

DOI: 10.31695/IJASRE.2019.33565

4. LITERATURE REVIEW

Data Mining represents a complex of technologies that are rooted in many disciplines: mathematics, statistics, computer science, physics, engineering, biology, etc., and with diverse applications in a large variety of different domains: business, health care, science and engineering, etc. Basically, data mining can be seen as the science of exploring large datasets for extracting implicit, previously unknown and potentially useful information, [4].

As Big Data takes center stage for business operations, data mining becomes something that salespeople, marketers, and C-level executives need to know how to do and do well. Generally, data mining is the process of finding patterns and correlations in large data sets to predict outcomes. There are a variety of techniques to use for data mining, but at its core are statistics, artificial intelligence, and machine learning. Companies and organizations are using data mining to get the insights they need about pricing, promotions, social media, campaigns, customer experience, and a plethora of other business practices. [1] describes data mining software that allow the users to analyze data from different emerging technique with the help of this one can efficiently learn with historical data and use that knowledge for predicting future behavior of concern areas. Growth of current education system is surely enhanced if data mining has been adopted as a futuristic strategic management tool. The Data Mining tool is able to facilitate better resource utilization in terms of human performance.

Several researchers have reviewed key contributions of data mining or analytic methods to public policy decision processes or delivery in the areas of energy planning, urban transportation planning, medical emergency planning, healthcare, social services, national security, defence, government finance allocation, understanding public opinion, and fire and police services [5].

Compared with their counterparts in the private sector, public sector decision-makers face several challenges. In particular, public sector problems typically involve making decisions for society at large. Indeed, policy makers in the public sector confront difficulties in deciding how public resources are to be allocated, since the whole underlying purpose of public policy and associated politics is about “deciding who gets what, when and how” [5].

4.1 Applications of Data Mining

Starting from the marketing forecast for large transnational companies and passing through the trend analysis of shares trading on the main Stock Exchanges, identification of the loyal customer profile, modeling demand for pharmaceuticals, automation of cancer diagnosis, bank fraud detection, hurricanes tracking, there are a growing demand for services provided by the data mining applications to the economic, medical, engineering, production and financial market.

(i) **Fraud detection** (e.g., in credit card transactions) is used to avoid as much as possible fraudulent use of bank cards in commercial transactions. For this purpose, available information on the use of cards of different customers is collected (e.g., typical purchases by using the card, how often, location, etc.). The labels illustrating the way the card is used (illegally, fair) are added, to complete the dataset for training.
After training the model to distinguish between the two types of users, the next step concerns its validation and, finally, its application to real-world data. Thus,the issuing bank can track, for instance, the fairness of transactions with cards by tracing the evolution of a particular account.

(ii) **Economics (business-finance)** - there is a huge amount of data already collected
in various areas such as: Web data, e-commerce, super/hypermarkets data, financial and banking transactions, etc., ready for analyzing in order to take optimal decisions

(iii) **Retail Stores** – use Data Mining to predict individual or customer behavior, customer prospecting, development of cross-selling and marketing strategies.

(iv) Government all over the world use Data Mining technology to forecast tax receipt and spending

(v) Venture capital firms use it to forecast market potentials and to evaluate business plans

(vi) Educational institutions apply Data Mining techniques to predict students enrollment and retention

(vii) Transportation and airline companies deploy DM techniques/applications to schedule flights, and forecast future flights

(viii) International financial organizations like the World Bank and International Monetary Fund utilize DM applications to forecast inflation and economic activities.

4.2 Related works

The term informal economy as, it is often used to denote informal sector, refers to all economic activities by workers and economic units that are not covered or are insufficiently covered by formal arrangements. The informal economy is largely characterized by: low entry requirements in terms of capital and professional qualifications; small scale of operations; skills often acquired outside of formal education; and, labour-intensive methods of production and adapted technology [6].
Virtually everywhere in the world, the informal economy is efficient and resilient, creating jobs. It is a major provider of

DOI: 10.31695/IJASRE.2019.33565

employment, goods and services for lower-income groups. It contributes significantly to GDP. Average incomes are lower in the informal economy than in the formal economy. A higher percentage of people working in the informal economy are micro entrepreneurs who hire others. The poorest are, typically, informal wageworkers, especially industrial outworkers.
The role of informal sectors on the level of unemployment and poverty level cannot be underrated in the world economy. According to [7], informal sector is the second largest employer of labour after agriculture in the world including Nigeria. The finding also indicated that the informal sectors is a refuge ground for retrenched workers from the formal sector, unemployed youth and fresh university graduates in the third world countries especially Nigeria. For many people in the developing world, the informal economy is an important source of livelihoods for individuals, families, and communities. In many developing countries informality is associated with a way of life, moulded by custom and tradition, than with a conscious decision to remain outside the formal sector"

In [8], the paper suggests the use of data-mining technique (Decision tree algorithm technique) to extract hidden information from large census data warehouse and geographical information system (GIS) as an integrating technology that gives geo-spatial distribution of the population. Decision tree learning is a method for approximating discrete-valued target function, in which the leaned function is represented by a decision tree. Decision tree algorithm was used to predict some basic attributes of the population in the census database. This study is an effort towards harnessing the power of data-mining technique to develop mining model applicable to the analysis of census data that could uncover some hidden patterns to get their geo-spatial distribution. **T**his work achieved the goal of applying data-mining technique to the analysis of census data. The result of this paper is a predictive attributes of a population to give geo-spatial distribution in Nigeria. The effort yielded the possibility of implementing the IDE3 decision algorithm in building decision trees from which attributes of a population can be predicted to give geo-spatial distribution.
The report [9] implements the CRISP-DM methodology which applies classification models to the problem of identifying individuals whose salary exceeds a specified value based on demographic information such as age, level of education and current employment type. The process involved the exploration, preparation, modelling and evaluation of the datasets.
The data mining objective of the study was to create a classification model which could predict

individuals whose salary exceeds fifty thousand US dollars by mining anonymised census data containing demographic information such as age, gender, education level and employment type. The original salary attribute in the census data had been anonymised to a binomial value indicating if a salary exceeds fifty thousand US dollars or not.

5. METHODOLOGY

5.1 Data description

To understand the nature of the data is very important in bringing to a successful data analysis. The data must be understood before applying data preprocessing techniques to extract more meaningful knowledge from a given data set. Therefore, the volume of data, the meanings of individual attributes and description of the initial format of the data are defined. The dataset for this study was provided by National Salaries, Incomes and Wages Commission (NSIWC), a federal government agency. It is a proprietary dataset comprising of data collected from the surveys conducted in some parts of the country. In this study, the International Standard Industrial Classification (ISIC) of all economic activities (Revision 3) was adopted. In applying the ISIC, the groups were slightly modified to accommodate the peculiar structure of informal sector enterprises operating in Nigerian economy. In the study, the following categories of economic activities were recognized:

(i) Agriculture
(ii) Manufacturing
(iii) Electricity, Gas and Water supply
(iv) Building and Construction
(v) Wholesale and Retail Trade, and Restaurant and Hotels
(vi) Transport, Storage and Communication
(vii) Community, social and Personal Services

The survey covered thirteen states and the Federal Capital Territory. Two states each was selected randomly for the study from each geo-political zone but considering the special status of both Lagos and the Federal Capital Territory, the two have to be co-opted into the study using judgmental discretion of their socio-economic and political importance in decision making process. In each of these activities, the information on the enterprise principal activities were collected. The principal activity of an economic enterprise is the activity that contributes most to the value added of the enterprise, or the activity the value of which exceeds that of any activity of the enterprise.

DOI: 10.31695/IJASRE.2019.33565

Many variables were involved, some of which were out rightly irrelevant. Data editing focused on detection of possible errors and outliers, correction and verification of data. Outliers were later removed and treated as missing data. Therefore, the initial phase of the project was an exploratory data analysis, including data preparation and understanding. We note that omitting important features in the training data hurts the learning performance, and on the other hand, including redundant ones leads to overfitting. Prior to analysis, a significant amount of processing and data cleaning was required, including feature selection, converting the format of the survey results for ingestion by WEKA software, mapping the survey results to survey questions. Next, the datasets for this study was saved as .arff (Attribute-Relation File Format) file. An .arff file is an ASCII text file that describes a list of instances sharing a set of attributes.

It is useful to be able to summarize a large data set and present it at a high conceptual level. Quality data improves the quality of data mining models. A good preprocessing process is needed to extract useful information from the data as well. This process can improve the performance of the model.

5.2 Choice of Tool

For data mining it is important to have good tools which combine advanced modeling technology with ease-of-use. Good tools help the discovery of interesting and valuable relationships within the data. The choice of data mining tool must be based on the application domain and its supported features. In this research, modeling process will be realized through WEKA Workbench which is one of the most widely used, efficient data mining tools.

One of the open source software designed for data analysis and knowledge discovering is WEKA. WEKA or Waikato Environment for Knowledge Analysis software is product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. It uses the GNU General Public License (GPL). The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results. It also has a general API, so WEKA can be embedded in other applications like any other library. WEKA has several standard data mining tasks, data preprocessing, clustering, classification, association visualization, and feature selection, [10]. Weka is one of the best tool to implement data mining concept, which has inbuilt data preprocessing tools and learning algorithms.

5.3 Choice of Machine Algorithm for the Problem

It is a well-known fact that no single machine learning scheme or algorithm is appropriate to all data mining problems. The universal learner is an idealistic fantasy so; to obtain accurate models the bias of the learning algorithm must match the structure of the domain. Data mining is an experimental science, [11].

5.3.1 Algorithm Accuracy

We compared all of the algorithm results to one base result. This was specified by the Test base option. The default is the first algorithm evaluated in the list, which in this case is ZeroR. We can see this by clicking the Select button next to Test base. Click the Perform test button in the Actions pane to perform the statistical test and produce some output we can review. See results in Fig 5.1

DOI: 10.31695/IJASRE.2019.33565

```
Tester:          weka.experiment.PairedCorrectedTTester  -G  4,5,6  -D  1  -R  2  -S  0.05  -result-matrix
"weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -
mean-width  0  -stddev-width  0  -sig-width  0  -count-width  5  -print-col-names  -print-row-names  -enum-col-
names"

Analysing:  Percent_correct
Datasets:   1
Resultsets: 8
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       04/07/19 11:56


Dataset                              (1) rules.Ze | (2) rules (3) bayes (4) funct (5) trees (6) lazy. (7)
meta. (8) trees
-----------------------------------------------------------------------------------------------------------
----
Rapid-incomes-26032018-fi(100)   52.54 |   54.24 v   20.45 *   55.95 v   83.03 v   86.91 v   80.54 v
90.90 v
-----------------------------------------------------------------------------------------------------------
----
                                                    (v/ /*) |   (1/0/0)   (0/0/1)     (1/0/0)   (1/0/0)
(1/0/0)       (1/0/0)   (1/0/0)

Key:
(1) rules.ZeroR '' 48055541465867954
(2) rules.OneR '-B 6' -3459427003147861443
(3) bayes.NaiveBayes '' 5995231201785697655
(4) functions.SimpleLogistic '-I 0 -M 500 -H 50 -W 0.0' 7397710626304705059
(5) trees.J48 '-C 0.25 -M 2' -217733168393644444
(6) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -
R
     first-last\\\"\"' -      3080186098777067172
(7) meta.Bagging '-P 100 -S 1 -num-slots 1 -I 10 -W trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0'
-115879962237199703
(8) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698

v - indicates that the result is significantly more/better than base classifier
* - indicates that the result is significantly less/worse than base classifier
```

Figure 5.1: Algorithm Comparison Results based on Classification accuracy.

The classification performance measures are calculated using a 10-fold cross validation methodology. In this experimentation methodology the dataset is first partitioned into 10 roughly equal-sized distinct subsets. For each experiment nine subsets were used for training and the one part is used for testing. This procedure was repeated for 10 times for each of the 8 model or algorithm types. Fig 5.1 is the algorithm comparison table with the results for 8 algorithms.

Each algorithm was run 10 times on the dataset and the accuracy reported is the mean and the standard deviation in rackets of those 10 runs.

We can see that SimpleLogistic,J48,k-Nearest Neighbors or KNN (lazy.IBk), meta.Bagging and trees.RandomForest algorithms all have a little **"v"** next to their results.

This means that the difference in the accuracy for these algorithms compared to ZeroR is statistically significant. We can also see that the accuracy for these algorithms compared to ZeroR is high,so we can say that these five algorithms achieved a statistically significantly better result than the ZeroR baseline. We can see that ZeroR, our base for comparison marked as (1) has the accuracy of 52.54% on the problem. This result is compared to the other 7 algorithms and indicated with a number and mapped in a legend at the bottom below the table of results. Note the * next to the bayes.NaiveBayes result.

This indicates that the result is significantly different from the ZeroR results, but the scores is lower. OneR does not have any character next to its result in the table, indicating that the results are not significantly different from ZeroR. If an algorithm had results larger than the base algorithm and the difference was significant, a little **v** would appear next to the results.

The value in brackets at the beginning of the **Rapid-incomes-26032018-fi row (100)** is the number of experimental runs: 10 times tenfold cross-validation

DOI: 10.31695/IJASRE.2019.33565

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
"weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -
mean-width 0 -stddev-width 0 -sig-width 0 -count-width 5 -print-col-names -print-row-names -enum-col-
names"

Analysing:  Percent_correct
Datasets:   1
Resultsets: 8
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       04/07/19 11:55

>-<   >   < Resultset
  7   7   0 trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
  5   6   1 lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A
                \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -3080186098777067172
  3   5   2 trees.J48 '-C 0.25 -M 2' -217733168393644444
  1   4   3 meta.Bagging '-P 100 -S 1 -num-slots 1 -I 10 -W trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L -1
-I 0.0' -115879962237199703
 -1   3   4 functions.SimpleLogistic '-I 0 -M 500 -H 50 -W 0.0' 7397710626304705059
 -3   2   5 rules.OneR '-B 6' -3459427003147861443
 -5   1   6 rules.ZeroR '' 48055541465867954
 -7   0   7 bayes.NaiveBayes '' 5995231201785697655
```

Figure 5.2: Summary of Algorithm Performance based on Rankings

5.3.2 Algorithm Ranking

We further analyzed and compared the performance of different machine learning algorithms by Ranking. Fig 5.2 is the ranking table showing the number of statistically significant wins each algorithm has had against all other algorithms on the dataset.

A win, means an accuracy that is better than the accuracy of another algorithm and that the difference was statistically significant. The first column of numbers shows the number of wins/losses by each algorithm while the second row depicts the actual Ranking of each algorithm. It can be seen that RandomForest algorithm has 7 wins, followed by lazy.lBk(KNN) which has5 wins. In the same vein, ZeroR suffered 5 losses (-5) followed by NaiveBayes with 7 losses (-7). Thus, RandomForest algorithm is considered the best contender for our experiment/study outperforming other algorithms in the comparison experiment. Similarly, NaiveBayes algorithm suffered statistically significant losses, thereby making it the worst machine learning algorithm for the purpose of this study.

Therefore, the choice of the ensemble machine learning technique of Random Forests to design our model is predicated on the fact that Random Forest outperforms other machine learning techniques in the comparison experiment with the highest classification accuracy and Random Forests are popular for its ability of not over-fitting when handling a large number of inputs.

6. ANALYSIS OF THE SURVEY DATA AND INTERPRETATION OF RESULTS

6.1 Data Pre-processing

Data cleaning is the process of removing noise and correcting inconsistencies in data. Missing values were handled and the data was checked for outliers and other inconsistencies. To begin, the correlation coefficient for each attribute was determined. The correlation coefficient is a number which measures the interdependence of two random variables or attributes. The correlation coefficient ranges between -1 and +1. Having a value of -1 means the two variables are in perfect negative correlation with each other. Having a value of 0 means the two variables have no correlation with each other.

6.2 Attributes Selection Procedure in the Weka Explorer

1. Open the Weka GUI Chooser and then the Weka Explorer.
2. Load the data/ INFORMAL-SECTOR INCOMES-SURVEY-DATA.arff dataset.
3. Click the "Select attributes" tab.
4. Click the "Choose" button in the "Attribute Evaluator" pane and select the "ClassifierAttributeEval". You will be presented with a dialog asking you to change to the "Ranker" search method, needed when using this feature selection method. Click the "Yes" button.
5. Click the "Start" button to run the feature selection method.

DOI: 10.31695/IJASRE.2019.33565

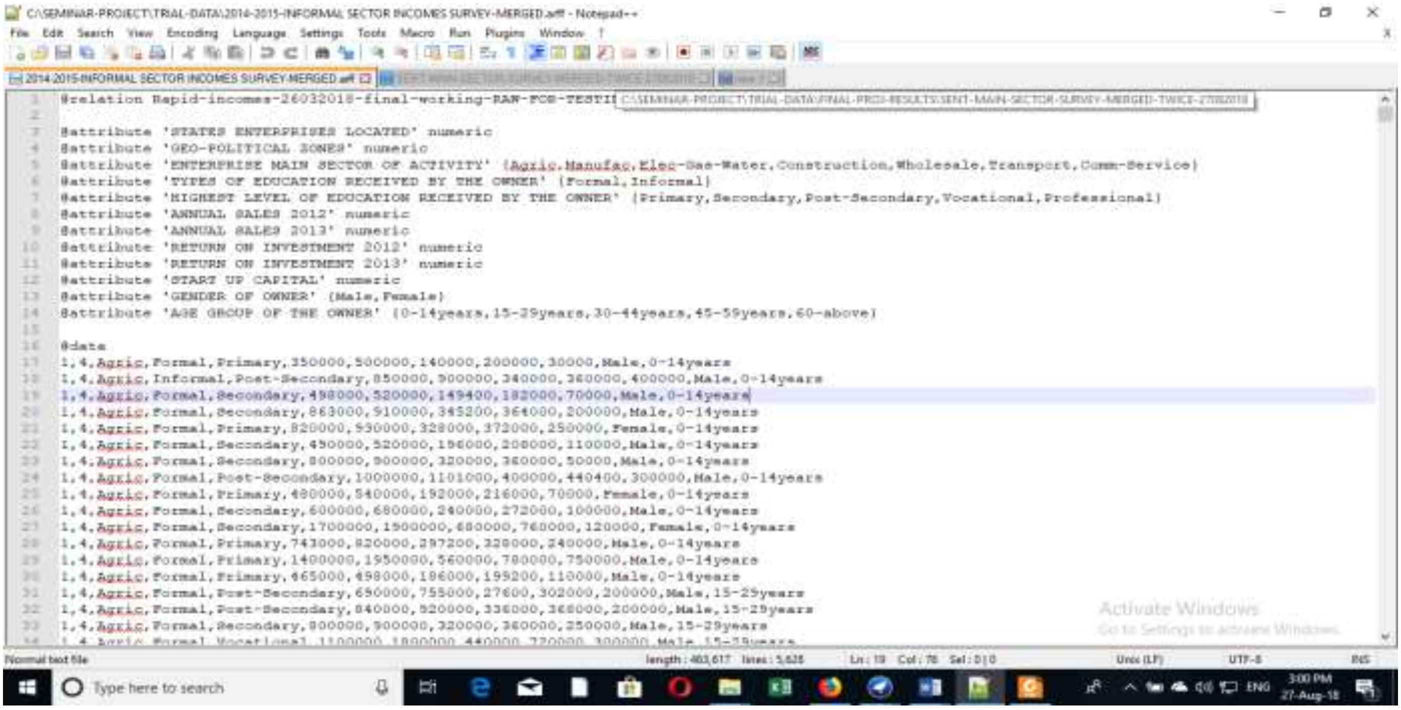
```
=== Run information ===

Instances:    5611
Attributes:   12
=== Attribute Selection on all input data ===
Search Method:Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 12 AGE GROUP OF THE OWNER):
        Classifier feature evaluator
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5
Ranked attributes:
 0.004099    3 ENTERPRISE MAIN SECTOR OF ACTIVITY
 0          11 GENDER OF OWNER
 0           2 GEO-POLITICAL ZONES
 0           5 HIGHEST LEVEL OF EDUCATION RECEIVED BY THE OWNER
 0           4 TYPES OF EDUCATION RECEIVED BY THE OWNER
 0           1 STATES ENTERPRISES LOCATED
-0.000891    8 RETURN ON INVESTMENT 2012
-0.003832   10 START UP CAPITAL
-0.004366    6 ANNUAL SALES 2012
-0.189628    9 RETURN ON INVESTMENT 2013
-0.198539    7 ANNUAL SALES 2013
Selected attributes: 3,11,2,5,4,1,8,10,6,9,7 : 11
```

Figure 6.1: Attribute Selection by Ranking

The dataset was then prepared in a format suitable for the Weka sofware i.e., in ARFF file format. We constructed a data set of **5611 instances with 12 features**, trained and tested them to predict gender of enterprise owner, age group of enterprise owner and enterprise main sector of activity.

We used WEKA Explorer to classify the training dataset, applying Random Forest algorithm with the goal to maximize Classification accuracy (percentage of instances classified correctly).


C:\SEMINAR-PROJECT\TRIAL-DATA\2014-2015-INFORMAL SECTOR INCOMES SURVEY-MERGED.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
2014-2015-INFORMAL SECTOR INCOMES SURVEY-MERGED.arff
C:\SEMINAR-PROJECT\TRIAL-DATA\FINAL-PROJ-RESULTS\SENT-MAIN-SECTOR-SURVEY-MERGED-TWICE-27082018
Normal text file
length : 403,617 lines : 5,628
Ln : 19 Col : 78 Sel : 0 | 0
Unix (LF)
UTF-8
INS
Type here to search
ENG
3:00 PM
27-Aug-18
Activate Windows
Go to Settings to activate Windows.

```
@relation Rapid-incomes-26032018-final-working-RAN-FOR-TESTII

@attribute 'STATES ENTERPRISES LOCATED' numeric
@attribute 'GEO-POLITICAL ZONES' numeric
@attribute 'ENTERPRISE MAIN SECTOR OF ACTIVITY' {Agric,Manufac,Elec-Gas-Water,Construction,Wholesale,Transport,Comm-Service}
@attribute 'TYPES OF EDUCATION RECEIVED BY THE OWNER' {Formal,Informal}
@attribute 'HIGHEST LEVEL OF EDUCATION RECEIVED BY THE OWNER' {Primary,Secondary,Post-Secondary,Vocational,Professional}
@attribute 'ANNUAL SALES 2012' numeric
@attribute 'ANNUAL SALES 2013' numeric
@attribute 'RETURN ON INVESTMENT 2012' numeric
@attribute 'RETURN ON INVESTMENT 2013' numeric
@attribute 'START UP CAPITAL' numeric
@attribute 'GENDER OF OWNER' {Male,Female}
@attribute 'AGE GROUP OF THE OWNER' {0-14years,15-29years,30-44years,45-59years,60-above}

@date
1,4,Agric,Formal,Primary,350000,500000,140000,200000,30000,Male,0-14years
1,4,Agric,Informal,Post-Secondary,850000,900000,340000,360000,400000,Male,0-14years
1,4,Agric,Formal,Secondary,498000,520000,149400,182000,70000,Male,0-14years
1,4,Agric,Formal,Secondary,863000,910000,345200,364000,200000,Male,0-14years
1,4,Agric,Formal,Primary,820000,930000,328000,372000,250000,Female,0-14years
1,4,Agric,Formal,Secondary,490000,520000,196000,208000,110000,Male,0-14years
1,4,Agric,Formal,Secondary,800000,900000,320000,360000,50000,Male,0-14years
1,4,Agric,Formal,Post-Secondary,1000000,1101000,400000,440400,300000,Male,0-14years
1,4,Agric,Formal,Primary,480000,540000,192000,216000,70000,Female,0-14years
1,4,Agric,Formal,Secondary,600000,680000,240000,272000,100000,Male,0-14years
1,4,Agric,Formal,Secondary,1700000,1900000,680000,760000,120000,Female,0-14years
1,4,Agric,Formal,Primary,743000,820000,297200,328000,240000,Male,0-14years
1,4,Agric,Formal,Primary,1400000,1950000,560000,780000,750000,Male,0-14years
1,4,Agric,Formal,Primary,465000,498000,186000,199200,110000,Male,0-14years
1,4,Agric,Formal,Post-Secondary,690000,755000,27600,302000,200000,Male,15-29years
1,4,Agric,Formal,Post-Secondary,840000,920000,336000,368000,200000,Male,15-29years
1,4,Agric,Formal,Secondary,800000,900000,320000,360000,250000,Male,15-29years
1,4,Agric,Formal,Vocational,1100000,1800000,440000,720000,300000,Male,15-29years
```

Figure 6.2: Screenshot of an .arff file

The datasets for this study was cleaned, pre-processed and saved as .arff (Attribute-Relation File Format) file. An .arff file is an ASCII text file that describes a list of instances sharing a set of attributes.

The performance summary is provided in Weka when a model is evaluated. In the Classify tab after an algorithm is evaluated by clicking the Start button, the results are presented in the Classifier output pane. This pane includes a lot of information, including:

- ➢ The run information such as the algorithm and its configuration, the dataset and its properties as well as the test option

DOI: 10.31695/IJASRE.2019.33565

- The details of the constructed model, if any.
- The summary of the performance including a host of different measures.

6.3 The Basic Random Forest Training Algorithm

Algorithm

The training algorithm for random forests applies the general technique of bootstrap aggregating,
or bagging, to tree learners.

1. Draw a random bootstrap sample of size n (randomly choose n samples from the training set with replacement).

2. Grow a decision tree from the bootstrap sample. At each node:

1. Randomly select d features without replacement.

2. Split the node using the feature that provides the best split according to the objective function, for instance, by maximizing the information gain.

3. Repeat the steps 1 to 2 k times.

4. Aggregate the prediction by each tree to assign the class label by majority vote in the case of classification trees or by averaging the predictions from all the individual regression trees. [12].

The Random Forest algorithm creates an ensemble model of decision trees. Each tree is trained on a randomly selected subset of the training data. This supervised learning method has a number of applications, including:

- Predicting genetic outcomes
- Financial analysis
- Medical diagnosis

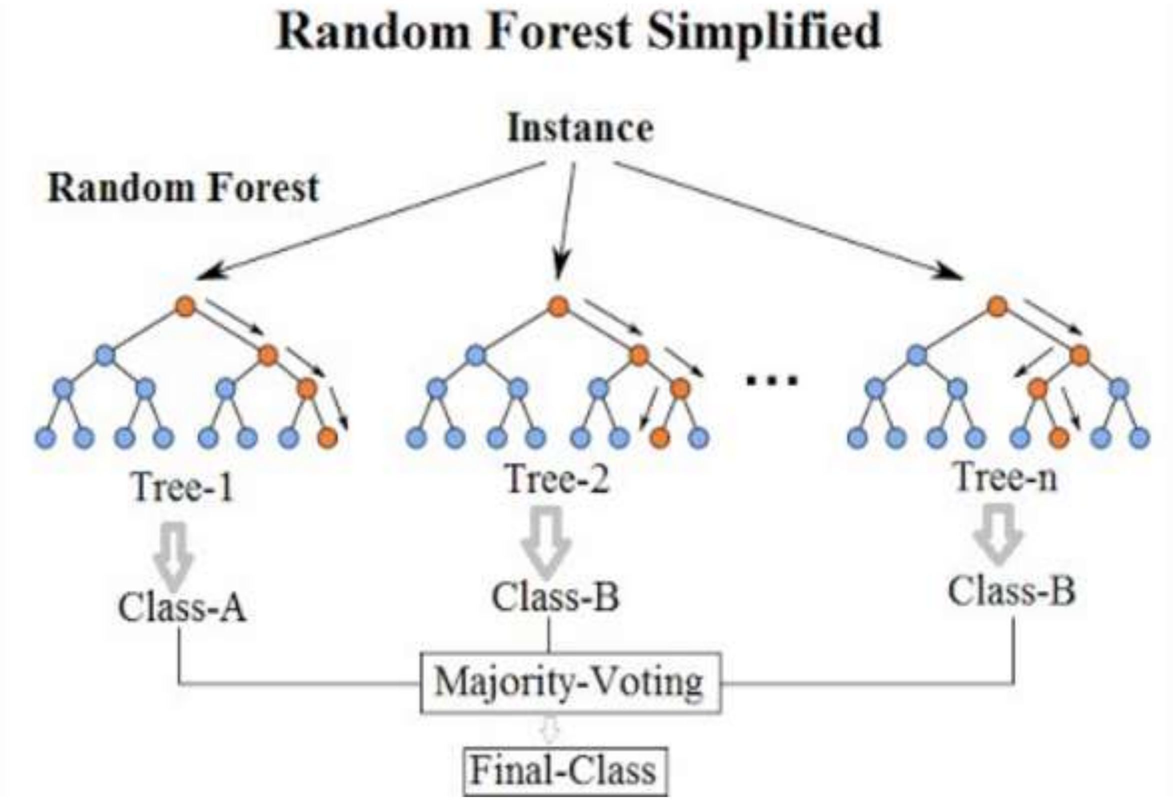


Figure 6.3: Random Forest Simplified

6.4 Experimentation and Evaluation

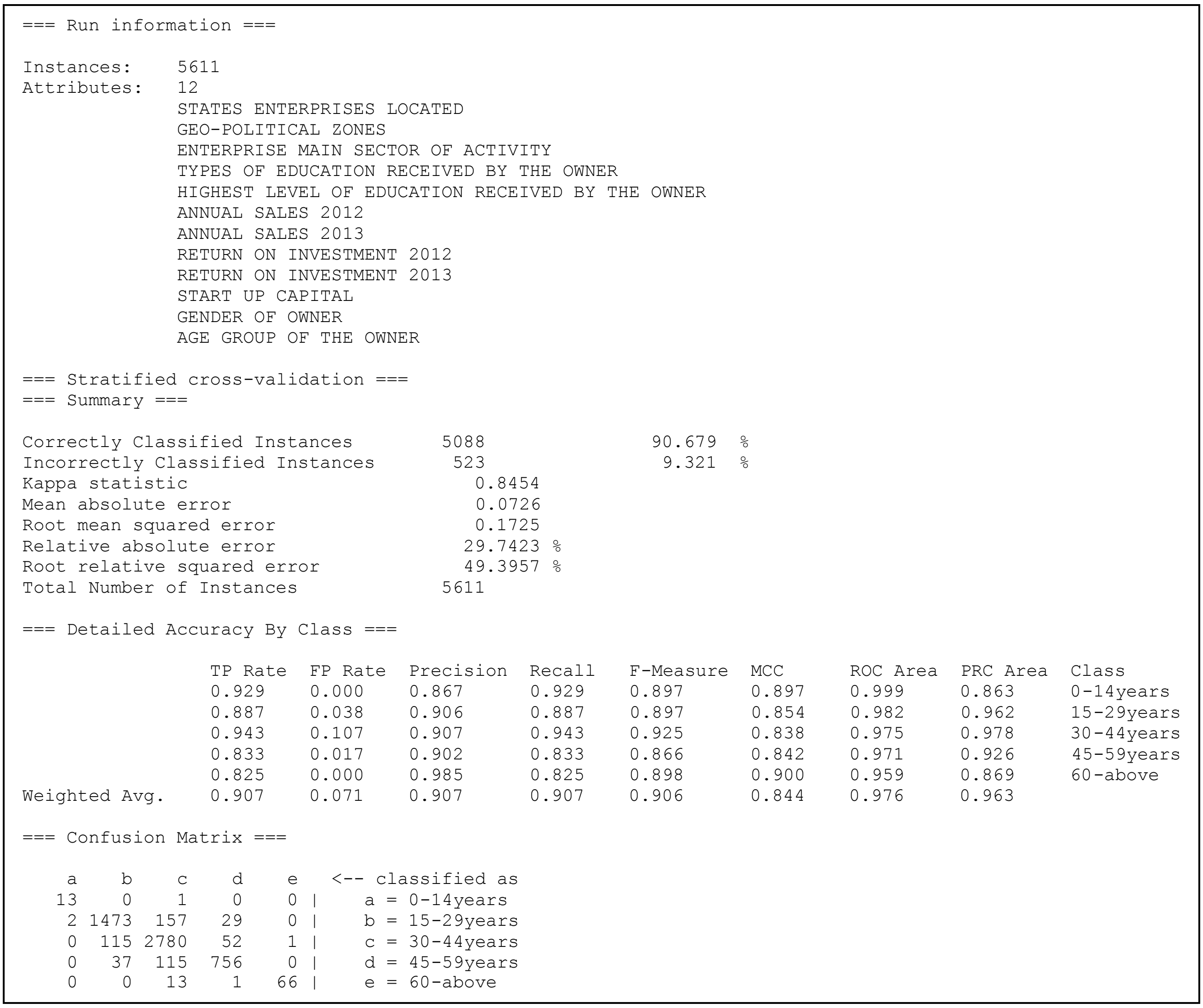
There are various ways of generating subsets of the data. The researcher used **Resample** to produce a random sample by sampling with replacement. The idea of the bootstrap is to sample the dataset with replacement to form a training set which will create a new random sample the same size as the training dataset, but will have a different composition. This is because the random sample is drawn with replacement, which means that each time an instance is randomly drawn from the training dataset and added

DOI: 10.31695/IJASRE.2019.33565

to the sample, it is also added back into the training dataset (replaced) meaning that it can be chosen again and added twice or more times to the sample, [13].

Resampling produces a random subsample of the dataset using either sampling with replacement or without replacement. The original dataset must fit entirely in memory. The number of instances in the generated dataset must be specified.

Experiment I: Predicting Age Group Attributes of the dataset



```
=== Run information ===

Instances:     5611
Attributes:    12
               STATES ENTERPRISES LOCATED
               GEO-POLITICAL ZONES
               ENTERPRISE MAIN SECTOR OF ACTIVITY
               TYPES OF EDUCATION RECEIVED BY THE OWNER
               HIGHEST LEVEL OF EDUCATION RECEIVED BY THE OWNER
               ANNUAL SALES 2012
               ANNUAL SALES 2013
               RETURN ON INVESTMENT 2012
               RETURN ON INVESTMENT 2013
               START UP CAPITAL
               GENDER OF OWNER
               AGE GROUP OF THE OWNER

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

Correctly Classified Instances        5088               90.679  %
Incorrectly Classified Instances       523                9.321  %
Kappa statistic                          0.8454
Mean absolute error                      0.0726
Root mean squared error                  0.1725
Relative absolute error                 29.7423 %
Root relative squared error             49.3957 %
Total Number of Instances             5611

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.929    0.000    0.867      0.929    0.897      0.897    0.999     0.863     0-14years
                 0.887    0.038    0.906      0.887    0.897      0.854    0.982     0.962     15-29years
                 0.943    0.107    0.907      0.943    0.925      0.838    0.975     0.978     30-44years
                 0.833    0.017    0.902      0.833    0.866      0.842    0.971     0.926     45-59years
                 0.825    0.000    0.985      0.825    0.898      0.900    0.959     0.869     60-above
Weighted Avg.    0.907    0.071    0.907      0.907    0.906      0.844    0.976     0.963

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   13    0    1    0    0 |    a = 0-14years
    2 1473  157   29    0 |    b = 15-29years
    0  115 2780   52    1 |    c = 30-44years
    0   37  115  756    0 |    d = 45-59years
    0    0   13    1   66 |    e = 60-above
```

Figure 6.3: Magnified output of Age Group prediction

In this experiment, the class label (Age Group of Owner) has norminal class attribute therefore the supervised version of Resample Instance Filter was applied followed by the Random Forest Algorithm. The screenshot of Fig 6.3 represents the summary of the algorithm performance results for the kappa statistics mean, absolute error, root mean squared error, correctly classified instances, and incorrectly classified instances from the total of 5,611 instances. The value of the Kappa statistic of 0.8454 means that statistical significance of the model is rather high. Therefore, a high value of Kappa depicts agreement of prediction with true class.

Precision is defined as the number of correct positive prediction over the total number of positive prediction;

Precision = TP/TP + FP

Recall is the TP rate (also referred to as sensitivity)

what fraction of those that are actually positive were predicted positive : TP / actual positives

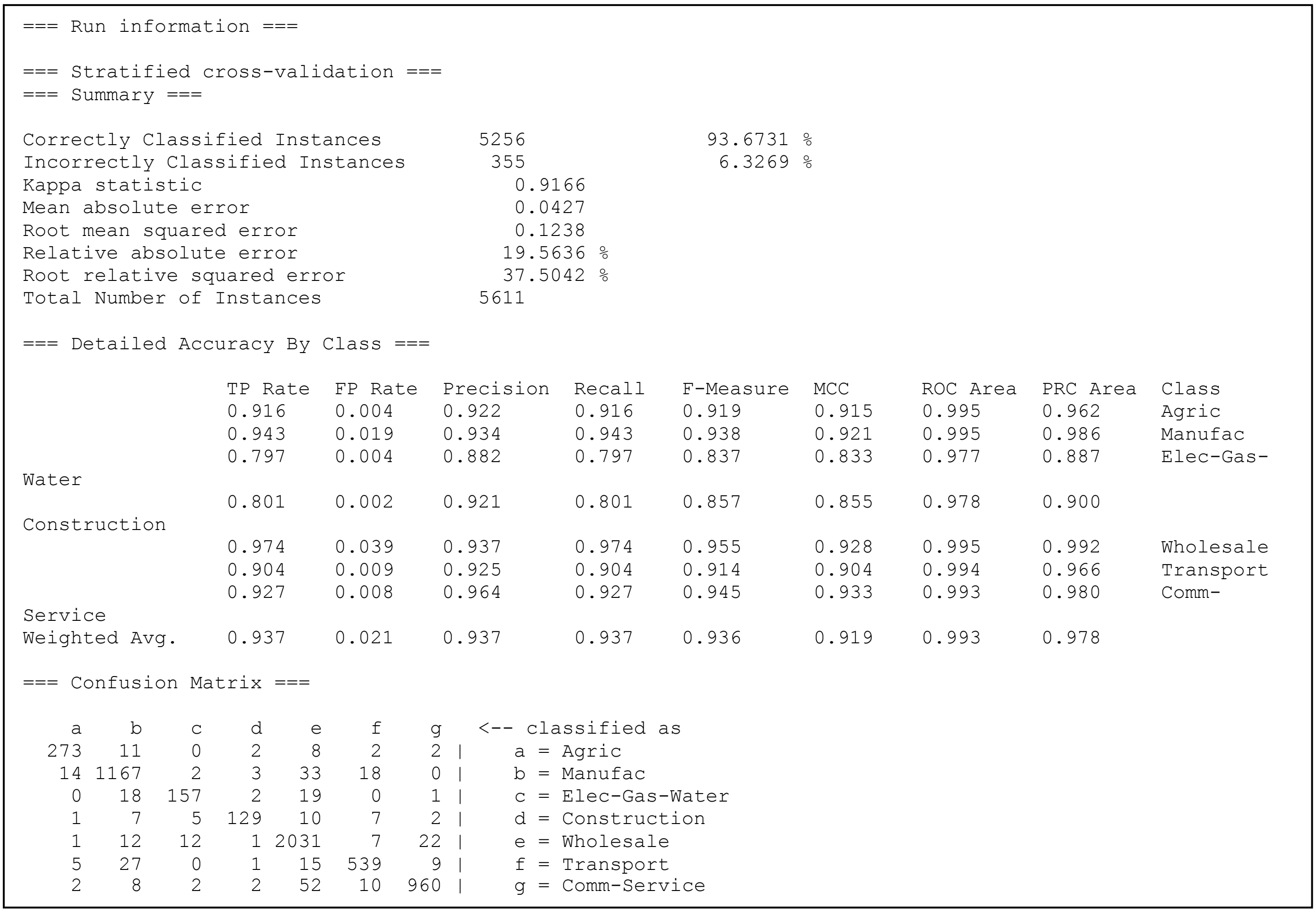
Recall = TP/TP + FN

DOI: 10.31695/IJASRE.2019.33565

A high Precision range of 86.7% to 98.5% indicates that the RandomForest algorithm returns more relevant results than irrelevant and the high Recall range of 82.5% to 92.9% means that most of the results returned by the algorithm are relevant.

The Random Forest prediction achieved a high predictive accuracy of 90.68% with 5,088 correctly classified instances and TP Rate of 0.943 for the dominant class of 30-44years.

Experiment II: Predicting Enterprise Main Sector Attributes of the dataset



```
=== Run information ===

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

Correctly Classified Instances       5256               93.6731 %
Incorrectly Classified Instances      355                6.3269 %
Kappa statistic                         0.9166
Mean absolute error                     0.0427
Root mean squared error                 0.1238
Relative absolute error                19.5636 %
Root relative squared error            37.5042 %
Total Number of Instances            5611

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.916    0.004    0.922      0.916    0.919      0.915    0.995     0.962     Agric
                 0.943    0.019    0.934      0.943    0.938      0.921    0.995     0.986     Manufac
                 0.797    0.004    0.882      0.797    0.837      0.833    0.977     0.887     Elec-Gas-
Water
                 0.801    0.002    0.921      0.801    0.857      0.855    0.978     0.900
Construction
                 0.974    0.039    0.937      0.974    0.955      0.928    0.995     0.992     Wholesale
                 0.904    0.009    0.925      0.904    0.914      0.904    0.994     0.966     Transport
                 0.927    0.008    0.964      0.927    0.945      0.933    0.993     0.980     Comm-
Service
Weighted Avg.    0.937    0.021    0.937      0.937    0.936      0.919    0.993     0.978

=== Confusion Matrix ===

    a    b    c    d    e    f    g   <-- classified as
  273   11    0    2    8    2    2 |    a = Agric
   14 1167    2    3   33   18    0 |    b = Manufac
    0   18  157    2   19    0    1 |    c = Elec-Gas-Water
    1    7    5  129   10    7    2 |    d = Construction
    1   12   12    1 2031    7   22 |    e = Wholesale
    5   27    0    1   15  539    9 |    f = Transport
    2    8    2    2   52   10  960 |    g = Comm-Service
```

Figure 6.4: Magnified output of Enterprise Main Sector prediction

This experiment is a multiclass classification problem with the class label (Enterprise Main Sector) being a multiclass norminal attribute as well. Similarly, the supervised version of Resample Instance Filter was applied followed by the Random Forest Algorithm. The screenshot of Fig 6.4 represents the summary of the algorithm performance results for the kappa statistics mean, absolute error, root mean squared error, correctly classified instances, and incorrectly classified instances from the total of 5,611 instances.

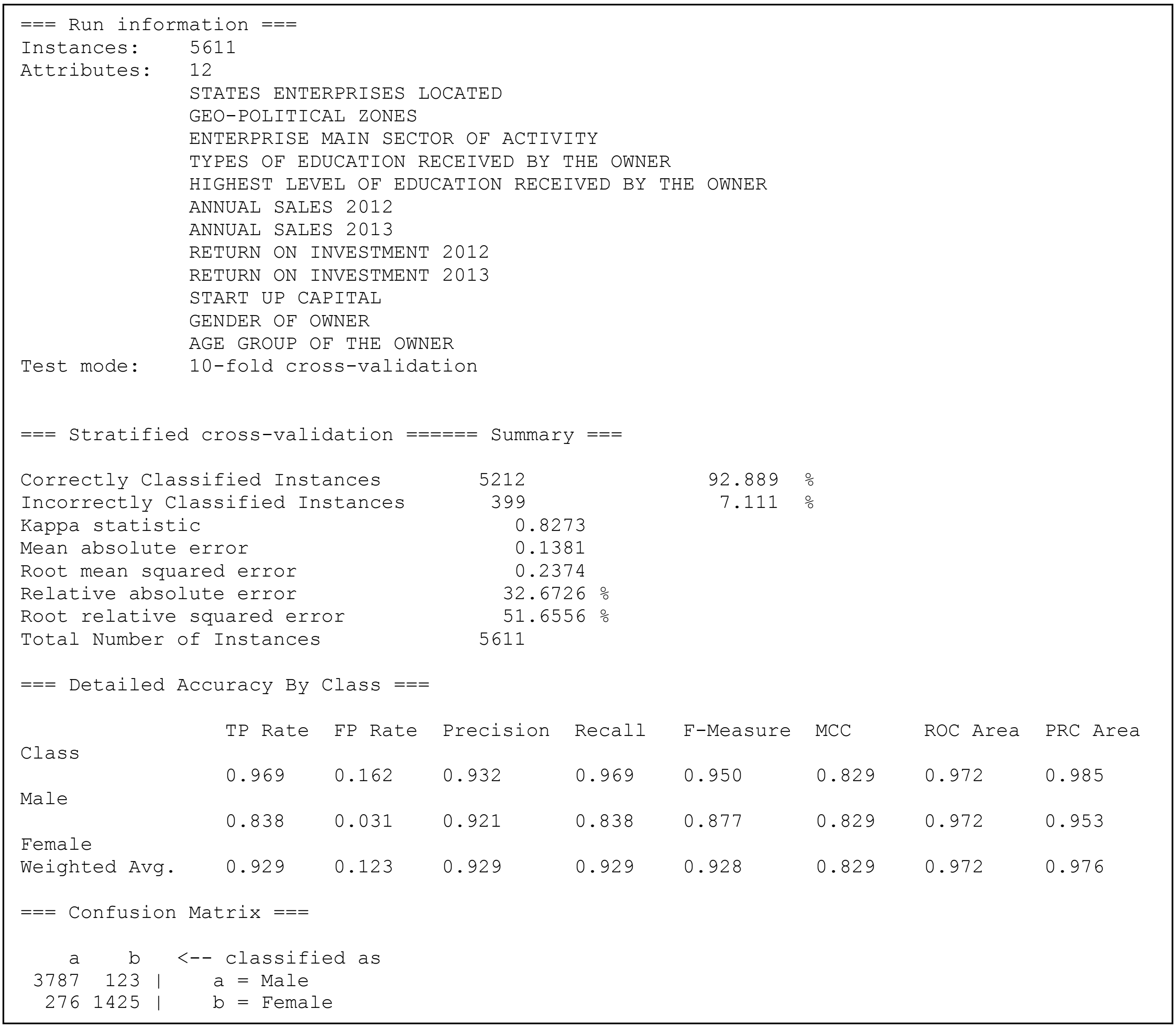
Kappa statistic of 0.9166 means the statistical significance of the model is high which depicts agreement of prediction with true class.

A high Precision range of 88.2% to 96.4% with a weighted average of 93.7% indicates that the RandomForest algorithm returns more relevant results than irrelevant and the high Recall range of 79.7% to 97.4% with a weighted average of 93.7% means that most of the results returned by the algorithm are relevant.

The Random Forest prediction achieved a high predictive accuracy of 93.67% with 5,256 correctly classified instances and TP Rate of 0.974 for the dominant class of wholesale.

DOI: 10.31695/IJASRE.2019.33565

Experiment III: Predicting Gender of Owner Attributes of the dataset



```
=== Run information ===
Instances:     5611
Attributes:    12
               STATES ENTERPRISES LOCATED
               GEO-POLITICAL ZONES
               ENTERPRISE MAIN SECTOR OF ACTIVITY
               TYPES OF EDUCATION RECEIVED BY THE OWNER
               HIGHEST LEVEL OF EDUCATION RECEIVED BY THE OWNER
               ANNUAL SALES 2012
               ANNUAL SALES 2013
               RETURN ON INVESTMENT 2012
               RETURN ON INVESTMENT 2013
               START UP CAPITAL
               GENDER OF OWNER
               AGE GROUP OF THE OWNER
Test mode:     10-fold cross-validation


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

Correctly Classified Instances        5212               92.889  %
Incorrectly Classified Instances       399                7.111  %
Kappa statistic                          0.8273
Mean absolute error                      0.1381
Root mean squared error                  0.2374
Relative absolute error                 32.6726 %
Root relative squared error             51.6556 %
Total Number of Instances             5611

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.969    0.162    0.932      0.969    0.950      0.829    0.972     0.985
Male
                 0.838    0.031    0.921      0.838    0.877      0.829    0.972     0.953
Female
Weighted Avg.    0.929    0.123    0.929      0.929    0.928      0.829    0.972     0.976

=== Confusion Matrix ===

    a    b   <-- classified as
 3787  123 |    a = Male
  276 1425 |    b = Female
```

Figure 6.5: Magnified output of Gender of Owner prediction

In this experiment, the class label (Gender of Owner) is a binary class label. The Supervised version of Resample Instance Filter was equally applied followed by the Random Forest Algorithm. The screenshot of Fig 6.5 represents the summary of the algorithm performance results for the kappa statistics mean, absolute error, root mean squared error, correctly classified instances, and incorrectly classified instances from the total of 5,611 instances.

Kappa statistic of 0.8273 means the statistical significance of the model is high which depicts agreement of prediction with true class.

A high Precision range of 92.1% to 93.2% with a weighted average of 92.9% indicates that the RandomForest algorithm returns more relevant results than irrelevant and the high Recall range of 83.8% to 96.9% with a weighted average of 92.9% means that most of the results returned by the algorithm are relevant.

The Random Forest prediction achieved a high predictive accuracy of 92.89% with 5,212 correctly classified instances and TP te of 0.969 for the dominant class of male owners of business.

DOI: 10.31695/IJASRE.2019.33565

7. SUMMARY, CONCLUSION AND RECOMMENDATIONS

7.1 Summary

In this project work, we reviewed the application of Data Mining tools and Machine Learning in the study of data collected on the informal sector of the Nigerian economy over the period 2014-2016. The Random Forest ensemble machine learning technique (algorithm) in particular were applied in these analysis and modeling of the data. We studied a proprietary dataset of NSIWC comprising 5611 instances with 12 features. We ran these analysis and generated results on **gender** of enterprise owner, **age group** of enterprise owner and **enterprise main sector** of activity.

7.2 Conclusion

The result shows that women participation in informal sector business is still low. The result further revealed that the female literacy level is poor compared to their male counterpart in the country, which lays credence to the fact that literacy level is a key factor to gender inequality of Nigeria. Participation of respondents in labor intensive enterprises like agriculture was particularly low and not encouraging. Similarly, the modal population age group of Nigerians working in the informal sector is in the age bracket of 30-44 years.

The distribution by gender indicates a particularly high share of men aged 30-44 years in the population. Labor force participation rates are typically lower for females than for males in all age category. One classic reason is the difference in life cycles of women and men still observed in many parts of the world, whereby women in the prime age tend to leave the labor force to give birth and raise children and men work to secure an income for the family.

7.3 Recommendations

There is an urgent need as a nation to drive a structural economic transformation with an emphasis on improving both public and private sector efficiency. This should be aimed at increasing national productivity and achieving sustainable diversification of production, to significantly grow the economy and achieve maximum welfare for the citizens, beginning with food, energy and security.

The results underscore the need for a national vision and comprehensive education and skills development as well as lifelong learning policies to improve the productivity and employability of citizens, and to sustain economic and social development in the long run. National policies should ensure quality education for all while strengthening an integrated education and training system relevant to the needs of people and the labor market.

Since low educational attainment and skills levels of women often lead to large gender imbalances in decent work opportunities and high rates of informal and vulnerable employment of women in the labor market, it is in the interest of governments to support the education of women and men equally.

The federal government should collaborate closely with businesses to deepen their investments in the agriculture, power, manufacturing, solid minerals and services sectors, and support the private sector to become the engine of national growth and development. In addition, science and technology should be effectively harnessed to drive national competitiveness, productivity and economic activities in all sectors. Government should be gender sensitive by encouraging women to be actively involved in all sectors of the economy.

To restore growth, government should diversify by focusing on the key sectors driving and enabling economic growth, with particular focus on agriculture, manufacturing, youth and women empowerment.

Digitization of services and open-government initiatives will engender operational efficiency improvements, reduce cost of governance, ensure citizen-centric policies and prompt service delivery. In the delivery of public services, governments must use the latest technologies to meet the rising expectations of hyper-connected citizens, while still reaching those offline - predominately the poor, the elderly and those living in areas with limited connectivity.

DOI: 10.31695/IJASRE.2019.33565

REFERENCES

[1]. Han J and Kamber, M. (2006): Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, San Franciso USA.

[2] Kantardzic, M. Data Mining: Concepts, models, methods, and algorithms, John Wiley and Sons,(2003)

[3] Aro-Gordon, S. (2014). Towards uncovering informal enterprises for financial inclusion: The hexagonal structural articulation approach. International Review of Research in Emerging Markets and the Global Economy, 1 (4), pp. 160 – 173. ISSN: 2311-3200.

[4] Gorunescu,F, (2011).: Data Mining: Concepts, Models and Techniques, ISRL 12, pp.57

[5] Daniell, K. A.,Morton, A. and Insua, D.R. (2016) Policy analysis and policy analytics. Annals of Operations Research, 236(1). pp. 1-13. ISSN 0254-5330, http://dx.doi.org/10.1007/s10479-015-1902-9

[6] Onyemaechi, J.O. (2013). "Role of the informal sectors in development of the Nigerian Economy: Output and employment approach". Journal of Economics and Development studies, 1(1), June, pp. 60-74.

[7] Chukuezi, C.O. (2010). "Urban Informal sectors and unemployment in third world cities: The situation in Nigeria". Asian Social Science, vol.6, No.8.

[8] Okeke,O.C. & Ekechukwu, B.C.(2013) Using Data-Mining Technique for Census Analysis to Give GeoSpatial Distribution of Nigeria. IOSR Journal of Computer Engineering (IOSR-JCE) Volume 14, Issue 2 (Sep. - Oct. 2013), PP 01-05

[9] Anonymous, "Predicting earning potential on Adult Dataset." Masters Thesis, Institute of Technology, May 2011.

[10] Llic, P.S.M and Veinovic, W.S.A., "Student's success prediction using WEKA tool", INFOTECH-JAHORINA, Vol.15, March 2016, pp. 684-688

[11] Witten, I. H.,Eibe, F. and Hall,M.A., Data mining : practical machine learning tools and techniques.—3rd ed., Morgan Kaufmann (2011)

[12] Raschka,S., (2015). Python Machine Learning, Packt Publishing Ltd., Birmingham,UK.

[13] Brownlee,J.(2016). "Supervised and Unsupervised Machine Learning Algorithm. Machine Learning Mastery, Volume 16, Issue 3, March 2016.

DOI: 10.31695/IJASRE.2019.33565