# Machine Learning Techniques for analysis of Egyptian Flight Delay

**Hanaa M. Mohamed[1], Shahinaz M. Al-Tabbakh [2], and H. El-Zahed[3]**

[1]Internet Dev.Dept. Manager of IT Sector,EGYPTAIR Holding Cooperation, Cairo, Egypt

[2]Computer Science Group, Faculty of Women for Sciences, A. and Education, Ain Shames University, Cairo-Egypt.

Shahinaz.Altabbakh@women.asu.edu.eg

[3] Faculty of Women for Sciences, A. and Education, Ain Shames University, Cairo-Egypt.

HElzahed@gmail.com

**Abstract**

*Flight delay has been the fiendish problem to the world's aviation industry, so there is very important significance to research for computer system predicting flight delay propagation. Extraction of hidden information from large datasets of raw data could be one of the ways for building predictive model. This paper describes the application of classification techniques for analysing the Flight delay pattern in Egypt Airline's Flight dataset. In this work, four decision tree classifiers were evaluated and results show that the REPTree have the best accuracy 80.3% with respect to Forest, Stump and J48. However, four rules based classifiers were compared and results show that PART provides best accuracy among studied rule-based classifiers with accuracy of 83.1%. By analysing running time for all classifiers, the current work concluded that REPtree is the most efficient classifier with respect to accuracy and running time. Also, the current work is extended to apply of Apriori association technique to extract some important information about flight delay. Association rules are presented and association technique is evaluated.*

**Keywords:**
Airlines, Flight delay, WEKA, Bigdata, Data mining, classification Algorithms , J48,Random Forest, Decision Stump, Ripper rule, Association rules, Apriori, Confusion matrix.

## 1. INTRODUCTION

Every day, there are thousands of people who travel by airplane to get to their destinations. Unfortunately for airline travellers, however, many of these flights do not leave on-time. Flight delay are a major problem within the airline industry. Flight delay is a challenging problem for all airline companies [1, 2], which will lead to Financial losses, Fuel losses and negative impact on their business reputation.

***Corresponding author**: hanaa_maher@egyptair.com

This work explores what factors influence the occurrence of flight delays. Predicting whether or not a Flight delay will be initiated at an airport, based on meteorological forecast and traffic demand, can alert traffic managers and airlines about potential congestion and necessitate strategies for mitigating these disruptions to air traffic.

Data mining can be defined as the extraction of useful knowledge from large data repositories. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining is ready for application in the business community because it is supported by three technologies including machine learning, statistics, and database systems for the analysis of large volumes of data. Because of Machine Learning techniques, Machine can be trained using the training samples from existing data sets and then a training model can be devised which can be applied to the future data to categorize it correctly. Such learning models are usually classifiers. Decision Tree, Artificial Neural Networks, Bayesian networks, and Support Vector Machines are examples of such classifiers. But if data is irrelevant, redundant, noisy and unreliable data, then there is no doubt that training phase is a very challenging task. Number of data can be misclassified if the training phase is not proper. Classification is one of the most popular data mining tasks. Hence in this paper the authors have tried to make a comparison of various Decision Trees approaches and ensemble Learning rules. For data-mining purposes, an open-source software package called WEKA developed using JAVA was used[3, 4] to help us in postulating a predictive models for flight delays based on various attributes of a particular flights. These models will make us able to predict when delays would be encountered and increase the knowledge for passengers advising them on the most efficient ways to travel.

In this paper, eight classification algorithms are investigated for their performance, as explained in section 3 with some theoretical aspects. In sections 2, related works are discussed. In sections 4, the types of classification algorithms including their performance measures will be investigated. Results and comparative analysis, and conclusions will be explained in sections 6 and 7 respectively.

## 2. RELATED WORKS

There are many data mining techniques proposed in the literature based on machine learning for building and extraction of predictive models using historical data. Namely, clustering, classification rules, association and regression. The structure of the trained predictive models can provide insight into factors that influence the initiation of a ground delay program at a given airport. Unsupervised data modeling techniques such as Principal Component Analysis, and clustering have been applied to classify days based on weather impact [5] and performance metrics [6] our contribution is discovery of the classification rules by more than one method and comparison of these methods for traffic flight data set of Egypt Airlines. It is Unique up to now, However there are some work done by:

Akpinar and Karabacak [7] provided a reviewed to Data Mining in Civil Aviation, so they implemented some data mining classification rules based on certain critical factors

including airlines, airports, cargo, passenger, efficiency and safety. Airline companies may use data mining in order to fuel cost optimization, planning take into consideration weather conditions, passenger analysis, cargo optimization, airport situation revenue per flight, profit per flight, cost per seat or more detailed one catering and handling expenses per seat.

Nazeri and Zhang [8] applied a data mining approach for analysis of whether impact on NAS performance. They applied decision tree learning algorithms C5.0 and K mean clustering algorithms. They discovered that weather pattern /rules had significant impact on NAS performance. Discovered rules may be used to predict if a day is good or bad performance based on its weather. They conclude rule relate between blocked flights and bypass distance.

Ha and it al [9] developed and implemented an experimental model based on the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology applied to Big Data Mining. They concluded that the arrival delay can be proposed for optimal airports. They classify the airports according to arrival delay.

Mukherjee and Sridhar[1] presented two supervised-learning models, logistic regression and decision tree to predict occurrence of a Ground Delay Program (GDP), at an airport based on traffic demand and meteorological conditions. The models are applied to predict GDP occurrence at two major U.S. airports: San Francisco International airport (SFO) and Newark Liberty International (EWR) airports. Historical hourly observations of meteorological conditions such as visibility, cloud height, wind, convection, precipitation, etc., and arrival traffic demand based on flight schedules are used to calibrate the models. The logistic regression model estimates the probability of a GDP occurrence during the hour. The decision tree model, on the other hand, classifies the hour as a GDP or non- GDP.

## 3. THEORETICAL OVERVIEW:

### 3.1. Machine learning approaches

Machine learning is a set of algorithms which are able to find potential patterns in data and use these patterns to predict unlabelled data. In 1959, Arthur Samuel [10] informally defined machine learning as a subject which tries to figure out how to make computer solve real problems automatically without programming detailed code. For example, he developed the Samuel Checkers-playing Program which had the ability to learn a reasonable way to play and managed to defeat many human players

### 3.2. Classification Models

Classification is one of the Data Mining techniques that is mainly used to analyse a given dataset and takes each instance of it and assigns this instance to a particular class with the aim of achieving least classification error. It is used to extract models that correctly define important data classes within the given dataset. It is a two-step process. In first step the model is created by applying classification algorithm on training data set.[22] Then in second step, the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy. So, classification is the process to assign class label for this

dataset whose class label is unknown. Various lists of techniques are available for classification like decision tree induction, Bayesian classification, and Bayesian network. Here under is the description of classifications algorithms to be investigated in this paper.

Decision Tree is a model which is used on most classifier in our work. It has a tree like structure in which all the internal nodes (including the root node) represent an attribute and the leaf nodes represent the classification categories as per the goal class [11]. The basic concept by which classification is done in Decision tree is that it takes multiple linear decisions to perform a nonlinear classification. A decision is taken at every level of the tree and finally we arrive at leaf node which leads the instance to belong to a particular category. According to the type of data being used, decision trees are categorized into two types:

**a) Classification Trees (For categorical data)**

**b) Regression Trees (For Numerical data)**

Entropy and information gain are the two parameters whose value determines the purity of a node and according to that only decision tree is constructed [24]. Most pure node lies at the bottom (as leaf nodes). DT use as a descriptive means for calculating conditional probabilities. A Decision Tree uses a top-down, divide and conquer as classification strategy and it partitions a set of given entities into further smaller classes on the basis of automatically selected rules .In addition , Weka implementation classes of classification trees are indicated by the following types . Our work is the first work to compare various tree classifiers in WEKA. The classifiers description are as following [3]:

**3.2.1-classifiers.trees.DecisionStump**

Class for building and using a decision stump. Usually used in conjunction with a boosting algorithm. Does regression (based on mean-squared error) or classification (based on entropy). Missing is treated as a separate value.

**3.2.2-classifiers.trees.J48**

decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data.this classifier is an open source java implementation of the C4.5 algorithm in the WEKA data mining tool.J48 also accepts both continuous and categorical attributes in building the decision tree. It use tree pruning that reduces misclassification errors and reducing the size of the decision tree. It is used to mitigate over fitting, where perfect accuracy on training data are also achieved [11, 12].

**3.2.3-classifiers.trees.RandomForest**

Class for constructing a forest of random trees. Random forests are collections of trees, all slightly different. Random Forest algorithm can be used to handle missing values.[27] Classify new data points by taking a (weighted) vote of their predictions that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees,

overfitting will be avoided by having many trees and so therefore more accurate also It run efficiently on large databases.

**3.2.4-classifiers.trees.REPTree**

Class for fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5) [8].

**Another Rules Classifiers in WEKA are as following:**

**3.2.5-classifiers.rules.PART** Class for generating a PART decision list. Uses separate-and-conquer. Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule [13].

**3.2.6-classifiers.rules.DecisionTable**

Class for building and using a simple decision table majority classifier. A decision table consists of a hierarchical table in which each entry in a higher level table gets broken down by the values of a pair of additional attributes to form another table. The structure is similar to dimensional stacking [14].

**3.2.7-classifiers.rules.OneR**

Class for building and using a OneR classifier; classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor, a frequency table is constructed for each predictor against the target. It uses the minimum-error attribute for prediction, discrediting numeric attributes also it is simple for humans to interpret [15].

**3.2.8-classifiers.rules.JRip**

It is the Weka implementation class of the algorithm Ripperk [8] that apply Ripper Rule (JRIP). Ripper Rule (JRIP) is used to generate various rules by adding repetitive datasets until the rules cover all data pattern according to the training dataset. In addition, once all rules are generated, some of them will be merged in order to reduce size[24]. In addition, this algorithm uses incremental reduced-error pruning in order to obtain a set of classification rules; k is the number of optimization cycles of rules sets.

**3.3. ASSOCIATION RULES**

Association rules are if/then statements that help uncover relationships between prominently unrelated data in a relational database [16]. Apriori algorithm is used to perform association rule mining. The Apriori algorithm was introduced in [AS94] as a way to generate association rules from market basket data. The Apriori algorithm is a two stage process: A frequent itemset (itemsets that satisfy minimum support threshold) mining stage and a rule generation stage (rules that satisfy minimum confidence threshold). WEKA allows the

resulting rules to be sorted according to different metrics such as confidence Eq(1) and lift.Eq(2).

If we have a given rule with 2 sides L => R

Confidence is the probability that L and R occur together, if the Confidence value is less than 1; L and R are negatively correlated, otherwise L and R positively correlated

Confidence = Pr(L,R) (1)

Lift (or improvement) is the ratio of the probability that L and R occur together to the multiple of the two individual probabilities for L and R or it is computed as the confidence of the rule divided by the support of the right-hand-side (RHS).

lift = Pr(L,R) / Pr(L).Pr(R) (2)

If this value is 1, then L and R are independent. The higher this value, the more likely that the existence of L and R together in a transaction is not just a random occurrence, but because of some relationship between them.

## 4. METHODOLOGY:

For achieving the goal of this research, we have put our methodology in phases as shown in figure.1. We used machine-learning algorithms embedded in WEKA on the data obtained from EGYPTAIR fleet-watch management

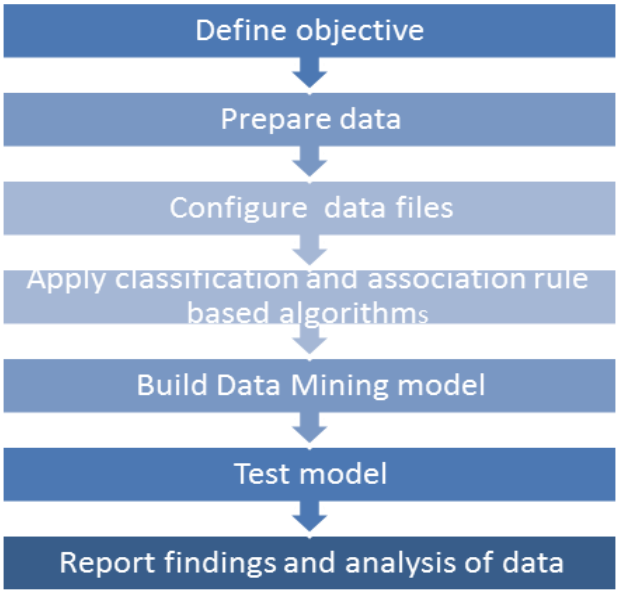


**Figure. 1. Phases of extraction of flight delay model**

### 4.1. WEKA ENVIRONMENT

WEKA (Waikato Environment for Knowledge Analysis) environment is a very popular tool programmed in Java that deals with machine learning and data mining. It contains a collection of visualization tools and algorithms available for data mining preprocess such as classification, association, clustering, modeling and evaluation.

### 4.2. DATA PREPARATION

The dataset for this research was obtained from online flight schedule repository [18]. The description of data set is shown in table 1, table 2. The dataset went through a number of transformations to prepare it for data-mining. First, we remove all flights that were missing data. A flight may be missing data because it was diverted, cancelled, or for some other related reason. Without complete values, these entries are not suitable to data-mining, so they were removed. The dataset itself contains data on both departure and arrival times, but for simplicity we decided to examine only the departures.

**Table1. Data set Description**

| | |
|---|---|
| Name of data set | Egypt Air flight delay data |
| Original Owner | Online repository [18] |
| Date received | JANUARY 2018 |
| Number of attributes | 9 |
| Type of attributes | Numeric + text |
| Number of records | 512 |
| Train: test | 66:33 |

**Table 2. Description of attributes for the data set**

| Attributes | Description | Type |
|---|---|---|
| Ser | serial | Integer |
| DepDatop | Date of operation of Departure | Date/Time |
| Flight | Flight Number | Integer |
| Origin | City From | Text |
| Dest | City To | Text |
| STD | Schedule Time of Departure | Integer |
| ATD | Actual Time of Departure | Integer |
| ArrDatop | Date of operation of Arrival | Date/Time |
| STA | Schedule Time of Arrival | Integer |
| ATA | Actual Time of Arrival | Integer |
| ACType | Aircraft type | Text |
| REG | Aircraft unique Registration letters | Text |

### 4.3. Data cleaning and transforming

Data cleaning is a very important step in data mining; here is where each dataset is cleaned. This means the model is prepared to fit the needed format, by eliminating unwanted values such as outliers, extreme values or values, which aren't wanted to be processed, transformed or adapted to the desired format[26]. If this step isn't done correctly, the final results may turn out to be incorrect or may vary a lot from the correct ones. That's why as mentioned in the first line of this paragraph, it is a very important step, and it doesn't take the same amount of time and effort to do it well than to do it wrongly. Figure 2. The data is configured to be in Arff format (WEKA) file format, with attribute values.

```
@RELATION FW

@ATTRIBUTE sTD NUMERIC

@ATTRIBUTE aTD NUMERIC

@ATTRIBUTE sTA NUMERIC

@ATTRIBUTE aTA NUMERIC

@ATTRIBUTE  DEPDATOP

@ATTRIBUTE  ArrDatop

@ATTRIBUTE

ATTRIBUTE cityFrom
{ABS,ABJ,ABV,ACC,ADD,AHB,ALG,AMM,AMS,ASM,ASW,ATH,AUH,BAH,BCN,BE
Y,BGW,BKK,BOM,BRU,BUD,CAI,CAN,CDG,CGK,CGN,CHR,CMN,CPH,DAR,DJE,DM
E,DMM,DOH,DXB,EBB,EBL,ELQ,FCO,FRA,GAE,GVA,HBE,HMB,HRE,HRG,IST,JED,J
FK,JNB,JUB,KAN,KRT,KUL,KWI,LCA,LHR,LOS,LXR,MAD,MAN,MCT,MED,MUC,M
UH,MXP,NBO,NDJ,OST,PEK,RMF,RUH,SAH,SHJ,SSH,SXF,TIP,TLV,TUN,VIE,YYZ}

@ATTRIBUTE city_To
{ABS,ABJ,ABV,ACC,ADD,AHB,ALG,AMM,AMS,ASM,ASW,ATH,AUH,BAH,BCN,BE
Y,BGW,BKK,BOM,BRU,BUD,CAI,CAN,CDG,CGK,CGN,CHR,CMN,CPH,DAR,DJE,DM
E,DMM,DOH,DXB,EBB,EBL,ELQ,FCO,FRA,GAE,GVA,HBE,HMB,HRE,
HRG,IST,JED,JFK,JNB,JUB,KAN,KRT,KUL,KWI,LCA,LHR,LOS,LXR,MAD,MAN,MCT,
MED,MUC,MUH,MXP,NBO,NDJ,OST,PEK,RMF,RUH,SAH
,SHJ,SSH,SXF,TIP,TLV,TUN,VIE,YYZ}

@ATTRIBUTE STATUS {A,D,R,S}

@ATTRIBUTE DAY NUMERIC

@ATTRIBUTE Class {'1','0'}
```

**Figure 2. The data is configured to be in (WEKA) file format, with attribute values**

The instance values of city_To and city_from is three-letter codes which are the standard codes used for airports. Their meaning can be found on any travel-reservation website. The instance values of Aircraft unique Registration letters is {A, D, R, S}, R: means returned, S: mean scheduled, D: means departed, A: Arrived . Some important notices during building model in WEKA:

When the dataset was imported into WEKA, a series of problems appear and to solve these problems some transformations had to be done to the data before being able to do anything else. As fixing model and clean data.

Every data mining software doesn't work the same way and doesn't have the same data requirements[27]. So before doing anything, some settings had to be adjusted before being be able to deal with the model without any more problems. Some variable types had to be changed, because WEKA automatically decides the variable type when you import a dataset.

Useless variables were removed using a WEKA filter. This removed the Origin and Origin city, and that's because it never changes, so the origin city is Cairo and the origin is CAI and both of these can be assumed. In addition to those variables, some variables that couldn't be known beforehand

### 4.4. Metrics of Comparisons

Classifiers in the open-source program WEKA were used to determine whether any algorithm stood out as being particularly well suited for these data[25]. The eight classifiers, which were tested, are readily available and represent some of the most widely used classification methods in aviation. However, they also represent classifiers that are diverse at the most fundamental levels. The eight classifier is explained previously in the theoretical overview part are decision stump, trees_J48, Random Forest, REPTree, PART, Decision Table , OneR and Ripperk. A distinguished confusion matrix was obtained to calculate sensitivity, specificity, precision, recall, F- measure, Roc and accuracy which are the metrics of evaluation of these classifiers. Confusion matrix is 2X2 matrix Representation of the classification results. The number of correctly classified instances is the sum of diagonals on the matrix; all others are incorrectly classified. Table 3 shows a representation of confusion matrix, where: TP = true positive, TN = true negative, FP = false positive, FN = false negative.

**Table 3 Confusion matrix**

| | **Classified as Delayed** | **Classified as not Delayed** |
|---|---|---|
| **Actual Delayed** | TP | FN |
| **Actual Not Delayed** | FP | TN |

The below formulas were used to calculate metrics of classifications namely, true positive rate, false positive precision, F score, Roc area and accuracy. We use the confusion matrix to see the performance of each classifier, the classifier used in the work is more efficient when more than 70% result have.

Number of correctly good classifier should have a high True Positive Rate (TPR) .TPR is the ratio of correctly predicted delays to the total number of actual delays; a TPR of 1 meaning that all delays were well predicted. TPR is equivalent to Recall.

$$\text{TPR}=\frac{\text{TP}}{\text{TP+FN}}\%100 \qquad (3)$$

False Positive Rate (FPR) which is the ratio of on-time samples predicted as delayed, to the total number of on-time samples; a FPR of 0 meaning that no on-time samples were predicted as delayed.

$$\text{FPR}=\frac{\text{FP}}{\text{TN+FP}}\ \%100 \qquad (4)$$

Accuracy compares how close a new test value is to a value predicted by if ... then rules, gives an overall evaluation. It is defined as the percentage proportion of correctly classified cases to all cases in the set. The larger the predictive accuracy the better the situation. This is written mathematically:

$$\text{Accuracy} = \frac{\text{TP+ TN}}{\text{FP+ FN+ TP+ TN}}\ \%100 \qquad (5)$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labelled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

$$\text{Precision } = \frac{\text{TP}}{\text{TP} + \text{FP}}\ \%100 \qquad (6)$$

F1Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

$$\text{F1 score} = 2 * \frac{\text{Recall} * \text{precision}}{\text{Recall} + \text{Precision}}\ \%100 \qquad (7)$$

ROC provides comparison between predicted and actual target values in a classification. It describes the performance of a model with a complete range of classification thresholds. ROC area varies between 0 and 1 interval. An increasing value indicates better classification, with an area of one representing perfect classification.

## 4. RESULTS AND DISCUSSION

### 4.1. Comparison of Classification Techniques

It is quite clear that the algorithm is a perfect classifier if at least on the training data, all instances were classified correctly as well as all errors are zero. However, it is not the case in reality. Therefore, we can admit that a best classifier is the algorithm with the maximum of correctly classified Instances or the minimum of incorrectly Classified Instances table 4 and Figure 3

**Table 4. Correctly classified and incorrectly classified instances for each algorithm**

| Algorithm | Correctly Classified | Incorrectly classified |
|---|---|---|
| Rules. PART | 424 | 88 |
| Rules. Jrip | 400 | 112 |
| rules. OneR | 406 | 106 |
| rules. DecisionTable | 415 | 97 |
| trees. J48 | 391 | 121 |
| trees. RandomForest | 406 | 106 |
| trees. REPTree | 409 | 103 |
| trees. DecisionStump | 403 | 109 |

It is worth to say that correctly classified instances is TP+TN, FN+FP is incorrectly classified

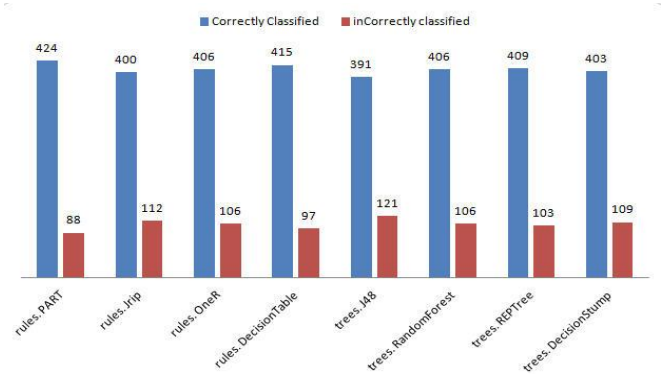


**Figure 3. Correctly classified and incorrectly classified instances for each algorithm**

From the previous analysis , It is found PART classification rules has the most correctly classified instances in the group of classification Rules techniques while REPTree has the most correctly classified instances in the group of tree classifiers. Best classifier is the algorithm with the minimum execution time. Decision Table and Part have the maximum execution time with respect to the Rules based classifiers. While Random forest classifier is the slowest among tree based classifiers. As shown in table 5 and Figure. 4

**Table 5. The running time comparisons with respect to all studied classification techniques**

| Algorithm | Running Time(second) |
|---|---|
| Rules. PART | 0.1 |
| Rules. Jrip | 0.01 |
| Rules. OneR | 0.01 |
| Rules. DecisionTable | 0.18 |
| Trees. J48 | 0.01 |
| Trees. Random Forest | 0.07 |
| Trees. REPTree | 0.01 |
| trees. DecisionStump | 0.01 |

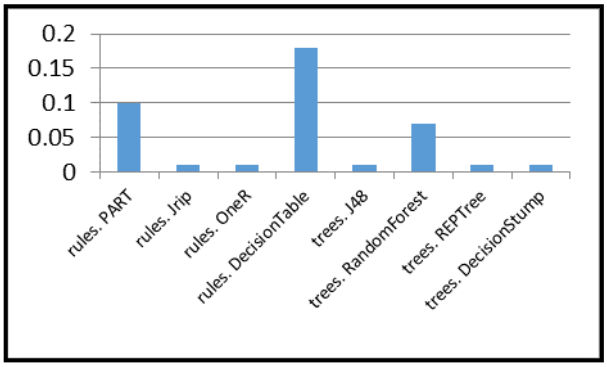


**Figure 4. Running Time (second)against name classifier**

More details can be shown in table .6. , there are more detailed about performance description via; precision, recall, true- and false positive rates, F score and Roc area. All these values are very important for comparing classify.

**Table 6 classification criteria values as taken from Weka.**

| Classifier Name | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| rules.PART | 0.828 | 0.154 | 0.845 | 0.828 | 0.831 | 0.868 |
| rules.Jrip | 0.782 | 0.241 | 0.788 | 0.782 | 0.784 | 0.787 |
| rules.OneR | 0.793 | 0.164 | 0.828 | 0.793 | 0.798 | 0.814 |
| rules.DecisionTable | 0.81 | 0.187 | 0.823 | 0.81 | 0.814 | 0.883 |
| trees.J48 | 0.764 | 0.377 | 0.763 | 0.764 | 0.747 | 0.88 |
| trees.RandomForest | 0.793 | 0.259 | 0.792 | 0.793 | 0.792 | 0.877 |
| trees.REPTree | 0.799 | 0.193 | 0.815 | 0.799 | 0.803 | 0.876 |
| trees.DecisionStump | 0.787 | 0.12 | 0.861 | 0.787 | 0.792 | 0.834 |

From table 6 and Figure 5 , Rules.Part is the best in rules classifiers with respect to TPR with percentage 82.8% and REPtree is the best for all tree classifiers with percentage 79.9%.

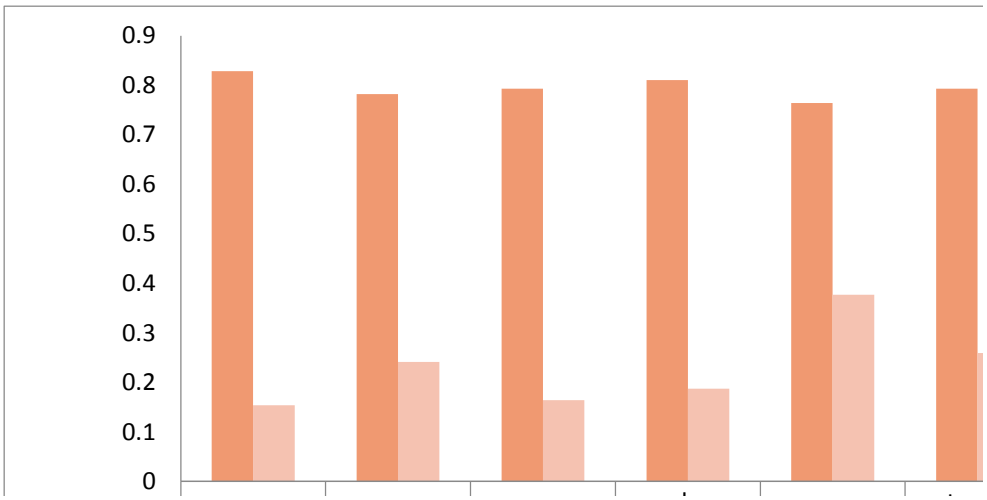


**Figure 5. Comparison of TPR and FPR for different classification techniques.**

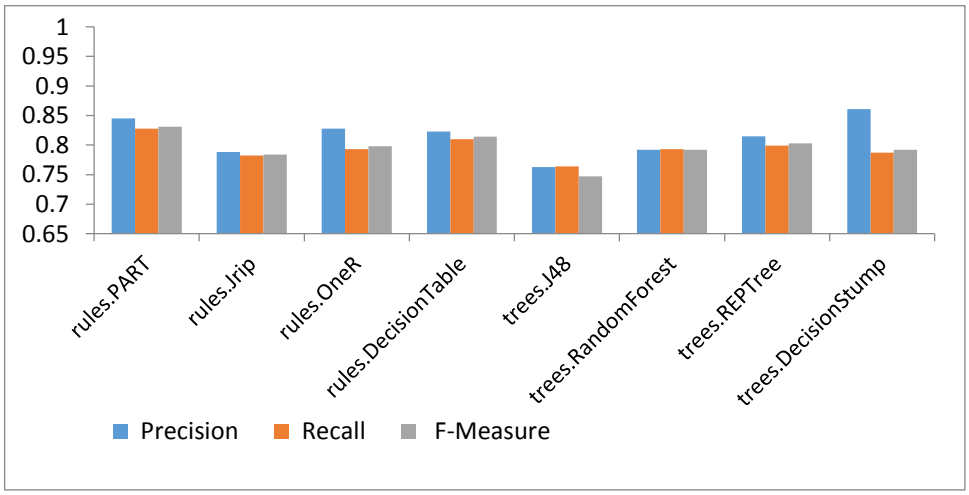


**Figure. 6. Comparison of Precision, Recall and F- score for the studied classifiers.**

F1 score is considered as the most important criteria for efficiency of classification technique. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. So F score is the best criteria for accuracy evaluation among all criteria. In our case, F1 score is the best for rules.PART with respect to rules classifiers with percentage 83.1%

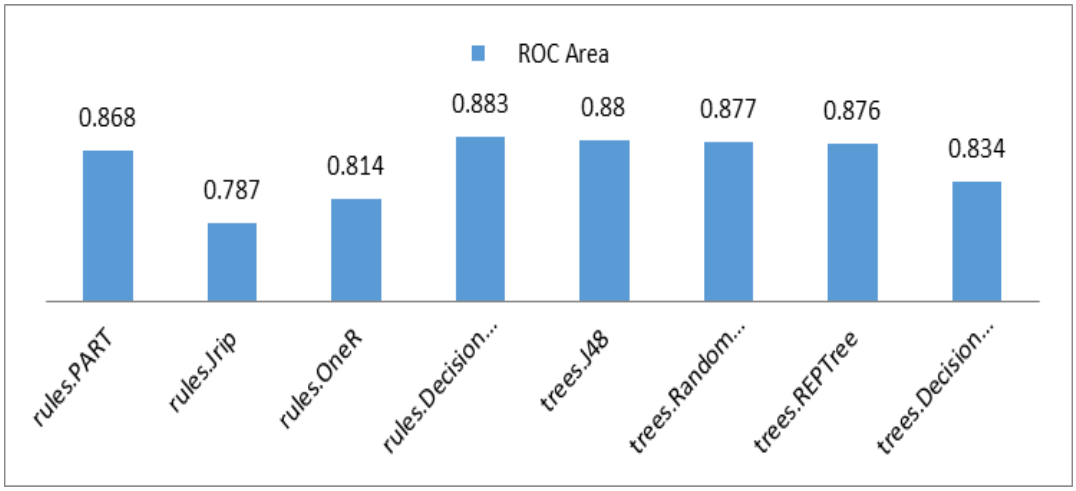


**Figure. 7. ROC Area comparison for classification techniques**

The Roc area is considered as the second important criteria for the efficiency of classification technique. It is found from figure 7 that Rules. decision has the best value for ROC area among the four rule classifiers with value0.883 and the tree.j48 has the best value for ROC area among the four tree classifiers .

### 4.2 Applying Apriori Algorithm

Here for predicting the existing data analysis, Apriori algorithm was used. In WEKA, minimum support is defined to 70% and minimum confident is 90% and the num Rule is define to 20. The resultant output show that some important information that also helpful for existing situation analysis as shown below:

IF the Flight-No=MS0910 and REG= GDL
THEN DELAY occurs,
Confidence=1 and lift=1.78.
So, they are positively correlated.

IF the Flight-No=MS0912 and REG= GDN
THEN DELAY occurs,
Confidence (0.8) lift:(1.83).
So, they are negatively correlated.

## 5. CONCLUSIONS

In this paper, the models performance in terms of classification accuracy for 4 rule-based algorithms and 4 tree-based algorithms using various accuracy measures like TP rate, F1 score and ROC area . Accuracy has been measured on the dataset by many criteria of evaluation but the most importance criteria of accuracy is F1 score and ROC area.

Thus It was found from the comparative analysis that classification Method Rules.PART performed best with classification accuracy of 83.1.%, DestionTable came out with classification accuracy of 81.4%, But OneR came out with classification accuracy of 78.4% , Jrip with 79.4% among all rule-based Classifiers. For Tree-based Classifiers , the Methods REPTree had 80.3% but method DecisionStump and random forest came out with classification accuracy 78.2% and J48 with 74.7% algorithms on flight delay dataset in WEKA.

In addition the analysis showing that JRip and OneR , j48 , REPTree and decision stump is the smallest running time on the studied classifiers with value .01 second. So we concluded from the previous analysis that REPtree is the most efficient classifier through all the studied classifiers with respect to accuracy and running time.

The accuracy of predictive model is affected by the selection of attribute. With this we can conclude that the different classification algorithms are designed to perform better for certain types of dataset. For the future work, we try to use a dataset with a huge number of instances. Certainly, it can lead us to manage big data mining technologies with the aim of achieving the precise and perfect predictions.

## Reference

A. G. S. a. S. B. Mukherjee, "Classification of Days Based on Weather-Impacted Traffic in the National Airspace System," in In Aviation Technology, Integration, and Operations Conference, Los Angeles, 2014.

A. S. R. G. a. B. S. Mukherjee, "Predicting Ground Delay Program at an airport based on meteorological conditions.," 14th AIAA Aviation Technology, Integration, and Operations Conference., 2014.

Akpinar, M. and Karabacak ,M. ,(2017). Data mining applications in civil aviation sector: State-of-art review .

Asencio, M. A.,(2012).Clustering Approach for Analysis of Convective Weather Impacting the NAS. In 12th Integrated Communications, Navigation, and Surveillance Conference, Herndon, Virginia.

B. a. M.-L. O. Wieder, ""The impact of Business Intelligence on the quality of decision making–a mediation model."," Procedia Computer Science 64, pp. 1163-1171, 2015.

Bandyopadhyay, R., J. and Guerrero, R. , “Predicting airline delays,” 2012.

Becker, B. G. (1998, October). Visualizing decision table classifiers. In Information Visualization, 1998. Proceedings. IEEE Symposium on (pp. 102-105). IEEE.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). Waikato Environment for Knowledge Analysis (WEKA) Manual for Version 3-7-8. The University of Waikato, Hamilton, New Zealand.

H. ,. J. N. a. H. P. S. Man, "Analysis of Air-Moving on Schedule Big Data based on CrispDm Methodology," ARPN Journal of Engineering and Applied Sciences, pp. 2088-2091, 2015.

http://www.cs.tau.ac.il/~fiat/dmsem03/Fast%20Algorithms%20for%20Mining%20Association%20Rules.ppt

http://www.saedsayad.com/oner.htm

https://transtats.bts.gov/

Kurniawan, R., Nazri, M. Z. A., Irsyad, M., Yendra, R., & Aklima, A. (2015, August). On machine learning technique selection for classification. In Electrical Engineering and Informatics (ICEEI), 2015 International Conference on (pp. 540-545). IEEE.

M. A. Asencio, "Clustering Approach for Analysis of Convective Weather Impacting the NAS.," in 14th AIAA Aviation Technology, Integration, and Operations Conference, 2014.

M. a. K. ,. Akpinar, in Data mining applications in civil aviation sector: State-of-art, 2017.

Man, H.,S. , Jung, N., and Hyun, P., S.,(2015). Analysis of Air-Moving on Schedule Big Data based on Crisp-Dm Methodology. In ARPN Journal of Engineering and Applied Sciences on(pp. 2088-2091).

Mukherjee, A., Grabbe, S. R., & Sridhar, B. (2014). Predicting Ground Delay Program at an airport based on meteorological conditions. In 14th AIAA Aviation Technology, Integration, and Operations Conference (pp 2713-2718).

Mukherjee, A., Grabbe, S., and Sridhar, B.,(2013).Classification of Days Based on Weather-Impacted Traffic in the National Airspace System. In Aviation Technology, Integration, and Operations Conference, Los Angeles.

N.-Y. a. J.-W. P. Kim, "A study on the impact of airline service delays on emotional reactions and customer behavior," Journal of Air Transport Management 57, pp. 19-25, 2016.

Nazeri, Z., Zhang, J. ,(2017). Mining Aviation Data to Understand Impacts of Severe Weather. In Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC.02).

Oza1, S. , Sharma, S. , Sangoi, H. , Raut, R., Kotak,V. C.,( April 2015)(FD Prediction System Using Weighted Multiple Linear Regression. In International Journal Of Engineering and Computer Science ISSN:2319-7242 Volume 4, Issue 4 on(11668-11677)

Palanisamy, S. K. (2006). Association rule based classification (Doctoral dissertation, Worcester Polytechnic Institute).

Pandey, P., & Prabhakar, R. (2016, August). An analysis of machine learning techniques (J48 & AdaBoost)-for classification. In Information Processing (IICIP), 2016 1st India International Conference on (pp. 1-6). IEEE.

R. R. F. E. H. M. K. R. R. P. S. A. &. S. D. Bouckaert, Waikato Environment for Knowledge Analysis (WEKA) Manual for Version 3-7-8., The University of Waikato, Hamilton, New Zealand.٢٠١٣ ،.

Rahman, M. S., & Waheed, S. (2017, February). Carbon emission measurement in improved cook stove using data mining. In Electrical, Computer and Communication Engineering (ECCE), International Conference on (pp. 83-86). IEEE.

Real-Time Business Intelligence at Continental Airline," [Online]. Available: http://akademi1.itu.edu.tr/oztaysib/DosyaGetir/117204/case%20II-2.pdf.

S. e. a. Choi, "Prediction of weather-induced airline delays based on machine learning algorithms," in Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th, 2016.

Sewaiwar, P., & Verma, K. K. (2015). Comparative study of various decision tree classification algorithm using WEKA. International Journal of Emerging Research in Management Technology, Volume 4, ISSN:2278-9359.

X. S. A. M. R. &. G. Wei, "Automatic structuring of it problem ticket data for enhanced problem resolution.," Integrated Network Management, pp. 852-855, 2007.

Y. J. H. &. L. D. Diao, "Rule-based problem classification in it service management. In Cloud Computing," CLOUD'09.IEEE International Conference, pp. 221-228, 2009.

Z. Z. J. Nazeri, "Mining Aviation Data to Understand Impacts of Severe Weather.," in Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC.02)., 2017.

## APPENDIX

### WEKA RESULTS

```
=== Run information ===
Scheme:       weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0
Relation:     FW
Instances:    512
Attributes:   6
              id
              FlightNo
              fdate
              REG
              Status
              Class
Test mode:    evaluate on training data
=== Classifier model (full training set) ===
Logistic model tree
------------------
Time taken to build model: 3.56 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0.01 seconds
=== Summary ===
Correctly Classified Instances         464               90.625  %
Incorrectly Classified Instances        48                9.375  %
Kappa statistic                          0.7976
Mean absolute error                      0.1462
Root mean squared error                  0.2649
Relative absolute error                 32.2157 %
Root relative squared error             55.6198 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      73.4375 %
Total Number of Instances              512
=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
ROC Area  PRC Area  Class
                 0.910    0.096    0.835      0.910    0.871
0.799    0.965     0.927     y
                 0.904    0.090    0.950      0.904    0.926
0.799    0.965     0.983     n
Weighted Avg.    0.906    0.092    0.910      0.906    0.907
0.799    0.965     0.964
=== Confusion Matrix ===
   a   b   <-- classified as
 162  16 |   a = y     32 302 |   b = n
```

## الملخص باللغة العربية

### " تقنيات التعلم الآلي لتحليل تأخر الرحلات الجوية المصرية"

**هناء ماهر محمد محمد بدوى ، شاهيناز محمود الطباخ ، هيام عبد العزيز علي الزاهد**

**قسم الفيزياء – كلية البنات – جامعة عين شمس**

لقد كان التأخر في الطيران مشكلة شائكة بالنسبة لصناعة الطيران في العالم ، لذلك هناك أهمية كبيرة للبحث عن نظام الكمبيوتر الذي يتنبأ ب تأخير الطيران. يمكن أن يكون استخلاص المعلومات المخفية من مجموعات البيانات الكبيرة من البيانات الخام إحدى الطرق لبناء نموذج تنبؤي. تصف هذه الورقة تطبيق تقنيات التصنيف لتحليل نمط تأخر رحلة الطيران في مجموعة بيانات طيران مصر للطيران.

في هذا العمل ، تم تقييم أربعة مصنّفات لشجرة القرار ، وأظهرت النتائج أن REPTree لديها أفضل دقة ٨٠.٣ ٪ فيما يتعلق بالغابات ، و Stump و J48. ومع ذلك ، تمت مقارنة المصنّفات الأربعة المستندة إلى القواعد ، وأظهرت النتائج أن PART يوفر أفضل دقة بين المصنّفات المدروسة المستندة إلى القواعد بدقة تصل إلى ٨٣.١٪. من خلال تحليل وقت التشغيل لجميع المصنفات ، استنتج العمل الحالي أن REPTree هو المصنف الأكثر كفاءة فيما يتعلق بالدقة ووقت التشغيل. أيضا ، يتم تمديد العمل الحالي لتطبيق تقنية اقتران Apriori لاستخراج بعض المعلومات الهامة حول تأخير الرحلة. يتم عرض قواعد الرابطة ويتم تقييم تقنية الارتباط.