European Journal of Computer Science and Information Technology
Vol.4, No.6, pp.15-24, December 2016
Published by European Centre for Research Training and Development UK (www.eajournals.org)

DYNAMIC DECISION TREE BASED ENSEMBLED LEARNING MODEL TO FORECAST FLIGHT STATUS

C. Ugwu[1], Ntuk, Ekaete[2]
[1]&2 Department of Computer Science
University of Port Harcourt Choba
Rivers State, Nigeria.
chidiebere.ugwu@uniport.edu.ng

ABSTRACT: *This paper explains the development of an enhanced predictive classifier for flight status that will reduce over fitting observed in existing models. A dynamic approach from ensemble learning technique called bagging algorithm was used to train a number of base learners using a base learning algorithm. The results of the various classifiers were combined, voting was done, by majority the most voted class was picked as the final output. This output was subjected to the decision tree algorithm to produce various replica sets generated from the training set to create various decision tree models. Object-Oriented Analysis and Design (OO-AD) methodology was adopted for the design and implementation was done with C# programming language. The result achieved was favorable as it was found to predict at an accuracy of 78.3% as against 68.2% accuracy of the existing systems which indicated an enhancement.*

KEYWORDS: Flight Status, Ensemble learning, Bagging Algorithm, Classification, Prediction.

INTRODUCTION

Flight delays have been a major concern in the aviation industry because of its effect to loss of productivity and confidence of passengers of the airlines involved and a major means of frustration to passengers causing loss of time and other resources [2]. This has a tremendous economic effect and dissatisfaction to airline companies and passengers. It is of best interest to an airline to ensure that flights are operated within the generally recognized 15 minutes frame, else such flight is considered delayed. Some causative factors such as man-made, mechanical or weather factors could easily be blamed for delays but most delays could actually be predicted and handled. It is significant to know that, for an airline, the importance of delay is not just its consequence on an individual airline carrier but its effect on the operational schedule [5]. [12][13] Stated that even though carriers typically blame adverse factors such as meteorological conditions for occurrence of delays, there are "systematic and predictable patterns to carriers' on-time performance", meaning that some delays are predictable and controlling them could be applied from initial stage of the plan. Tukey Twicing in 1977 started the ideology of ensemble learning approach when he combined it to linear regression models. The philosophy of the ensemble classifier is that another base classifier compensates the errors made by one base classifier. However, training the base classifier in a straight-forward manner is not going to solve this problem.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

An ensemble classifier performs better than its base counterpart if the base classifiers are accurate and diverse [3][1]. Ensemble techniques also can be used for refining the quality and robustness of clustering algorithms [6][1]. When developing models that would be used for prediction, most times, the use of very large data causes over-fitting, that is, when attributes needed than necessary are used to develop the model, that can affect the accuracy of the generated output of such models. The objective of this paper is to develop a model based on an ensemble learning approach called bagging algorithm with decision tree classifier for an enhanced prediction of flight status.

RELATED WORKS

Predicting and analyzing the causes of delay have long been a major research area in air traffic management and airline decision making. Different scholars have studied this problem from various perspectives and the following were reviewed [19], in their paper titled a theoretic research on on-time performance of airplane turnings in a web of airspace. "The propagation of unforeseen delays in aircraft rotations were found to be significant when short-connection-time policy was used by an airline at its hub airport". [13], discovered that nearly 67 million flights over 12 years of some US airline carriers did not alter their schedules to incorporate foreseeable movements in push back delays. Whereas normal scheduled time was almost equal to the average time between pushing back from the gate on departure to pulling up to the arrival gate. Carrier's schedules does not justify the fact that the usual flight leaves almost ten minutes late.

[11], published a case study of air transport delays in europe and observed that the reactionary delays is the main 'culprit' in the cause of delays amounting to 40% of the departure delay in Europe. His conclusion was that the series of reactionary delays starting in the morning have a higher magnitude and effect when compared to the ones starting in the afternoon as they propagate on average on more subsequent flights. [20], proposed a neural network model for predicting flight delays. A comprehensive data of delayed flights in the months of January, June and November in 2012 was used. Sixty (60) flight delay reasons were considered, and for each of this reason, a code and description is assigned to it. The result showed that the flight number, the airport that the flight took off from, the delay reasons, the number of delays and the aircraft type had about 26%, 24%, 24%, 17% and 10% effect on the flight delay respectively. The results proved that the considered factors have an important weight on the flights delays and that their successful management can reduce the delays flights times.

The existing techniques have not been able to solve generalization problems (capability of classifying patterns with other known patterns that share distinct features, this affect the accuracy of the output generated) and scaling problem. This paper is aimed at solving these problems by adopting an algorithm that is cable of classifying patterns with other known patterns that shares distinct features there by reducing overfitting. The algorithm is found to have potentials of averaging results from a substantial number of bootstrap samples.

MATERIALS AND METHODS

The data was obtained from the Bureau of Transportation Statistics, a Federal Agency of the United States of America, drawn from https://github.com/datailluminations/PredictingFlightsDelay . The dataset made up of records of all USA domestic flights of major carriers, for January 2004, which focused on the departure flights only. The dataset contained many attributes of which some are

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

irrelevant, the irrelevant attributes were pruned during extensive preprocessing. The resulting data was partitioned into training and test sets. A bootstrap replica of samples from the training set was generated by random sampling with replacement and multiple estimators from bootstrapped samples were also generated. The bootstrapped samples were subjected to decision tree algorithm to generate classifiers for each bootstrap replica. The classifiers voted to generate the best for the model. The model was tested with the test set to determine the accuracy of the learnt classifier model as it classified the test set into different categories. Figure 1 & 2 depicts the architectural and high level model of the system showing the processes briefly explained above. We approach this problem as a classification problem, predicting two classes - whether the flight will be delayed, or whether it will be on time. Broadly speaking, in machine learning and statistics, classification is the task of identifying the class or category to which a new observation belongs, on the basis of a training set of data containing observations with known categories. The two letter codes used for carriers as obtained from the data is shown in table 1.

Table 1: Two-Letter Codes Used To Abbreviate Carrier

Code	Carrier
US	US Airways, Inc.
RU	Continental Express Airline.
CO	Continental Air Lines, Inc.
DL	Delta Air Lines, Inc.
MQ	American Eagle Carriers, Inc.
UA	United Air Lines, Inc.
OH	Comair, Inc.
DH	Atlantic Coast Airlines.

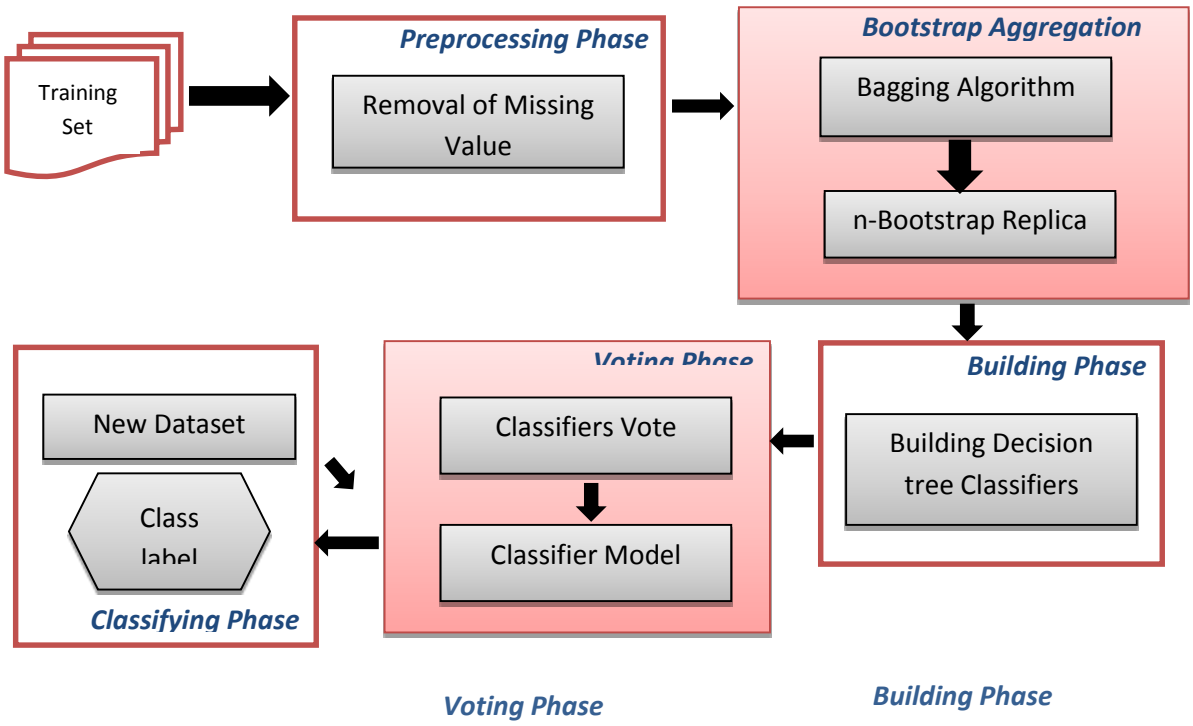

Figure 1: Architecture of the System.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

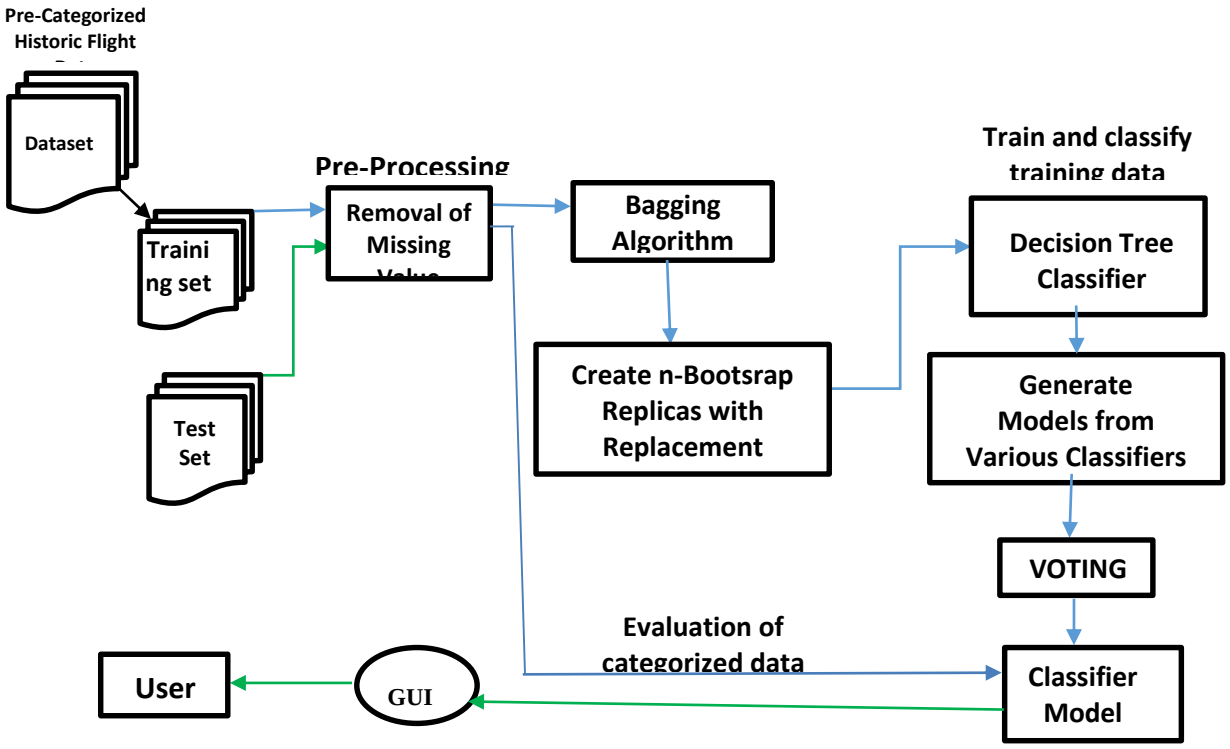

Figure 2: High level model of the system

The algorithm for these processes are outlined as follows

Step 1: Input training dataset.
Step 2:
a. Create replica sets of the same by random selection of training examples from the dataset.
b. Learning of the classifier by the bagging algorithm based on the actual training set.
Step 3: Compound classifier is created as the aggregation of particular classifiers and an example d_i is classified to the class c_j in accordance with the number of votes obtained from particular classifiers H_M.
Step 4: Train the data using Ensemble learning to produce a classifier model.
Step 5: Input: Test Set.
Step 6: Obtain test data and classify using the Ensemble built model.
Step 7: Initiate query for Flight Delays using classifiers.
Step 8: Output:

The entropy and information gain were determine from the frequencies of attributes for on-time and delay fight status from the processed data sets using C4.5 algorithm as shown in table 2. The frequency for on-time status was 32 while that of delay was 28. Table 3 shows the respective entropies and information gain of the attributes as determined from table 2.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

Table 2: The frequencies of flight status as derived from processed data

Flight Status	
On time	Delayed
32	28

Table 3: Entropies and information gain of the flight status attributes.

	ATTRIBUTES	ENTROPY	INFORMATION GAIN
1	Flight status	0.996	
2	Flight status, Carrier	0.835	0.161
3	Flight status, Day of week	0.876	0.120
4	Flight status, Destination	0.959	0.037
5	Flight status, Weather	0.955	0.041
	Flight status, Origin	0.979	0.017

A decision tree was constructed from entropy and information gain of the attributes for the processed data set as shown in figure 3. The carrier had the highest information gain and was used as the root node, the partitioning follows the next attribute with the highest information gain, this process continues until the construction of the tree terminates. The interpretation of the tree from the left subtree is that if the carrier is CO, destination EWR and origin is DCA the flight status is delayed. Also, if the carrier is DH, destination is EWF origin is IAD the flight status is on-time and so on.

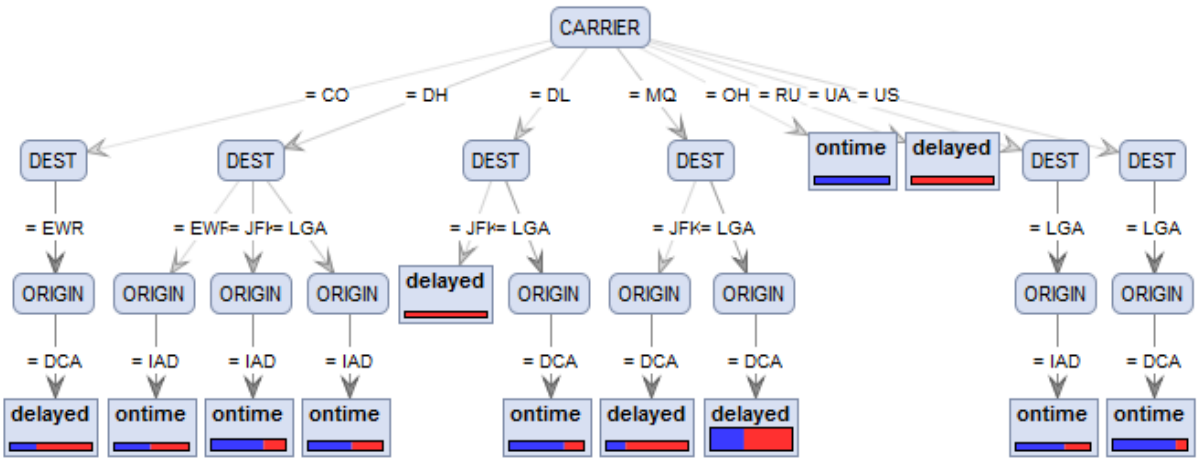

Figure 3: Constructed Decision Tree

EXPERIMENTS, RESULTS AND EVALUATION

Experiments were conducted on the model using a set of historic flight data with 1688 samples obtained from www.gittransfer.com/dataillumination s. The dataset was divided into 70% for training and 30% for testing. The evaluated results of the test set shows that the number of accurate prediction exceeds

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

the number of "miss" prediction as depicted in figure 4. This indicates the balanced nature of the predictive model.

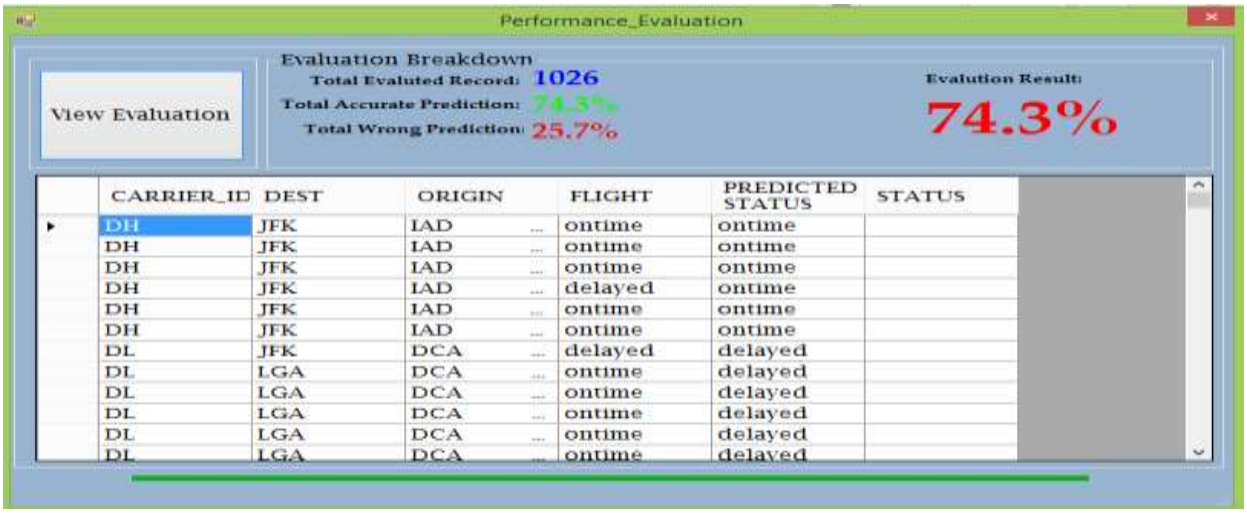

Figure 4: Performance Evaluation Results

Figure 5: On-time Flight Status Predicted Result

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

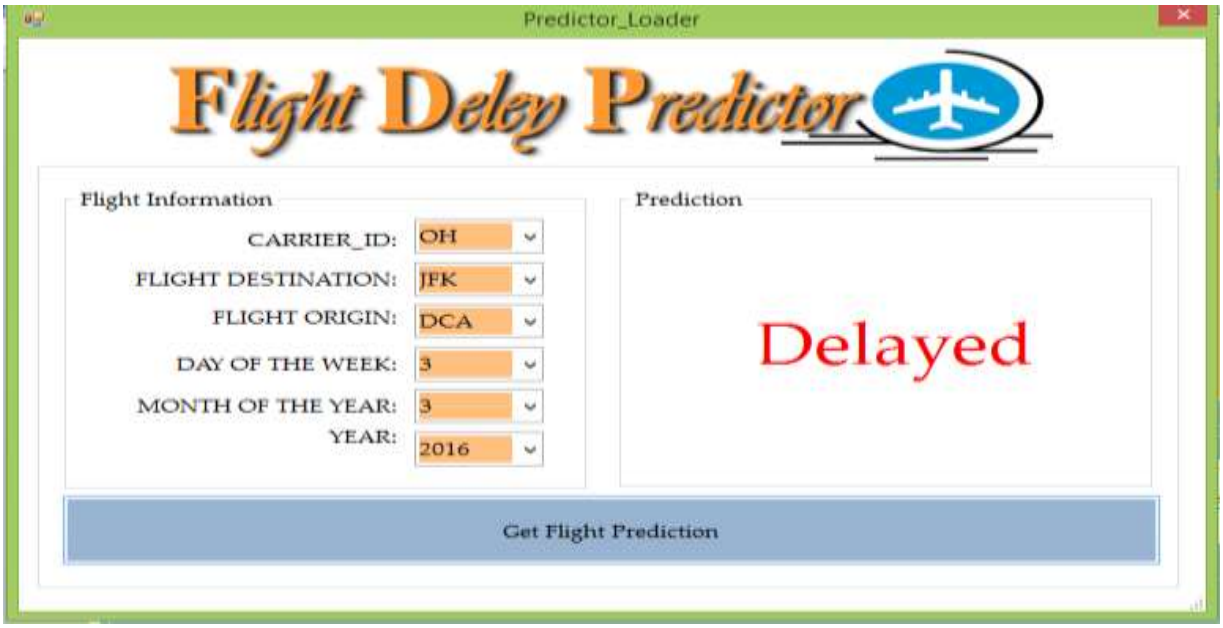

Figure 6: Delay Flight Status Flight Status Prediction

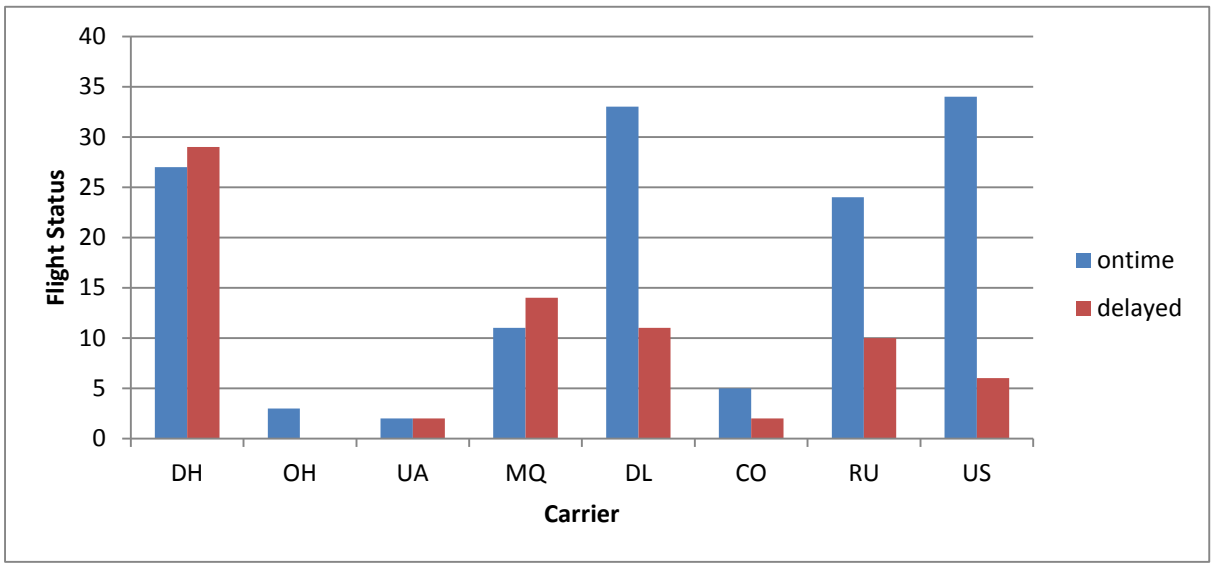

Figure 7: A graph of actual flight status against the carriers

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

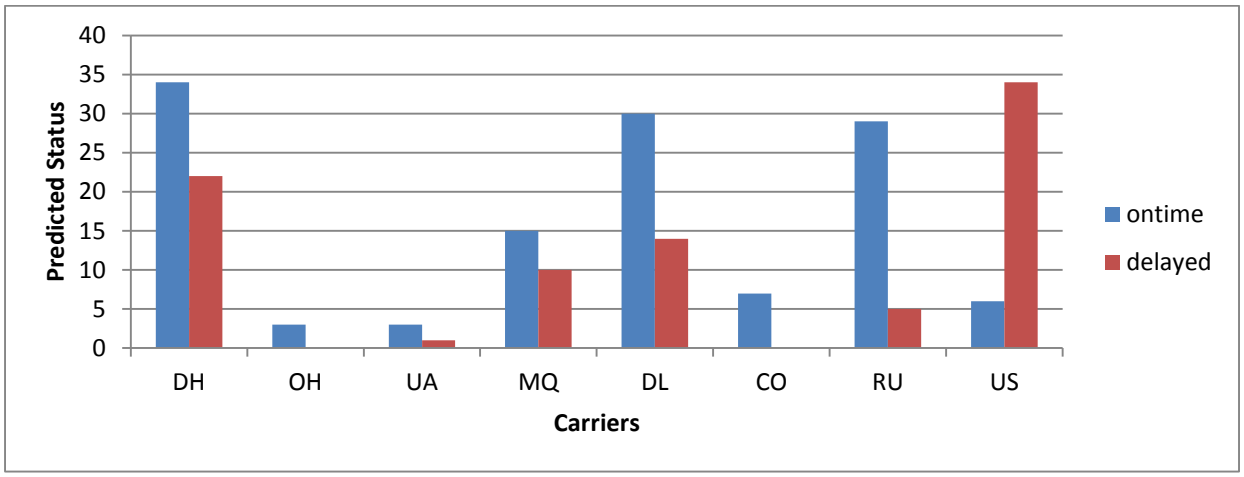

Figure 8: A graph of Predicted Flight status against the carriers

DISCUSSION OF RESULTS AND IMPLICATIONS

Figure 4 shows the performance evaluation to verify the efficacy and effectiveness of the classifier developed. It was discovered that the number of accurate predictions exceeds the number of "miss" predictions. The percentage accuracy of the system is 74.3% as against existing system as 68.6% [20]. Experimentation was done on different carrier to determine the level of accuracy of the model. The results as obtained was represented graphically in figures 7 and 8. The predicted value for each carriers US, RU, CO, DL, MQ, UA, OH, DH for on-time status is as follows 14, 20, 5, 18, 16, 0, 3 and 36 and delay status is as follows 21, 14, 2, 26, 9, 4, 0 and 20 respectively.

The predicted value against Destinations LGA, EWA, JFK of ontime status of 45, 32, 64 and delay status of 32, 10, 31 respectively. The predicted value against Origin DCA, BWI and IAD for ontime status of 48, 40, 60 and delay status of 29, 2, 35 respectively.

Figures 4- 6 shows the flight prediction application screenshots showing the predictions for ontime and delays after accepting the following inputs the carrier ID, Flight Destination, Flight Origin, Day of the week, Month of the Year and Year. A click on "Get Flight prediction" button generates the predicted output as in the interfaces shown. This shows that the application will seamlessly predict for on-time and delay of flights which can recast the confidence of the passengers on the air carries and optimize resources for both the passengers and carries.

CONCLUSION

In conclusion, the classification model developed was able to identify flights status as on-time or delayed. It was observed that many factors have impact on flight delays or on- time such factors are weather, airline carriers, and departing airports (origin) and arriving airport (destination). Following therefore the results from the model, airline carriers had the highest impact on predicting for on-time and delay for flight status. The voting ability of ensemble learning approach and extensive potentials of interpretability of decision tree model for flights delays shows a bright future for predictive models since the weaknesses of each classifier is surmounted during the averaging of the classifiers.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

REFERENCES

1. Aakash T,. & Aditya P. (2014). Improving classification of J48 algorithm using bagging, boosting and blending ensemble methods on SONAR dataset using WEKA. *International Journal of Engineering and Technical Research.* 2(9), 207-209
2. Ahmad S., C. A. (2008). Analysis of the potential for delay propagation in passenger airline networks. *In Journal of Air Transport Management*, 221-236.
3. Akhlagur, T & Sumaria, R (2014). Ensemble Classifier and Their Applications; A Review. International Journal of Computer Trends and Technology (IJCTT) India. 10, 31-35.
4. Asmita, S., & Shukla, K. (2014). Review on the Architecture, Algorithm and Fusion Strategies in Ensemble Learning. *International Journal of Computer Applications, 108*(8), 0975-8887.
5. Beatty R, Hus R, Berry J, (1998). "Preliminary evaluation of flight delay propagation through an airline schedule", Proceedings of the 2nd USA/Europe Air Traffic Management R & D Seminar, Orlando, USA. 1-4.
6. Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40*(2), 139-157.
7. Elena, I. (2012). Algorithms for Learning Regression Trees and Ensembles on Evolving Data Streams. PhD thesis. Jozef Stefan International Postgraduate School, Ljubljana, Slovenia.
8. Gopika, B. (2014). An Analysis on Esemble Methods in Classification Tasks. *International Journal of Advanced Research in Computer and Communication Engineering*. 12-14
9. Liu Y. et al. (2008). Flight delay propagation research based on Bayesian Network. *Computer Engineering and Applications*, *17*(44), 242-245.
10. Martina J (2009) The Propagation of Air Transportation Delays in Europe. Aachen Germany: Thesis Submitted to the Department of Airport and Air Transport Research, RWTH AACHEN UNIVERSITY, Germany.
11. Mayer. C, Sinai.T, (2003). "Network Effects, Congestion Externalities and Air Traffic Delays: or Why Not all Delays are evil". American Economic Review, 93, 1194 - 1215.
12. Mayer. C, Sinai.T, (2003). "Why do airlines systematically schedule flights to arrive late"? The Wharton school University of Pennsylvania, USA.
13. Montroy T.E, Ade P. A. R, Bock J.J, (2005) Cormell University Library a Measurement of CMB EE spectrum from the 2003 flight a Boomerang, New York. 1,1-17
14. Mueller, E. (2002). Analysis of aircraft and departure delay characteristics. *Proceedings of AIAA Aircraft Technology, Integratio, and Operations (ATIO) Conference.* Los Angeles, CA.
15. Shim, K. (2000). A decision tree classifier that integrates building and pruning. *Data Mining and Knowledge Discovery, 4*(4), 315-344.
16. Steiner, M., Bateman, R., Megenhardt, D., & Pinto, J. O. (2009). Evaluation of Ensemble-based probabilistic weather information for air traffic management. *Aviation, Range and Aerospace Metrology Special Symposium on weather*, 12.
17. Tu, Y. (2008). Estimating flight departure delay and delay propagation in a busy hub-airport. *The Proceedings of the IEEE ICNC'08: 2008 Fourth International Conference on Natural Computation*, (500-505).
18. Wu & Caves (2010) Flight Schedule Functionality Control and management: a Stochastic Approach. Journal Transport Planning and Technology, Abingdon. 26(4), 313 - 330.
19. ZeinEldin R (2014). A Neural Network model for flight delay: classification and prediction. Global Journal of Engineering science and Researches, India 1,13-22.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

20. Zhai, C. (2008). Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies, 1*(1), 1–141,.
21. Zhou, Z.-H. (2004). *Ensemble Learning.* Nanjing: National Key Laboratory for Novel Software Technology, Nanjing University, 1-5.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)