SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 1, NO. 2, A1, AUGUST 2018

A Comparative Study of the Efficient Data Mining Algorithm to Find the Most Influenced Factor on Price Variation in Oman Fish Markets

Amal Al-Hatali, Dr. Arockiasamy Soosaimanickam

Amal Al-Hatali; alhatali9991@gmail.com
Research Scholar, University of Nizwa, College of Economics, Management and Information Systems, Sultanate of Oman

Dr. Arockiasamy Soosaimanickam; arockiasamy@unizwa.edu.om
Associate Professor and Dean, University of Nizwa, College of Economics, Management and Information Systems, Sultanate of Oman

Received 23 June 2018; Accepted 02 August 2018; Published online 03 August 2018

Abstract

In oman, Fishing is one of the oldest professions that provides significantly to the national economy and for creating more job opportunities, especially, where many people completely depend on this income as an important source of living. The customers dealing with Fish markets in Oman need a good and innovative software platform to help them to deal with the problem of increasing of fish prices. This study aims at analysing various factors behind increasing prices in Oman fish markets using some data mining algorithms, by means of studying the old data that kept on the database that will assist to make a proper decision. The research has been conducted for the data collected from 29 fish markets in Sultanate of Oman and 15 fish species in each markets have been considered. To analyse the data, data mining algorithms, namely J48 algorithm, Decision Stump, and Random Tree has been chosen to perform the classification of data to find the most affected factor in fish price variations. The suitable algorithm has been chosen based on the good performance, which has been used for building an application. The result of the study shows that the Time is the major factor for price variations follwed by Place and then the quantity. This application model will help customers to get different information about prices in fish markets in Oman.

Keywords: Oman Fish Market, Factor Analyses, Attribute Selection, Information Gain Classification algorithm.

1. Introduction

In the last 47 years, Oman fisheries have dramatically improved in various aspects. In 2006, fisheries output is reported to have risen to 280,000 tones (Alwatan newspaper, 2017). This shows that compared to other animal food producing sectors, fisheries market continues to grow more rapidly. The average growth between 2011 and 2016 is about 12%; where it was 158 thousand tons in 2011, but it increased to 280 thousand tons in 2016 compared to the total tons in 2011 as it is shown in Figure 1. Request for fisheries products continues to rise to meet the needs of consumers, reflecting authorization of the dietary advantages of fish and shellfish in both developing and developed countries.

There is a variation in the fish prices in the market that make Omani customers suffer from changing fish prices. An analytical study will make to choose the most affected factor that has an impact on fish price. Data mining is a group of methods to find unobserved and useful information from large databases of various business domains. For identifying the interesting patterns (manner) and correlation and to get benefits from the data warehouse, Factor Analysis and Information Gain methods are used (PandyaJalpa P, 2017). Factor analysis reveals interesting correlation and/or relationships among a large set of data items. Factor Analysis shows attributes value conditions (options) that happen frequently together in a given dataset (Akash Rajak, 2012).

The data that is stored in the Agriculture and fisheries databases show an increase in prices often, there is a need to take advantage of this data by applying data mining techniques such as Factor analysis, Information Gain, and others.

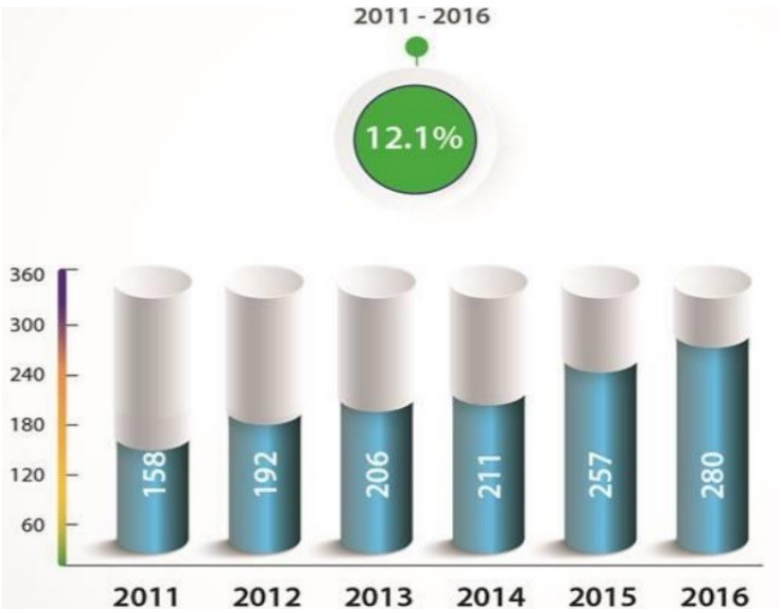

Figure 1: Fish Production statistics for the period 2011 – 2016

(Source: Ministry of Agriculture and Fisheries, 2016)

The discovered knowledge can be used to classify and analyze fish attributes, and to find the relationship between factors that affect fish' price and shows variation. The purpose of this study is to apply a classification (analysis) model for data and make a comparison based on the accuracy of data for different classifying algorithms and then find information gain of three experiments and understand entropy concept in order to develop an implementation that could help to decrease the issue of raising the prices.

2. Background of Fish Markets in Oman

The Government of Oman has been working hard to improve food production and security to overcome problems caused by war and famine. In addition, Oman plans to establish various investments in agriculture, horticulture, aquaculture and fishing, as the country is looking for possible solutions to support population growth and export promotion. In the fisheries sector alone, the government is trying to increase production from 257,000 tons per year in 2015 to 480,000 tons by 2020. Omani consumers are not only aware of the shortage of fish in the Sultanate's markets but also of rising prices. Oman is a major consumer of fish at about 28 kg per person per year. Fish prices have risen because of the growing population inside the country. Also, demand from neighboring countries such as UAE, Saudi Arabia and other countries. Another reason, fish are used extensively in tourist facilities due to high demand for tourists in the country.

This filed of fishing in oman is considerable contributing to improve oman national economy and creating new jobs. (fish production in Oman is growing by 5.4%, 2017). Also, there have been improvements made by the government in the recent years significantly in this area. In addition, it is considered as one of the most important and effective economic sectors that contribute remarkable to increase GDP growth. According to NCSIO in 2016, the GDP of fisheries increased by 18.4% compared to 2015 as shown in Figure 2. Moreover, it is not surprising to see the considerable achievements in this area during past few decades, and this sector has been classified as the second in the Arab world and the 26th in the world in the field of food security in 2016 (Ministry of Agriculture and Fisheries, 2016). The future of this fisheries sector will be highly fertile, specifically in the future direction of the government to activate other renewable energy based second level industries and to abandon oil and its derivatives as an important resource.

Figure 2: Gross Domestic Product value of the Agricultural and Fisheries Sector in Oman

(source: Ministry of Agriculture and Fisheries, 2016)

The social aspect has a significant impact on the fisheries sector where it is dependent on a large segment of residents, and is considered by many people as a source of income to provide the requirements of life, so daily volatility in prices directly affect them. There are obviously different prices for fish of the same species at different locations at the same time as in Ministry of Agriculture and Fisheries (2016). The reasons attributed are by several factors affecting fish prices, for example, supply and demand, climate, oil, fuel, gas, and many others (Qatan, 2010, Maribeth P. et al., 2016).

3. Related Work

At present, the research community is paying more attention to topics related to the factor analysis, which can be attributed to the active contribution to the growth of economies in governments and institutions. There are different ways to study factors (feature analyzes), and some papers are also presented here. It has been classified to six areas under the analysis of various factors including travel, economy, finance (equities), fisheries, agriculture, energy and minerals (electricity, oil and gas), sales and marketing (gold, retail and real estate). (Wohlfarth, et al., 2011) focuses on the field of travel and the classification models used are the classification tree and the random forest to analyze data on travel and travelers.

(Anita Bay, 2015), focuses on the area of economics, where factor analysis has been used: the classification tree. Data are collected in 20 countries, showing the economic ranking of countries (Kuwait, Germany, Iceland, Belgium, Denmark, Taiwan, Qatar, Ireland, Sweden, Luxembourg, Austria, Singapore, Norway, Netherlands, Hong Kong, Canada, and Australia). There are some research papers highlighting the use of different data mining algorithms to analyze factors in the field of fisheries and agriculture. According to (T SaiSujana, 2017), a comparative study of nine different sets of data with multiple unbalanced categories and associated with other meta-heuristic algorithms was performed. The results show that the proposed approach provides high accuracy for classification with a subset of attributes having fewer properties. These models are compared based on corresponding error measurement values from RMSE, MAE, and MAPE to see the improvements in the performance of those algorithms.

In the finance area, different algorithms have been used for feature selection, for instance, (O Villacampa, 2015) presented a comparative study among Classification Model. The authors used details about car services performed and car sales at over 200 auto agents. He concluded that Decision tree model provides better results than other model, in particular, the values of RMSE, MAPE and MAE.

In addition to the previous research results, there are different data mining algorithms to be explored in the field of fisheries. (Hu, et al., 2005) discusses about the hybrid model using various algorithms such

as the Wavelet Neural Network, the genetic algorithm, and the decision tree to predict the prices of aquatic products in a particular time periods, such as one day, one week and one month etc. As per the obtained results, when the prediction range is expanded, the accuracy of the prediction is not going downwards. As mentioned in (Failler, 2006), a comparative study has been performed on three time series functions namely Autoregressive, Moving Average and Autoregressive Moving Average. They have also used twelve species in England (Cornwall) to predict fishery prices for various time periods. These models are compared with similar error statistics, such as the Theil coefficient, the Root Mean Square Error, the absolute error ratio and the absolute percentage error. The better prediction method has been chosen based the smaller error value.

The same way, different algorithms were used to analyze factors in the field of agriculture as mentioned in (Seyed Jamal F, 2013), where they have examined the factors affect the development of nanotechnology in the agricultural sector of Iran. The methodology used in this study include a combination of descriptive and quantitative research. Also, it uses a factor and descriptive analysis as data processing methods. The research population includes researchers in the field of nanotechnology in the West Azarbaijan Province with 74 samples. The data collected by making face to face interview with respondents and analyzed by using the factor analysis technique. Based on the responses of the respondents, about 50% of the total common variance is explained by research, educational and informative factors, where the majority of it has been explained by the research factor (19.43%).

In energy filed, (Fengyan Fan and Yalin Lei, 2017) studies the reason behind Carbon emissions in China that affect the air and make pollution problems. Energy intensity was the main factor limiting carbon emissions, and the effect of inhibition increased every year. To control carbon emissions, Beijing should continue to adjust the way economic development and properly control the size of the population with improved energy efficiency.

In sales and marketing, (Usha Ananthakumar, 2017) presents a study which is based on analysis of the retail market in different parts of Mumbai, in India. Factors are analyzed on the data collected to the most important factors influencing sales in the region. The aim from this study to help new retailer who is interested in starting this business.

The results show that J48 algorithm provides a significant improvement in reducing the errors and dealing with missing values, and it concluded Decision Tree models give better results than others models and have a good performance to get a correct results.

4. Classification algorithm

There are different classification algorithms incorporated in the Weka tool. The chosen algorithms are J48, Random tree, and Decision Stump.

4.1 The J48 algorithm

J48 decision tree is related to ID3 algorithm. It use to decide the target value of a new sample based on different attribute values of the available data. The additional features of J48 are accounting for missing (unavailable) values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 writes and implement with Java language. The WEKA tool provides a number of chooses associated with tree pruning. The different attributes denoted by the internal nodes of a decision tree, the branches between the nodes tells us the possible values that these attributes can have in the experimental results, while the terminal nodes tell us the final value of the dependent variable. This algorithm generates the rules from which particular identity of that data is generated. The objective is gradually come bigger of a decision tree until it gains equipped of flexibility and accuracy.

4.2 Random Tree algorithm

The overlooking Classifier and ensemble learning algorithms generate lots of individual learners. It also suggests significant ideas to construct a random set of data for building a decision tree. In general, in a

standard tree every node is divided using the best split among all variables. These algorithms use this procedure for divide selection and thus create reasonably balanced trees where one global setting for the ridge value works across all leaves, thus simplifying the optimization procedure (Liaw, 2012).

4.3 Decision stump algorithm

It is a type of machine learning model which is consisting of one unit of the decision tree, i.e. a decision tree with one root which is directly connected to the other terminal nodes. A decision stump makes a forecasting based on the value of just a single input attribute.

5. Methodology

In general, customers and suppliers face many problems with price variations in fish markets without knowing the proper reasons or factors that affect fish prices. Hence the researchers in the research work wants to explore various options to provide an appropriate solution through proposed approaches, which are likely to find the most influential factors on fish prices to avoid any losses to satisfy their expectation as much as possible. The researchers aimed at this research to conduct a comparative study and analysis of the models commonly used in the area data mining classification algorithms. The slected algorithm based on good performance accuracy and the small error values will be selected for building an application to commonly satisfy the large nymber customers in this field. At the end of this reaserch, it is proposed to build an application model that will help the customers / client to reduce the problem of increased prices as much as possible. For this study, Weka software is chosen for the implementation.

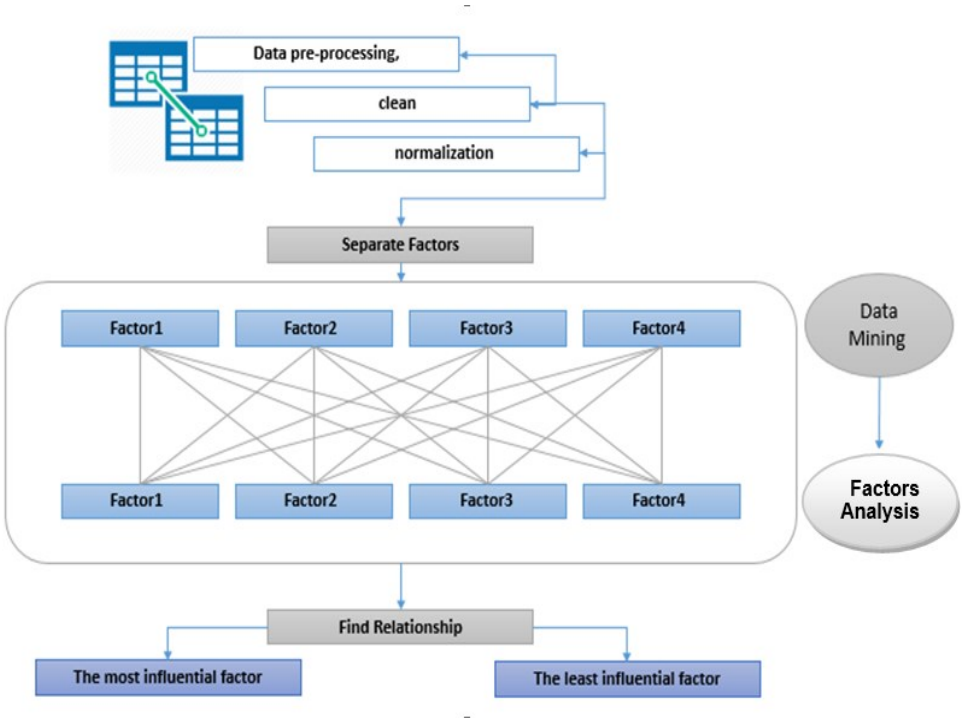

Figure 3: Proposed Model

As showed in Figure 3, there are five main steps, which are data gathering, data preprocessing, classification process, estimation and anatomy, and finally developing the implementation. First, data were collected manually from 29 markets across Oman during the period from November 2015 to October 2016, with prices of 15 species of fish. Second, in data processing, data is cleaned to organize the data to be used in the classification process. Third, the Weka workbook is used to perform the classification process. Fourth, in the evaluation process, an analytical study was conducted based on the results of absolute error average in the models used, namely the J48 algorithm, Decision Stump, and Random Tree. Al the end, the algorith which provides a good performance algorithm with minimum errors have been chosen to develop an application.

6. Comparison between algorithms

For comparison, the first experiment is performed on Weka with 10 fold cross-validation, the training set and split percentage (66%). The first step is to find the Confusion Matrix of the fish dataset using Random Tree, Decision stump, and J48 classification algorithms. In the next step, experiment calculates the classification accuracy and Mean absolute error.

Table 1: the Confusion Matrix for Random Tree Algorithm

Test Options		ALL Data
Fold Cross Validation 10	TP Rate	0.997
	FP Rate	0.001
	Precision	0.997
	Recall	0.997
	F-Mea sure	0.997
	ROC Area	1.000
Time Taken (Sec)	0.17	
Training Set	TP Rate	1.000
	FP Rate	0.000
	Precision	1.000
	Recall	1.000
	F-Mea sure	1.000
	ROC Area	1.000
Time Taken (Sec)	0.12	
Spilt percentage (66%)	TP Rate	0.999
	FP Rate	0.000
	Precision	0.999
	Recall	0.999
	F-Mea sure	0.999
	ROC Area	1.000
Time Taken (Sec)	0.18	

Table 2: the Confusion Matrix for J48 Algorithm

Test Options		ALL Data
Fold Cross Validation 10	TP Rate	0.999
	FP Rate	0.000
	Precision	0.999
	Recall	0.999

	F-Measure	0.999
	ROC Area	1.000
Time Taken (Sec)	0.06	
Training Set	TP Rate	1.000
	FP Rate	0.000
	Precision	1.000
	Recall	1.000
	F-Measure	1.000
	ROC Area	1.000
Time Taken (Sec)	0.01	
Spilt percentage (66%)	TP Rate	0.999
	FP Rate	0.000
	Precision	0.999
	Recall	0.999
	F-Measure	0.999
	ROC Area	1.000
Time Taken (Sec)	0.01	

Table 3: the Confusion Matrix for Decision stump Algorithm

Test Options		ALL Data
Fold Cross Validation 10	TP Rate	0.187
	FP Rate	0.179
	Precision	0.398
	Recall	0.187
	F-Mea sure	0.104
	ROC Area	0.508
Time Taken (Sec)	0.09	
Training Set	TP Rate	0.191
	FP Rate	0.181
	Precision	0.200
	Recall	0.191
	F-Mea sure	0.072
	ROC Area	0.508
Time Taken (Sec)	0.03	
Spilt percentage (66%)	TP Rate	0.185
	FP Rate	0.177
	Precision	0.204
	Recall	0.185
	F-Mea sure	0.067
	ROC Area	0.507
Time Taken (Sec)	0.04	

The simulation result shows that the highest correctly classified instances is (99%) out of 24572 instances by J48 Decision Tree and the lowest correctly classified instances is (19 %) by Decision Stump algorithm. The Random Tree Algorithm shows a closed result to J48 Algorithm which makes it in the second position. As the figure 4 shows that J48 takes less time to classify data with 0.027 second average for all three test options. Moreover, Decision Stump takes 0.053 second average for all test options, but Random Tree takes more time to classify 24572 instances for about 0.157-second average.

Table 4: Accuracy and Mean absolute error

Algorithm	Test Options	Accuracy	Mean absolute error
J48	Fold Cross Validation 10	0.999	0.000
	Training Set	1.000	0.000
	Spilt percentage (66%)	0.999	0.000
Random Tree	Fold Cross Validation 10	0.997	0.001
	Training Set	1.000	0.000
	Spilt percentage (66%)	0.999	0.000
Decision stump	Fold Cross Validation 10	0.187	0.139
	Training Set	0.191	0.139
	Spilt percentage (66%)	0.185	0.139

The J48 algorithm and Random Tree algorithm both gives 99% accuracy in fold Cross Validation 10. In fact, the highest accuracy belongs to the J48 Decision Tree classifier, followed by Random Tree algorithm, and Decision stump Tree Classifier.

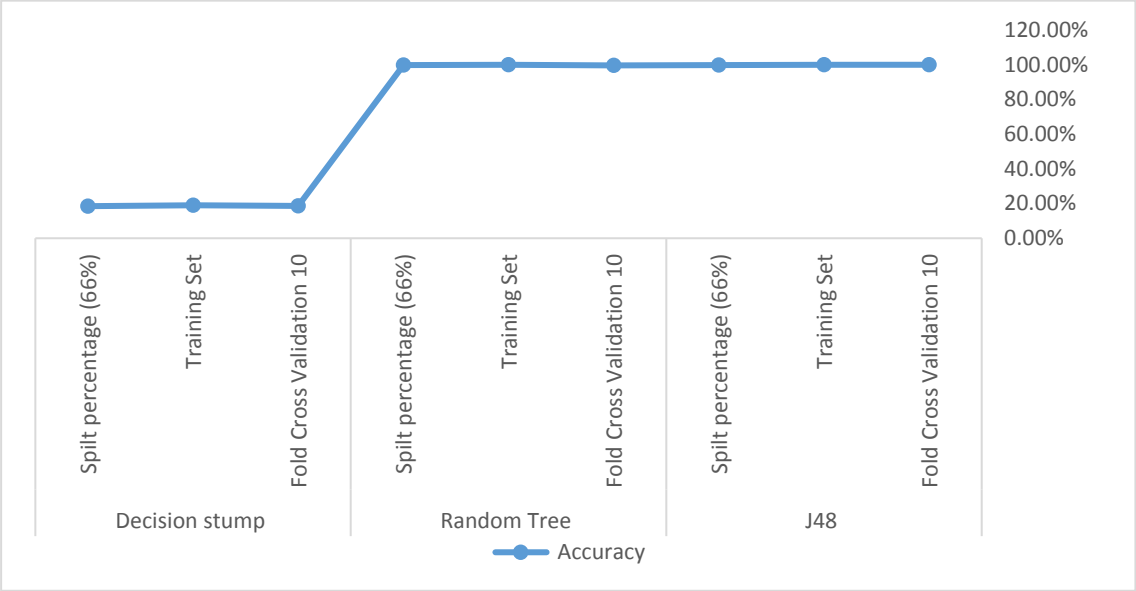

Figure 4: Accuracy Comparison

The average of mean absolute error of J48 algorithm for all test options is 0.0001 % and the average of mean absolute error of Random tree algorithm for all test options is 0.00023 %. But, the average of mean absolute error of Decision stump algorithm for all test options is 0.1392 %, which have more error than other algorithms.

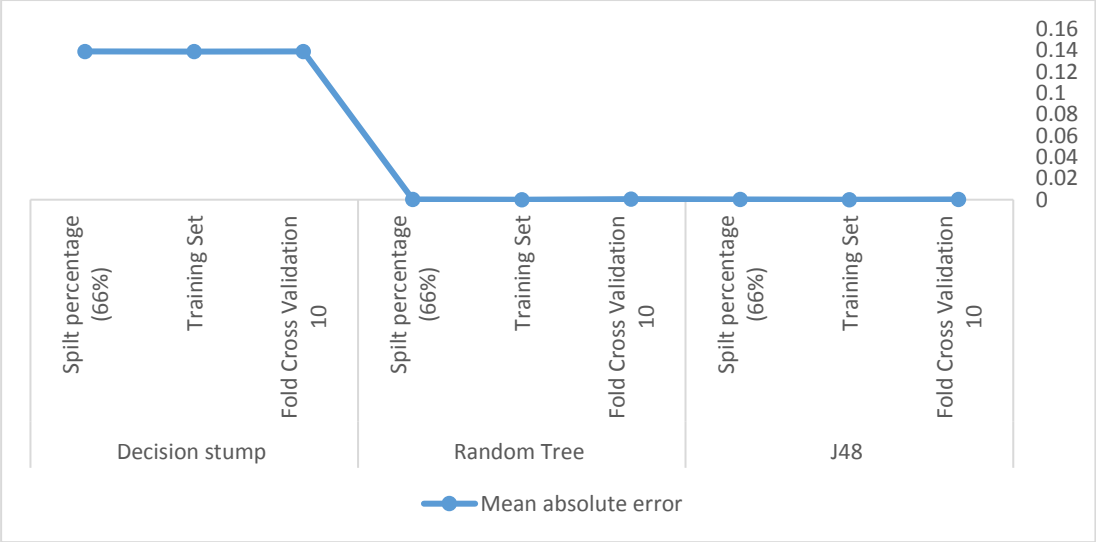

Figure 5: Error Comparison

7. Result and discussion

Three experiments have been performed and tested to find percentage of correctly classified instances and information gain value for each factor. First, the whole dataset classified by the J48 algorithm and calculates information gain for each factor in the database. Second, the split data depend on year Quarter. Finally, the split data for each location with information gain calculation for each location. The purpose of dividing the dataset is to check the validity of the result of entropy for four factors that selected to study.

7.1 Experiment A

The accuracy percentage of whole dataset is 99% in Fold Cross Validation (10) and Spilt percentage (66%). On other hand, the accuracy percentage of whole dataset is 100 % in Training Set.

Table 5: Accuracy Detailsby Class Weighted Average for Experiment A

Test Options	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
10 Fold Cross Validation	0.999	0.000	0.999	0.999	0.999	1.000
Training & Test Set	1.000	0.000	1.000	1.000	1.000	1.000
Spilt percentage (66%)	0.999	0.000	0.999	0.999	0.999	1.000

Ranked attributed are displayed according to the attribute selection that 2.7849 is with lead rank shown in first attribute name as Time and stand the first rank, the second attribute is the location with 0.2016, price and quantity take third and fourth rank position with 0.0348 and 0.0144 respectively.

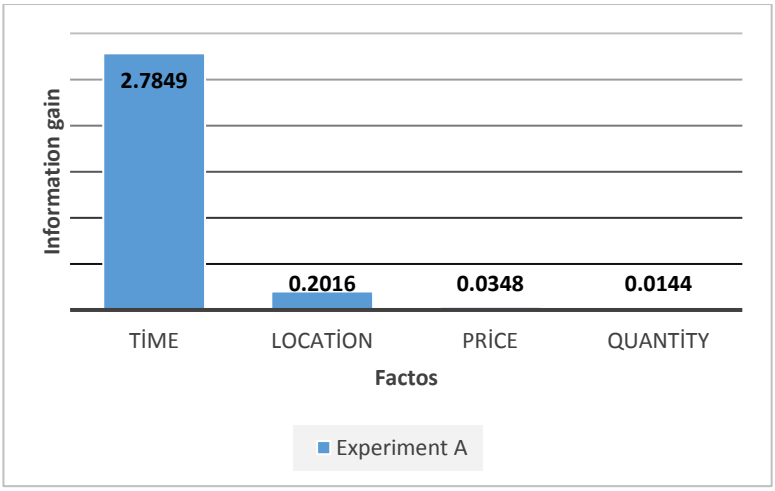

Figure 6: IG for whole data

7.2 Experiment B

The accuracy percentage of Experiment two for the first Quarter is 100 %. The accuracy percentage of Experiment two for the second quarter is 100 %. The accuracy percentage of Experiment one for the second quarter is 97%. The accuracy percentage of Experiment one for the second quarter is 96%.

Table 6: Accuracy Detailsby Class Weighted Average for Experiment B

Quarter	Test Options	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Quarter 1 (from November to January	Fold Cross Validation (10)	1.000	0.000	1.000	1.000	1.000	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	1.000	0.000	1.000	1.000	1.000	1.000
Quarter 2 (from February to April)	Fold Cross Validation (10)	1.000	0.000	1.000	1.000	1.000	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	1.000	0.000	1.000	1.000	1.000	1.000
Quarter 3 (from May to July):	Fold Cross Validation (10)	0.979	0.020	0.980	0.979	0.979	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	1.000	0.000	1.000	1.000	1.000	1.000
Quarter (from August to October):	Fold Cross Validation (10)	0.970	0.026	0.972	0.970	0.970	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.956	0.036	0.960	0.956	0.955	0.997

For experiment two, the result showed that Time has high rank all over other factors which are location, price and quantity. The result is almost similar to the result of first experiment. The following figure 7 shows the average Ranked attribute values for all Quarters of year:

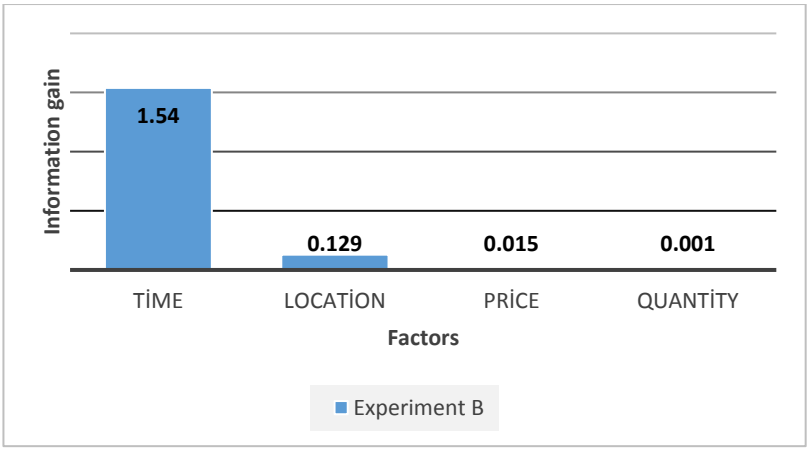

Figure 7: IG for divided Dataset by Time

7.3 Experiment C

For experiment three, the result showed that Time has high rank all over other factors which are location, price and quantity. The price takes second rank and quantity takes the third rank position.

Table 7: Accuracy Details by Class Weighted Average for Experiment C

Location	**Test Options**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Al Ashkara	Fold Cross Validation (10)	0.906	0.084	0.921	0.906	0.905	0.992
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.667	0.267	0.810	0.667	0.629	0.978
Al Meerah Al Suwaiq, Aljazir, Lulu Darsait Lulu Nizwa Sohar Suq aljumla	Fold Cross Validation (10)	1.000	0.000	1.000	1.000	1.000	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	1.000	0.000	1.000	1.000	1.000	1.000
Althermd	Fold Cross Validation (10)	0.991	0.002	0.991	0.991	0.991	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.966	0.008	0.971	0.966	0.966	1.000
Barka	Fold Cross Validation (10)	0.998	0.000	0.999	0.998	0.998	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.998	0.000	0.998	0.998	0.998	1.000
Buraimi Sinew	Fold Cross Validation (10)	0.999	0.000	0.999	0.999	0.999	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	1.000	0.000	1.000	1.000	1.000	1.000
Dibba Duqm	Fold Cross Validation (10)	0.998	0.000	0.998	0.998	0.998	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	1.000	0.000	1.000	1.000	1.000	1.000
Ibra	Fold Cross Validation (10)	0.987	0.006	0.988	0.987	0.987	1.000

	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.975	0.009	0.978	0.975	0.976	1.000
Ibri	Fold Cross Validation (10)	1.000	0.000	1.000	1.000	1.000	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.985	0.005	0.986	0.985	0.985	1.000
Izki	Fold Cross Validation (10)	0.793	0.057	0.838	0.793	0.790	0.967
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.650	0.022	0.899	0.650	0.709	0.916
Jalan Abu Ali	Fold Cross Validation (10)	0.995	0.001	0.996	0.995	0.995	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.996	0.002	0.996	0.996	0.995	1.000
Masirah	Fold Cross Validation (10)	0.963	0.012	0.968	0.963	0.963	0.999
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.890	0.031	0.927	0.890	0.894	0.996
Mirbat	Fold Cross Validation (10)	0.994	0.002	0.994	0.994	0.994	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.994	0.002	0.994	0.994	0.994	1.000
Muttrah	Fold Cross Validation (10)	0.999	0.000	0.999	0.999	0.999	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.990	0.002	0.991	0.990	0.990	1.000
Nizwa	Fold Cross Validation (10)	0.984	0.003	0.985	0.984	0.983	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.986	0.002	0.987	0.986	0.986	1.000
Qurriyat	Fold Cross Validation (10)	0.991	0.002	0.991	0.991	0.991	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.936	0.014	0.953	0.936	0.938	0.998
Rustaq	Fold Cross Validation (10)	0.994	0.002	0.995	0.994	0.994	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.963	0.008	0.969	0.963	0.963	1.000
Salalah	Fold Cross Validation (10)	0.941	0.015	0.955	0.941	0.937	0.998
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.927	0.013	0.951	0.927	0.922	0.998
Seeb	Fold Cross Validation (10)	0.974	0.005	0.978	0.974	0.973	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.966	0.006	0.972	0.966	0.965	1.000

Shinas	Fold Cross Validation (10)	0.999	0.000	0.999	0.999	0.999	1.000
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.986	0.003	0.987	0.986	0.986	1.000
Sur	Fold Cross Validation (10)	0.967	0.008	0.954	0.967	0.956	0.999
	Training Set	1.000	0.000	1.000	1.000	1.000	1.000
	Spilt percentage (66%)	0.950	0.010	0.912	0.950	0.928	0.997

The result is same as the result of first experiment. The following figure 8 shows the average Ranked attribute values for all locations:

Figure 8: IG for divided Dataset by location

As a measure of the success of the model, the classification rate was used on the fold Cross Validation 10 test sample. For composing a decision tree model, J48 algorithms were used, where their functioning is high accuracy with fewer errors.

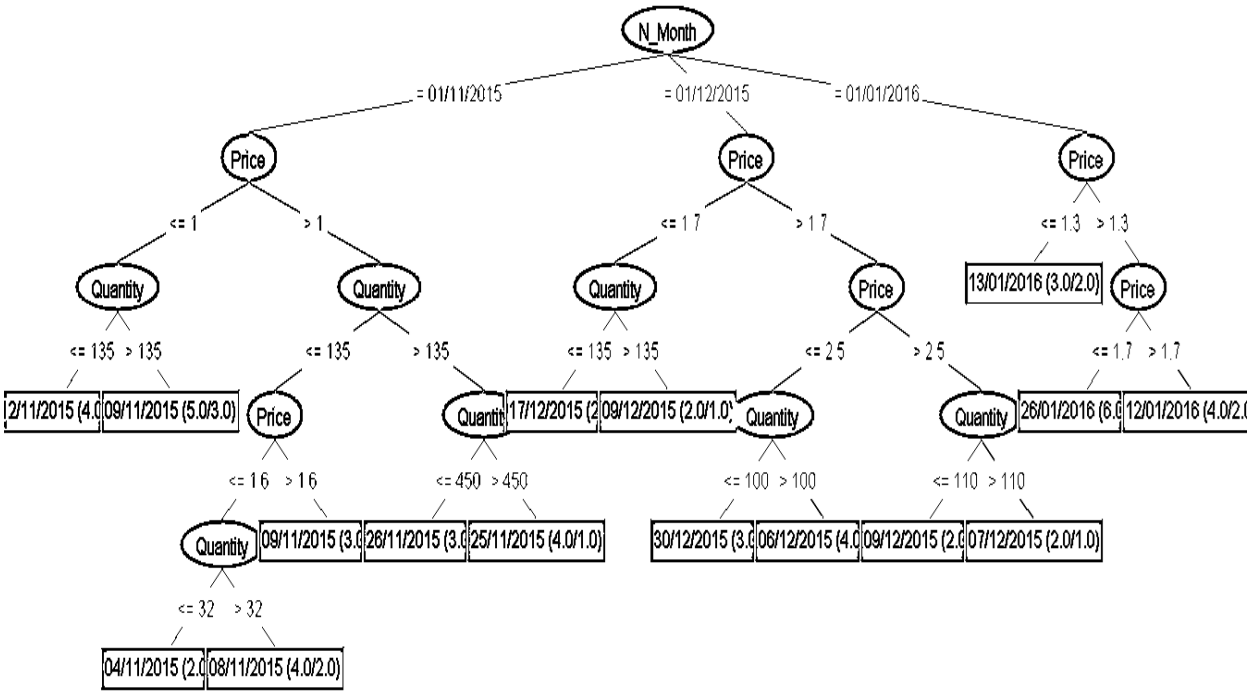

Figure 9: J48 decision tree in Weka

7.4 Comparative Analyses

Since the goal of this research was to find the factor behind variation of fish price in Omani market. To achieve the goal, the J48 algorithm used to check the accuracy of data and we have succeeded in achieving our target by using this algorithm to classify the data accordingly. We find there is a relationship between J4 algorithm and attribute selection which known as information gain method and give the same result. The previous comparison shows that the J48 algorithm had the highest classification accuracy rate of 99.92%. By using the Weka environment to test the three experiments. The information gain for each attribute shows that Time has a direct impact on fish price and make it change over time. The location, where fisherman or customer catch or buy fish has a second impact of fish price. The quantity has less impact on fish price. The figure 9 shows the ranked attributes in order before it tested by attribute selection algorithm. It is proved by this algorithm the order of each factor and how it impacts on fish price.

The main factor influencing the change in fish prices is the Time as shown in table 8 and the reason is that there are different fishing seasons that need to work out a schedule for fishing or increase fish production in ponds by raising fish in artificial ponds. Here, the lack of fish at the time increased the value of fish prices and vice versa.

Table 8: Information gain for three experiments

Experiment	Time	Location	Price	Quantity
Experiment 1	2.785	0.202	0.035	0.014
Experiment 2	1.540	0.129	0.015	0.001
Experiment 3	2.493	0.000	0.085	0.049

The second factor is the place where the price of fish differs in the possibility of fishing for fish selling places in the markets and we notice the increase in prices in markets and decline in fishing places. The third factor is the quantity, the higher the quantity the lower the price and the smaller the quantity, the higher the demand and the higher the price of fish.

8. Conclusion

The classification algorithm J48 has been chosen based on the different test results for building the application model. The test has been done to the feature selection and a weighted average which is calculated in percentages. Although cross validation test and split percentage test shows vast differences, the training test and cross-validation set almost produced the approximately similar result.

Based on the decision tree produced by the J48 algorithm, it is concluded that the most important factor that has effect on fish prices are a temporal factors (Time) only. This is due to different seasons of fish prices according to the four seasons and the possibility of the presence of a particular type of fish during the season. The more fish available in a season, the lower price of fish is identified and the less availability of fish, where the price is higher.

The application model designed to support the idea of factor analysis that can affect the difference in fish prices in a particular area. The application also provides information on fish in different areas and according to the price and quantity it needs.

Acknowledgments

This research was supported by a grant from TRC (The Research Council), Sultanate of Oman.

References

Akash Rajak and Mahendra Kumar Gupta (2012)."Association Rule Mining: Applications in Various Areas", Krishna Institute of Engineering & Technology, Pages 3-7.

Alapan, Maribeth P.(2016)."Factors Affecting the Market Price of Fish in the Northern Part of Surigao Del Sur, Philippines",Journal of Environment and Ecology, Issue number: 2,Volume number:7, Pages:34.

Altmeyer, Sebastian, and Robert I. Davis. (2014). "On the Correctness, Optimality and Precision of Static Probabilistic Timing Analysis". In Design, Automation & Test in Europe Conference & Exhibition (DATE), 2014, Pages 1–6. New Jersey: IEEE Conference Publications. doi:10.7873/DATE.2014.039.

Amal Al Mugrashi, Arockiasamy Sosoaimanickam. (2018). "A Comparative Study of the Efficient Data Mining Algorithm for Forecasting Least Prices in Oman Fish Markets", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 11, pp. 8751-8758

Anita Bai (2015). "An Application of Factor Analysis in the Evaluation of Country Economic Rank",Procedia Computer Science, Volume number: 54, Pages:311-317.

Cucu-Grosjean, Liliana, Luca Santinelli, Michael Houston, Code Lo, Tullio Vardanega, Leonidas Kosmidis, Jaume Abella, Enrico Mezzetti, Eduardo Quinones, and Francisco J. Cazorla (2012). "Measurement-Based Probabilistic Timing Analysis for Multi-Path Programs. In 2012 24th Euromicro Conference on Real-Time Systems", Pages 91–101. IEEE. doi:10.1109/ECRTS.2012.31.

D. Koller and M. Sahami(1996). "Toward optimal feature selection", In Proceedings of the Thirteenth International Conference on Machine Learning, Pages 284–292, 1996.

Muscat Newspaper, Article title:"6% rise in fish production; ministry targets half a million tonnes by 2020 – Oman", Website title: Muscat Daily News,URL: https://www.muscatdaily.com/Archive/Oman/6-rise-in-fish-production-ministry-targets-half-a-million-tonnes-by-2020-52dq.

Madhu Sudana Rao Nalluri , SaiSujana T , Harshini Reddy K ,Swaminathan V (2017).An Efficient Feature Selection using Artificial Fish Swarm Optimization and SVM Classifier. 2017 International Conference on Networks & Advances in Computational Technologies (NetACT). Pages 20-22,on July 2017

Ministry of Agriculture and Fisheries (2016), "أداء القطاع الزراعي والسمكي - المؤشرات الإنتاجية والإقتصادية",Website title: Maf.gov.om,URL: http://www.maf.gov.om/pages/PageCreator.aspx?lang=AR&DId=0&I=0&CId=0&CMSId=800746.

Farm and fishery vital for Oman's economy (2014), Avaliable On: http://timesofoman.com/article/31777/Oman/Farm-and-fishery-vital-for-Oman's-economy.

Fengyan Fan (2017),"Factor analysis of energy-related carbon emissions: a case study of Beijing,Journal of Cleaner Production ,Volume number: 163, Pages: S277-S283.

Fish production in Oman grows by 5.4% (2017), Fish production in Oman grows by 5.4%. Times of Oman. Retrieved 5 September 2017, from http://timesofoman.com/article/112688.

G. VamsiKrishna (2015),"An Integrated Approach for Weather Forecasting based on Data Mining and Forecasting Analysis",International Journal of Computer Applications, Issue number: (11), Volume number:120, Pages:26-29.

H. Almuallim and T. G. Dietterich (1994), "Learning boolean concepts in the presence of many irrelevant features", Artificial Intelligence, vol. 69, no. 1-2, Pages 279–305, 1994.

Hacer, E Aykut, E Halil and E Hamit, 2015,"Optimizing the monthly crude oil price forecasting accuracy via bagging ensemble models", Journal of Economics and International Finance, Issue number:5, Volume number:7, Pages:127-136.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009), "The WEKA data mining software", ACM SIGKDD Explorations Newsletter, vol 11(issue 1), Pages: 10. http://dx.doi.org/10.1145/1656274.1656278.

Hu Jing; Yang NingSheng; Ouyang HaiYing; Sun YingZe; Chen BaiSong (2013), "Application progress on data mining in field of fishery production", Journal of Agricultural Science and Technology (Beijing), Vol.15, issue No.4 ,Pages 176-182.

Kosmidis, Leonidas, Eduardo Quiñones, Jaume Abella, Tullio Vardanega, Carles Hernandez, Andrea

Gianarro, Ian Broster, and Francisco J. Cazorla. (2016), "Fitting Processor Architectures for Measurement-Based Probabilistic Timing Analysis" . Microprocessors and Microsystems 47. Elsevier B.V.: Pages 287–302. doi:10.1016/j.micpro.2016.07.014.

Liaw, Andy (2012), Documentation for R package random Forest.Pages 55-60.

M.Vasantha & Dr.V.Subbiah Bharathy (2010), "Evaluation of Attribute Selection Methods with Tree based Supervised Classification-A Case Study with Mammogram Images". International Journal of Computer Applications (0975 – 8887) Volume 8– issue No.12, October 2010.

Osiris Villacampa (2015), Feature Selection and Classification Methods for Decision Making: A Comparative Analysis.

PandyaJalpa P., MorenaRustom D (2017), "A Survey on Association Rule Mining Algorithms Used in Different Application Areas". Volume 8, issue No. 5, May-June 2017, Pages 1430-1436.

Qatan, Salim (2013), "Operating a wholesale fish market in the sultanate of Oman analyses of external factors" . Iceland: UNU-Fisheries Training Programme.Pages 40-45.

Sudhir , J., & Kodge, B., (2013), WEKA. Census Data Mining and Data Analysis Using WEKA. 1-6.

Tao Chen, Chu Zhang and Lifeng Xu (2016), "Factor analysis of fatal road traffic crashes with massive casualties in China", Advances in Mechanical Engineering 2016, Vol. 8(issue no. 4) Pages 1–11.

Seyed Jamal F. Hosseini and Nioushah Eghtedari (2013), "A confirmatory factorial analysis affecting the development of nanotechnology in agricultural sector of Iran". African Journal of Agricultural Research. Vol. 8(issue no. 16), Pages 1401-1404.

Trupti A. Kumbhare and Prof. Santosh V. Chobe (2014), "An Overview of Association Rule Mining Algorithms", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (issue no.1) , 2014, Pages 927-930.

Usha Ananthakumar & Ankita Mridha (2017). "Application of factor analysis in analyzing sales of the retail stores". Retrieved from https://ieeexplore.ieee.org/document/7919597/

W. Duch, T. Winiarski, J. Biesiada, J, and A. Kachel (2003), "Feature Ranking, Selection and Discretization", Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), Pages 251 – 254, 2003.

Wohlfarth, T., Clémençon, S., Roueff, F., & Casellato, X. (2011). "A data-mining approach to travel price forecasting", In Machine Learning and Applications and Workshops (ICMLA), 2011 10Th International Conference On, 1, Pages 84-89. URL: https://hal.archivesouvertes.fr/hal-00665041/file/ICMLA.pdf

Zahra Karimi (2013). "Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods",International Journal of Computer Applications,Issue number(4),Volume number:78, Pages:21-27.