EXTRACTING A PHYSICAL DATA MODEL BASED ON SOME DATA MINING TECHNIQUES IN AIRLINES INDUSTRY

By

Hanaa Maher Mohamed Mohamed Badawı

A thesis submitted in partial fulfilment of the

requirement for the degree of

M.Sc. In (Physics – Computer Science and Application)

to

Physics Department

Faculty of women for Arts ,Science and Education

Ain Shames University

2018

ACKNOWLEDGEMENTS

My sincere thanks goes first to **Prof Dr. Hayam El Zahed**, for encouraging me and giving me the guidance and unlimited support during my work. I would like to express my sincere gratitude to my advisor **Dr. Shahinaz El Tabakh**, for the continuous support through my master studying, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis.

Besides my advisors, I would like to thank my examiners: **Prof. Dr. Ibrahim El henawy** and **Prof. Dr. Mohamed Hashem Abdel Aziz**, for their constructive suggestion and valuable comments, which improved my thesis quality.

My sincere thanks go to my colleagues in egyptair holding company for their wonderful collaboration. They supported me greatly and were always willing to help me.

My deepest gratitude go to **Mr Ashraf Ahmed** (GM of Customer Services Development Dept) for his leading ideas , continues support, great encouragement and brilliant thoughts to improve the We-care system. Many thanks go to **Dr Mohamed Salah** and **Eng. Hossam Said** (GM in IT Sector), for their great ideas and fruitful discussions in aviation industry.

Special thanks to my colleague, **Eng. Sarah Mofid Shafiq** (IT Developer in IT Sector), for the hard working she did in creating We-care Project , for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last three years.

Nobody has been more important to me in the flow of this work than my husband, **Eng Yasser Abdel Monem Ahmed**. I really want to thank him for his support and patience. I would like to thank my mother, **Mrs Kamilia Abdel Monem**, whose love and guidance are with me in whatever I do. Most importantly, I wish to thank my three wonderful children, **Eng Amira , Eng Omer and Mr Mostafa**, who provide unending inspiration.

ABSTRACT

In the airline industry, large scale data is generated every hour, it could be in structured format (uniform data table) or unstructured format (emails, SMS, photos, tweets, etc.) distributed among many repositories. It is very essential to get useful information from these enormous data sources to help in building promising marketing strategies. These datasets has to be combining together to build a successful marketing plan. Understanding the customer's needs is crucial to business growth. But some Customers might experience a problem during their flight, for example: the flight delay in station, fraud payment in booking process and missing bags on arrival. These customers are becoming critical obstacle in success roadmap in today's airline business.

Data mining can help in getting useful knowledge from customer's data to extract information from customer emails or to predict flight delays. In this work we choose to study two issues as subject to data mining process, these two issues affect mostly on customer satisfaction; we had the issue of *flight delay* as sample of structured Data. Also the research had the issue of *grievance handling* dataset to study as sample of unstructured.

In part one; we focused on applying classification techniques for analysing the Flight delay pattern in Egypt Airline's Flight dataset. We compared eight classifiers algorithms of the WEKA Data mining tool , four decision-tree-based classifiers (REPTree, Forest, Stump and J48) and other four rule-based classifiers (PART,DecisionTable ,OneR ,JRip). The four decision tree classifiers are evaluated and results showed that the REPTree have the best accuracy 80.3% with respect to Forest, Stump and J48. However, four decision rules based classifiers were compared and

results show that PART provides best accuracy among studied rule-based classifiers with accuracy of 83.1%.

In part two, we built a model to extract the useful information from customer grievances data, to be used as business guide. Customer grievance system in EGYPTAIR called We-care , has large feeds of data which can be collected in datasets through various channels such as e-mail, website or mobile App, call centre phone call, etc. Then the incoming datasets are analyzed and assessed by organization's support teams in order to fulfil the grievance request. And then it is assigned to related department through manual classification system. Finally, it provides solution for the issue. As grievance categorization was handled manually, it is time-consuming process; we decide to improve **We-care** system's workflow, by categorizing the grievance automatically, for better performance. Classifying methods are used to identify data into groupings of categories across the variable touch points. The system has more than 150 categories of problems, but for experimental purposes we decide to study 6 categories only. We have applied four commonly used classifiers (SVM, KNN,Naive Byes and Decision-Tree) on our dataset to classify the new grievances , and then selected the best of them to be the Grievances Classifier in our system. Among four classifiers applied on the dataset, KNN achieved the highest average accuracy (98%), then SVM (SMO) with accuracy of (97%).

The benefits of performing a thorough analysis of problems include better understanding of service performance.

Keywords-Airlines Grievance; Flight Delay; WEKA, classification ,Data Mining Text mining.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

Competitive pressure is very strong in the airline industry. The homogeneous nature of the airlines makes product differentiation very difficult and costly. As a result, airlines have shifted their focus towards understanding their customer's behaviours better that enables them to quickly respond to their individual needs and wants. To understand the customer's needs and issues is essential to business survival and growth.

The company can do it by serving the customer as special as it can, with more personalized actions, good experience or sending a follow up e-mail with any extra offers, etc. Unfortunately, many times passengers have issues about their journey, for example: *some fraud payment in booking process, the flight delay in station and missing card in the frequent flyer system*. So, it is a must to have a method to handle these issues (grievances) quickly and professionally. Fortunately there is an intelligent way to do so, by (DM) Data Mining Process. DM is getting useful knowledge from large data. In this work we had the issue of **flight delay** to study as subject to DM process of structured data .Also the research had the issue of **grievance handling dataset** to study as unstructured data sources on DM

The researcher focused on data mining techniques for predicting flight delays and classifying customer's emails as application on airlines industry. The research proposed to build models that utilized by data mining and text processing solutions that could uncover these trends. The experiment shows the comparisons between used methods according to the results to achieve the best performance. Each method was evaluated using precision, recall, F-measure.

1.2 PHYICAL DATA MODEL FOR AIRLINE INDUSTRY

Physical data model represents how a model can be built in the relational database. A physical database model describe structure of containing tables, including column details, column data type, primary key , foreign key constraints and relationships between tables. Features of a physical data model includes: description of all tables and fields, keys for indexing and are used to identify relationships between tables. Physical considerations of

data model could be quite different from the logical data model. For example, incoming data format for our work could be structured, semi-structured and unstructured data format.

For focusing on airline industry physical data model, we had the below diagram as an example as shown in figure 1.1.

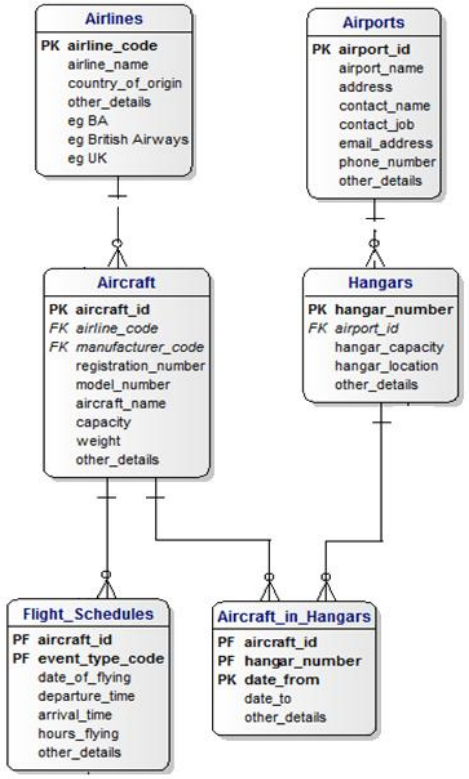

FIGURE 1.1 AIRPORT DATA MODEL, [1]

1.2 PHYICAL DATA MODEL FOR AIRLINE INDUSTRY

1.2.1 Data model for flight delay

Flight delay has been the largest problem to the world's aviation industry, so there is very important significance to research for computer system for predicting flight delay propagation. Extraction of hidden information from large data sets of raw data could be one of the ways for building predictive computer system. Application of classification algorithms for flight delays is very important issue for aviation experts. This research describes the application of different classification techniques for analyzing the flight delay pattern in Egypt National Airline's Flight database. The research aim to build predictive models for flight delays based on various attributes of particular flights. These models will make us able to predict when we are most likely to encounter delays and increase the knowledge for passengers advising them on the most efficient ways to travel. The method steps is shown in Figure 1.2

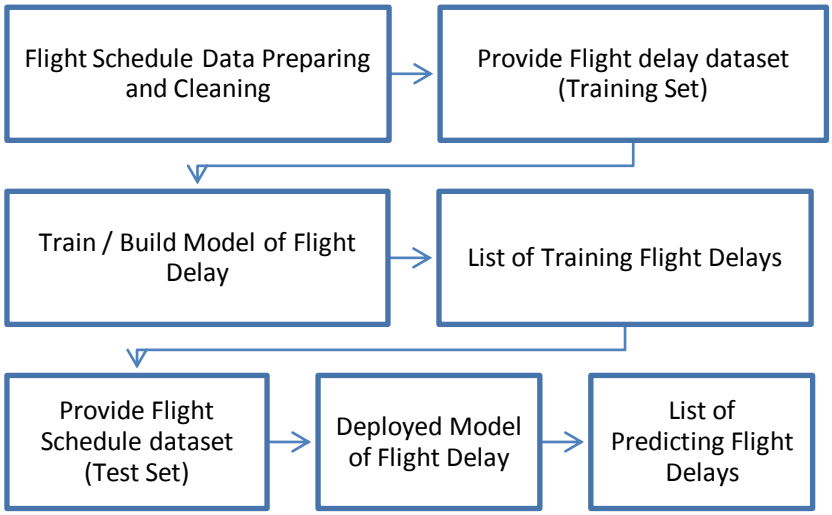

FIGURE 1.2 DATAMINING EXTRACTING MODEL FOR FLIGHT DELAY

1.2.2 Data model for grievances handling

A grievance handling system is a system that manages the process of how organizations handle, manage, respond and report to client's grievances. The manual categorization of the large number of grievances is extremely difficult, time consuming, expensive, is often not feasible and lead to un-satisfaction of the customer. So to improve the quality of service the system need to minimize the processing time by replacing the manual categorization with automatic categorization, there must be an intelligent method to do so. In order to extract a physical data model for airlines grievances, the research determines the input, process and output phase of this model as shown in figure 1.2.

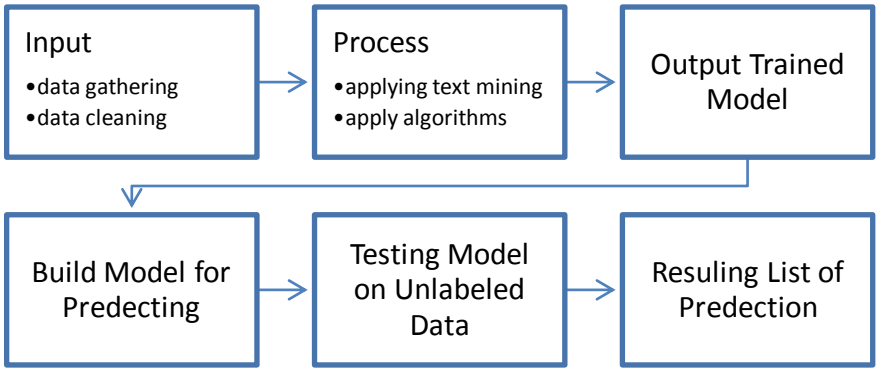

FIGURE 1.3 DATAMINING EXTRACTING MODEL FOR GRIEVANCES HANDLING

1.3 RESEARCH GOALS

1.3.1 Flight delay issue

Using machine-learning algorithms, we developed two types of models. First, we used classification learning to create models that predict whether or not a given departure will be delayed or on-time. Ultimately, we compare eight algorithms for building models that can predict flight delays with high accuracy.

1.3.2 Grievance handling issue

Provide high quality system to manage the customers' grievances in EGYPTAIR's. Pass the limits of the existing manual grievances systems and quick responding and minimize the required time of processing the incoming grievances by using automated categorization

that analyze the English text contents and predict the category. Also predict which problem is most likely to happen, that will focus on the weak point.

1.4 RESEARCH METHODOLOGY

We used data mining techniques to classify, build models and predict results for 2 datasets (Flight delays, Grievances Content). we analyze them and study their limits, and then design and test the models for each one Then comparing some data mining algorithms and select the best algorithms that achieved the best performance. The work process involves following steps

1. Collect dataset from different sources
2. Clean dataset (fully described in literature chapter)
3. Create training data set.
4. Identify useful attributes for classification (Selection analysis).
5. Learn a model using training examples in Training set.
6. Use the model to classify the unknown data samples.
7. Evaluate best method or algorithm that gives best results.

1.5 THESIS ORGANIZATION

The thesis is composed of five chapters. Chapter 1 presents problem introduction. Chapter 2 presents Literature Review for theoretical aspects of the research and related works. Chapter 3 flight delay problem description, proposed methodology and Results. Chapter 4 presents the proposed grievances handling model, Methodology, experimental results. Finally, Chapter 5 shows conclusion and the possible future work.