Designing a Predictive Model for Train Arrival Time Management, Using Data Mining Approach

A Thesis Presented

By

Yonas Abebe Tamru

To

The Faculty of Informatics

Of

St. Mary's University

In Partial Fulfillment of the Requirements

For the Degree of Master of Science In

Computer Science

August, 2019

St. Mary's University
School of Graduate Studies
Faculty of Informatics
Department of Computer Science

Designing a Predictive Model for Train Arrival Time Management, Using Data Mining Approach

Name and Signature of Member of the Examiner:

Advisor Million Meshesha (PhD) Signature_____Date_____

Examiner Tibebe Beshah(PhD) Signature_____Date_____

Examiner Getahun W/Mariam(PhD) Signature___ _Date_____

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been fully acknowledged.

Yonas Abebe Tamru
Student

_____

Signature
Addis Ababa
Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Dr. Million Meshesha
Advisor

_____

Signature
Addis Ababa
Ethiopia

Aug 2019

ACKNOWLEDGMENTS

First and foremost, I would like to thank my GOD for all of things happened in my life. Secondly, the success of this thesis is credited to the extensive support and assistance from my advisor Dr. Million Meshesha. I would like to express my grateful gratitude and sincere appreciation to him for his guidance, valuable advice, constructive comments, encouragement and kindness to me throughout this study.

Thirdly, I want to thank Dr. Getahun from computer science department head of st mary who helped me to get the assistance for this study. And my special thanks go to Addis Abeba light railway transit controlling staff; specially Bisrat Mehary and Mikayas Alemayew; who helped me in providing necessary information and materials which are crucial in this study and gives me support in different area during the research.

Fourthly, I would like to thank all of you who support me to complete this thesis work as well as my study, even if I didn't mention your name here!

Finally, I would like to thank my family who supported and encouraged me throughout the time of my study and the research work.

Thank you!

Table Contents

List of Figures

List of Tables

List of Acronyms

AALRT: Addis Ababa Light Railway Transport

ANN: Artificial Neural Network

ARFF: Attribute-Relation File Format for processing

ATS: Automatic Train Supervision

C&RT: Classification and Regression Trees

CNR: Chinese train manufacturer

CRISP-DM: Cross-Industry Standard Process for Data Mining

CSV: Comma Separated Value

DM: Data Mining

DT: Decision Tree

E-W: East–West

FIFO: First-In-First-Out

FOFI: First-Out-First In

GNSS: Global Navigation Satellite System

GNU: General Public License

KDD: Knowledge Discovery in Databases

KM: Kilo Meter

MPC: Model-Predictive Control

MS: Micro Soft

OCC: Operating Controlling Centers

PTC: Positive Train Control

QOS: Quality of Services

SEMMA: Sample, Explore, Modify, Model, and Assess

S-N: South-North

WEKA: Waikato Environment for Knowledge Analysis

Abstract

Access to transport service is critical to the development of all aspects of a nation train arrival time management including staff behavior, affordability, and ticket payment system and also somewhat satisfied with reliability, comfort, safety and security accessibility and availability. However, this transport services are not free from problems. Passenger loading is the main problems of all railway services operators. This research therefore aims to design a predictive model that can determine Train Arrival Time Management of Addis Ababa light transit operating control center data. To overcome the drawback of simple statistical method, we proposed the use of data mining techniques, for the data analysis for train arrival time management.

The study follows hybrid data mining process model. After experiment survey for problem understanding, selected around 20,000 records of three years from OCC data. After eliminating irrelevant and unnecessary data, a total of 15040 datasets with 12 attributes are used for the purpose of conducting this study. Data preprocessing was done to clean the datasets. After data preprocessing, the collected data has been prepared in arff format suitable for the DM tasks.

The study was conducted using WEKA software version 3.8 and three classification techniques; namely, J48 algorithm from decision tree, Naïve Bayes and JRIP, algorithm from rule induction. As a result, J48 decision tree algorithm with Percentage split (66%) registered better performance of 95.5612% accuracy.

As a result, the study showed that scoring high value in speed, headway time and passenger loading attributes in train arrival time management are determinant factors for the arrival time success in the AALRT. Besides, the study revealed that other regions train arrival time management is more associated with success rate.

Keywords*:* Data Mining, Knowledge discovery, OCC, QoS, Classification, Hybrid.

CHAPTER ONE

INTRODUCTION

1.1. Background

The railway industry plays a vital role in many countries. All railway companies try to achieve more regular and reliable train services, in order to satisfy their customers [1].In railway transport, the railway timetabling is one of the main factors to have a good traffic management system. Generally speaking, a good train timetable can enable to use resources optimally, (like time, human power, electric power consumption, the rail infrastructure, the trains, etc). It also minimizes possible traffic accidents, increases the attractiveness of railways, minimizes possible delay; announce train services to potential customers, there by the customers at large can be satisfied [2].

Having flexible and reliable train scheduling and routing is a central part of the planning process to have good traffic management, comfort, costs, and to maintain the quality of service demand for a railway company. Its design is concerned with the problem of selecting a set of lines and determining the headway, the arrival time, the traveling time and the departure time for a set of trains at a sequence of stations [3]. Railway timetable is a program for space and time-wise running of railway passenger and/or freight traffic on a railway line. A timetable for a railway line or railway network, at least contains a list of stations per railway line with the arrival and departure times for trains. Operating economy wise the timetable is the result of the traffic production planning for a given time period i.e. the validity period for the timetable [2]. Obviously, the timetable is not the only plan that needs to be composed in order to operate a railway system but also areas like demand estimation, rail line planning, rolling stock scheduling and crew scheduling too. This also indicates the dependencies between the timetabling process and other railway planning processes.

Currently Ethiopia is constructing a railway transport system, which is considered as a solution for the existing transportation problem. Railway infrastructures, as any other kind of infrastructure, are affected by the time waiting process. Therefore, accuracy and reliability of such time waiting inspections are imperative, and advanced quantitative methods to locate the

passenger overflow are of pressing importance. In many other nations modern security systems used by time waiting protection applications including a set of different references technologies integrated by appropriate management systems. Such systems are highly dependent on human operators for supervision and operation. Process mining is a method for discovering processes and extracting information about them from event data using a process model [4].It combine's data mining with domain knowledge about the specific processes that are analyzed. The principle idea of the concept is to extract the necessary information from large data sets and obtain an output containing clean and structured data ready for analysis.

This study aims to employ data mining technology. Rygielski, Wang and Yen, [5] defined data mining as a sophisticated data search capability that uses statistical and machine learning algorithms to discover patterns and correlations in data. Data mining can find out and extract useful patterns from information which are hidden in large databases to reveal the unseen relationships. Wang and Yen [5] stated that data mining is one of the step who knowledge discovery process. Data mining help to construct predictive and descriptive models [5]. Predictive modeling permits the value of one variable to be predicted from the known values of other variables. Classification, Regression, Time series analysis, Prediction etc. Are some examples of predictive modeling. As Tan et al [6] indicated many of the DM applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or Classes. It is a supervised learning method because the classes are predefined before the examination of the target data.

1.2. Addis Abeba Light Railway

Addis Abeba Light Rail Transit Project is a semi-closed urban rail transit system. To effectively solve the problem of urban transportation, especially that of the downtown area, the government of Ethiopia decided to build a light rail in the city of Addis Abeba. Currently this project has operationalized two lines, the east–west (E-W) line and the south-north (S-N) line. About 3 km is the sharing Section for both (E-W) route and (N-S) route, which has the greatest passenger [7]. The E-W line starts from Ayat and ends at Torhailoch. The total length is 17.4 km. There are 22 stations, among which 5 are elevated stations, 1 underground station and 16 ground stations. The control center (commonly used by both lines) is temporarily considered to be placed inside the

depot. The S-N line phase starts from Menelik II Square and ends at Kaliti. The total length is 16.689 km [7]. There are 22 stations, among which 9 are elevated stations (5 common stations at the common line), 2 underground stations and 11 ground stations [7].

A longer stopping time or a greater number of stops increases the total travel time, which may decrease passengers' satisfaction and transportation efficiency. This could lead to lower fare revenue and a reduction in the total profit to the operator of the railway line. Several studies have reported that the total loss of travel time for passengers is linearly related to the train service level [8]. Hence, the profit of a rail line or network is expected to have an inverse relationship with the number of stops and the stopping time. Being railway scheduling such a rich and complex problem, it is necessary to define all the model's limitations, assumptions and inputs. This model considers a single railway line that serves trains travailing in both directions. The railway is formed by track segments, which make the connection between all the meet points. In this context, meet points include not just stations, but also siding or any location where two trains may cross simultaneously. So, for this model, trains are only able to meet or pass at meet points.

1.3. Statement of the Problem

The real-time dispatching process can be approached by retiming trains, i.e., by using running time supplements that are included in the timetable. The train sequence at junctions and merging points may also be adjusted (reordering) to the actual delay situation for example, the trains can be rescheduled in the order they arrive. From the dispatcher to running time for train every 15 minute intervals and the passenger serves as 30 second form the station. A further degree of freedom is to change locally the route used by a train (rerouting), for example, an empty platform can be used instead of causing a delay while waiting for a still occupied track [9].

Delay in train arrival time occurs due to variability of process time, capacity, passenger overflow, single train, power off, one way of railroad, lack-of train, over loading the train, signal problem, the actual delay of a train and synchronization processes, and dependence on availability of infrastructure, rolling-stock and crew [9]. Small deviations from the scheduled process times as a consequence of variability result in disturbances. This described disturbances as structural deviations that reflect stochasticity of process times due to internal and external factors. The issue of minimizing their impact on timetable reliability is addressed both in the tactical and operational

control and planning levels. Time supplements and buffer times are added in the process of timetable construction Moreover; operational traffic control aims to minimize deviations from the timetable during real-time operations. On the other hand, disruptions are caused by major deviations of timetable and logistic schedules due to failures of infrastructure, rolling-stock, line blockages, extreme, weather conditions.

Major disruptions in general do not happen frequently and they are resourced by special disruption and incident management strategies [10]. Primary delay is an extension of the scheduled process time caused by a disruption within the process. Primary delays may result in secondary delays. Occurrence of secondary delays is called delay propagation. Secondary delays occur as a result of interdependence's between trains, i.e. due to route conflicts or waiting for scheduled connections. They may be a consequence of primary and secondary delays but also due to early trains and timetable errors. Capacity constraints are a common reason for secondary delays. Extended running time of a train may cause knock-on delays to successive trains on the saturated line. Similarly, extended dwell time in a station often results in consecutive delays of other trains in busy stations due to occupied platform track or station routes.

The problem is first reduced by exploiting the fact that a relative train order cannot change on open track lines. This model is further extended with a speed coordination component [11]. The speed profiles of hindered trains are adjusted to model braking and reacceleration. The conflict resolution and speed coordination components are integrated into a closed-loop framework where the feasibility of the solution computed by the conflict resolution part is verified after computing the adjusted speed profiles. The model was applied on a realistic case study of a busy traffic control area. Optimal results for different kinds of disruption scenarios were obtained in a short time. Finally, the relevant contributions from the field of real-time rescheduling were discussed from the perspective of their applicability for rescheduling traffic over large scale networks.

After predicting the expected conflicts and delays that make the planned timetable infeasible, traffic control needs to find a new feasible schedule for train operations. That procedure is called real-time rescheduling [11]. It is performed both on the level of local and network traffic control. Operational requirements of rescheduling tasks of Traffic controls are summarized network traffic controller’s deal with disruptions and disturbances with effects that can incomplete and affect the global network performance. They need to take into account macroscopic constraints of railway

traffic, such as running times of trains between timetable points, dwell times, minimum headway times between successive dependent events in timetable points, and synchronization constraints. The objectives of rescheduling on this level depend on the traffic situation and the magnitude of disruption. They vary from minimizing the deviations from the published timetable in case of disturbances, to maintaining passenger over flows and maximizing throughput in case offline blockages and major incidents.

An important task of network traffic controllers is to coordinate the controllers on the passenger over flow whose cognitional awareness is limited to their own area, and try to minimize delay propagation in multiple areas. Apart from changing the scheduled times and relative train orders defined in the timetable, network traffic controllers may reroute trains over different lines, cancel or add trains, implement short turns, skip-stop operation. Local traffic controllers manage route conflicts, delays and disturbances within their control area. Microscopic train routes, signaling and interlocking principles need to be considered by traffic controllers on this level. The dispatchers and signaler in a local traffic control area implement rescheduling decisions. That includes changing the relative order of trains that simultaneously claim the same block (platform track or station route), changing a train route in a station area or modifying departure times. The objective is to minimize the deviation from a target trajectory set by the hierarchically higher network control level. Process mining is a method for discovering processes and extracting information about them from event data using a process model [4]. It combines data mining with domain knowledge about the specific processes that are analyzed. The principle idea of the concept is to extract the necessary information from large data sets and obtain an output containing clean and structured data ready for analysis.

This study therefore tries to apply Data Mining techniques for constructing a predictive model that helps in determining the Train arrival time management. To this end, the study attempts to explore, investigate and answer the following main research questions.

- What are the suitable attributes to describe the problem under study?
- Which classification algorithm is suitable for train arrival time predictive modeling?
- To what extent the model works in the prediction processes?

1.4. Objective of the Study

1.4.1. General objective

The general objective of the research is to design a predictive model for train arrival time management of Addis Abeba railway transit by using data mining techniques.

1.4.2. Specific objectives

In order to achieve the general objective of the study, the following specific objectives are formulated,

- To understand the train arrival time management based on the review of select predictive of Addis Abeba railway transit area.
- To understand and get familiar with the data, identify data quality problems and prepare quality data for experimentation.
- To select data mining classification algorithm for train time table data analysis.
- To design an optimal predictive model through an extent experimentation.
- To design develop and evaluate a prototype.

1.5. Scope and limitation of the study

The scope of this research is to designing a predictive model that designing train arrival time management from Addis abeba light railway transit data.

The coverage of this research would be on Addis Abeba railway transit only. Train arrival time management has been using “traffic solution” with underlying oracle database for the last 3 years from Decmber.20, 2016 to Decmber.20, 2018.

In this research a predictive model is used. A predictive model makes a prediction about values of data using known results found from different historical data. Prediction methods use existing variables to predict unknown or future values of other variables. Predictive model includes classification, prediction, regression and time series analysis [6]. In this research classification data mining approach is used to predict train arrival time management from AARLT OCC.

This research was aimed to include all important information for solving the study problem. However, some attributes like, Ballast, Sleepers, Rail, Curves, Gauge, Turnout, Rail Welding, Discount Fee Passenger, Lrv Driver, Engineering Driver, and Block time are not included in this study because the data was not available. As described above this research only attempted to

apply DM techniques in predicting and take final table 3.7 data. Train arrival time data in AALRT was selected because of its wide coverage from the other services that the company provides. However, researches can also be conducted in other sections of the company other than train arrival time management.

Although this research was aimed to include all branches of the company throughout the country, the data used in this study was only collected from the only Addis Ababa light rail transit. This only branch was chosen because of their activeness in the company currently. Due to time and financial matters this research didn't include the data from the regional branches of the company. So, further research can be conducted including the data from these branches.

1.6. Significance of the Study

In this research, the applicability of data mining techniques in the railway industry to build models that can be used for traffic control based on the value they contribute. Based on this, the subsequent benefits can be gained from the finding of this study.

Primarily, the researcher gained an experience of conducting a research as this study was conducted for academic purpose; hence, the finding of this study, could motivate other researchers to conduct further researches in the area.

Secondly, the result of the study could help train arrival time to manage customer follows and gain business advantage. It may also improve the railway business process. Thus, the railway, in this study train arrival time management, besides, has done of the study can be used by regulator body dealing with railway businesses.

Thirdly, the findings of the study can be used by other organizations dealing with train arrival time control and in this research how the application of data mining technology is of great significance to ensuring the safety of railway operation and enhancing the core competitiveness of railway.

The finding of this study can be used by railway to increase the quality of service given to its policy holders in order to maintain the standard. Moreover, the study findings can provide insight for further researches to apply data mining technologies to advance railway industry.

1.7. Methodology of the study

The methodology is the general strategy that outlines the way in which the research is to be undertaken and, among other things, identifies the methods to be used in it. Methodology in science and research because of how the scientific method is structured.

It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge. Typically, it encompasses concepts such as paradigm, theoretical model, phases and quantitative or qualitative techniques [12].

For conducting this research the WEKA (Waikato Environment for Knowledge Analysis) version 3.8.0 (for Mac OS) DM software is chosen. Weka is chosen because of its widespread application in different DM researches and familiarity of the researcher with the software. Weka, a machine-learning algorithm in visual C#, is adopted for undertaking the experiment. Weka constitutes several machine learning algorithms for solving real-world DM problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from one's own Java code. Weka is open source software issued under the GNU General Public License. The Weka DM software included classification, clustering, association rule 'learner, numeric prediction and several other schemes. In addition to the learning schemes, Weka also comprises several tools that can be used for datasets pre-processing [13].

In Data Mining, Methodology is a way that deals with data collection, analysis and interpretation that shows how to achieve the objective and answer the research questions. Hence, in order to achieve the general and specific objectives of the study the following methods are used.

1.7.1. Research design

This study follows experimental research. Experimental research, often considered to be the "gold standard" in research designs, is one of the most rigorous of all research design. In this design, the researcher can attempt to maintain control over all factors that may affect the result of an experiment. We use the six-step process of Hybrid data mining process model. This model was developed, by adopting the CRISP-DM model to the needs of academic research community. Unlike the CRISP-DM process model, which is fully industrial, the Hybrid process model is both academic and industrial.

According to Cios, Swiniarski and Kurgan [14], the hybrid data mining process model are enhanced the knowledge discovery process by combining the academic and industrial models in

data mining research. The development of hybrid models was adopted from the CRISP-DM model as its can be used for academic research. Thus, this model is research-oriented, which present data mining step than the modeling step. The six steps of hybrid process model allow a number of feedback mechanisms. Moreover, the knowledge discovered in the final step for a specific domain may be applied in other domains.

The six steps of hybrid data mining process model include understanding of the problem domain, understanding of the data, preparation of the data, data mining for predictive modeling, evaluation of the discovered knowledge and use of the discovered knowledge.

1.7.2. Understanding of the problem domain

This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

In this research, in order to identify, define, understand and formulate the problem domain different discussion points reflect the train arrival time management are used, so as to closely works with the domain experts of AALRT, then determine attribute feature selection and understanding business processes. In collaborating with the domain experts, the Operating Controlling Center (OCC) data is selected as the main source of data collection. Based on the insight and knowledge gained about the domain of railway business, data mining problem is defined.

1.7.3. Understanding of the data

This step is used for collecting sample data and deciding which data is important. Data are checked for completeness, redundancy, missing values, plausibility of attribute values. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

In this research, in order to understand the data, brief discussion on the OCC data was conducted with the domain experts of AALRT. The discussion includes listing out initial attributes, their respective values and evaluation of the importance of the OCC data for this research.

Oracle Database: The train arrival time records were very huge data and it is saved in damp file. To read this file, Oracle Database is used and retrieves necessary data in collaboration with domain experts.

Out of the total 15040 datasets, 67% of them belong to the class of Punctual and 33% of them are Delay classes. With respect to track line where they found Up line and Down line holds 45%, and 55% respectively. To understand the nature of the data, descriptive statistics is used. Based on the appropriateness of the problem domain, document analysis, literatures reviewed in chapter two and based on the information obtained from domain experts attributes which are relevant to the study are chosen.

1.7.4. Preparation of the data

In this step, the data going to be used are prepared to apply the DM methods. It consist of tasks such as sampling, testing the correlation and significance of the data, cleaning the data, checking the completeness of the tuples, handling noisy and missing values. Then, the dimensionality of the data is reduced by feature extraction and selection algorithms. This step also comprises the derivation of new attributes, summarization of the data. Finally, the datasets that meet the input requirements of DM tools stated in the first step are selected for modeling purpose. WEKA data mining tool is selected for preprocessing task such as discretization, normalization and attributes selection. With WEKA, numeric attributes are discredited in order to replace the labels attribute by intervals, which is easy to interpret and consistent to apply different DM techniques. In addition, numeric attributes are also normalized to prevent bias when attributes have very different ranges. Besides, the dimensionality of the data was evaluated using information gain evaluation method of WEKA. MS-Excel is also used for data preparation; pre-processing and analysis by using it functions such as sort, formulas (to compute CLV), filter, find and replace and so on. Besides, it is used for documentation purpose. The data sets in MS-Excel should be converted CSV (Comma Separated Value) and ARFF (Attribute-Relation File Format for processing). The input of the data into data mining applications proved to be simple with the conversion of an Excel spread sheet datasets into a CSV file format and then an ARFF file format for modeling purposes.

1.7.5. Data mining for predictive modeling

The main purpose of this research is to develop a predictive model for identifying the train arrival time management using data mining techniques. In this research classification technique is selected because the datasets in OCC data has clear and simplified labeled class.

WEKA version 3.8.1 DM tool has been used to create models using the classification algorithms, such as decision tree and rule induction. WEKA version 3.8 is chosen because [27]:

- It is easy to use by a novice user due to the graphical user interfaces it contains
- It is very portable because it is fully implemented in the visual C# programming language and thus runs on almost any computing platform
- It contains a comprehensive collection of data preprocessing and modeling techniques.

1.7.6. Evaluation of the discovered knowledge

In this research different classification models are developed and evaluated using training and testing dataset. The experimental output of the classification models is analyzed and evaluated for performance accuracy using confusion matrix.

After performing the confusion matrix, the results are evaluated by measuring its Accuracy and Error Rate. Furthermore, the effectiveness and efficiency of the model is also computed in terms of recall and precision.

Here in collaboration with domain experts of train arrival time managements, understanding the results of the models, checking whether the discovered knowledge is new and interesting, and checking the contribution of the discovered knowledge is evaluated.

1.7.7. Use of the discovered knowledge

After evaluating the discovered knowledge, the last step is using this knowledge for the industrial purposes. In this step the knowledge discovered is incorporated in to performance system and take this action based on the discovered knowledge.

In this research the discovered knowledge is used by integrating the user interface which is designed by visual C# programming language with a Weka system in order to show the prediction of train arrival time management.

Visual C# programming language is chosen because [15]:

- Instead of a lot of noise (EJB, private static class implementations, etc) you get elegant and friendly native constructs such as Properties and Events.

- It supports native resource-management idioms (the using statement). Java 7 is also going to support this, but C# has had it for a way longer time.
- It's deeply integrated with Windows, if that's what you want.
- It has Lambdas and LINQ, therefore supporting a small amount of functional programming.
- It allows for both generic covariance and contra variance explicitly.
- It has dynamic variables, if you want them.
- Better enumeration support, with the yield statement.
- It allows you to define new value (or non-reference) types.

.

CHAPTER TWO

LITERATURE REVIEW

2.1. Overview

Literature review has been conducted to assess concepts, techniques and applications of data mining technology and to get domain knowledge about the problem. In order to get a deeper understanding of how Train arrival time management is performing its transport activity, Train arrival time procedures, relevant documents and Train arrival time management website were reviewed. In order to select modeling techniques that best suit the problem different data mining books, research works, journals and published articles on the application of data mining in train arrival time and time prediction were reviewed. An extensive survey of relevant train arrival time scheduling and rescheduling approaches can be found in [16].

2.1.1. What is Data Mining?

Data mining is the process of extracting or mining knowledge from large data sets. But, knowledge mining from data can describe the definition of data mining even if it is long. data mining have similar or a bit different meaning with different terms, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [17]. According to Olson [18], Data mining also considered as an exploratory data analysis. Generally, Data mining uses advanced data analysis tools to find out previously unknown (hidden), valid patterns and relationships among data in large data sets. It is the core field for different disciplines such as database, machine learning and pattern recognition.

It is a common practice to refer to the idea of searching applicable patterns in data using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms KDD and data mining are used widely [19].

Knowledge discovery was coined at KDD to emphasize the fact that knowledge is the end product of a data-driven discovery and that it has been popularized in the artificial intelligence and machine learning fields. According to Fayyad [19], KDD and data mining are two different terms. KDD refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in the process. Furthermore, data mining is considered as the application of specific algorithms for extracting patterns from data.

2.1.2. Why Data Mining?

Nowadays, massive amount of data is produced and collected incrementally. The possibility of gathering and storing huge amount of data by different organizations is becoming true because of using fast and less expensive computers. When organizational data bases keep growing in number and size due to the availability of powerful and affordable database systems the need for new techniques and tools became very important. These tools are used for helping humans to automatically identify patterns, transform the processed data into meaning full information in order to draw concrete conclusions. In addition, it helps in extraction of hidden knowledge from huge amount of digital data [20].

2.2. The Data Mining Task

Data mining is applicable in predictive modeling, descriptive modeling and exploratory data analysis, discovering pattern and rules [21]. Each of these applications will be explained in brief as follows.

2.2.1. Predictive Modeling

It is building a model for the dependent variable from one or more independent variables. The value of the dependent variable will be predicted from the known values of other variables. Time table prediction in train and other fields and function approximation are some of the applications areas. In predictive modeling one identifies patterns found in the data to predict future values. Classification and regression are two forms of data analysis that can be used to predict future data [13].

Classification methods create classes by examining already classified cases and inductively finding the pattern typical to each class. Regression uses the historical relationship between an independent and a dependent variable to predict the future values of the dependent variable. The difference between classification and regression is the type of output that is predicted; classification predicts class membership, whereas regression models continuous valued functions [13]. Traffic control use regression to predict future train and punctual rates.

Among different data mining tasks, Classification model is one of them. It is also known as supervised learning. This supervised method of data mining technique can predicts the continuous numeric and nominal attributes values. The attributes are further divided in two

namely; the input and output fields. Thus, the inputs are used to predict the outcome of the output fields [22].

Classification encompasses two levels: classifier construction and the usage of the classifier constructed. The former is concerned with the building of a classification model by describing a set of predetermined classes from a training set as a result of learning from that dataset. Each sample in the training set is assumed to belong to a predefined class, as determined by the class attribute label. The model is represented as classification rules, decision trees, or mathematical formula. The later involves the use of a classifier built to predict or classify unknown objects based on the patterns observed in the training set [23].

Classification maps data into predefined group or classes. Because the classes are determined before examining the data, classification is often considered as supervised learning. Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data which are already known to belong to the classes [24].

2.2.2. Descriptive Modeling

It describe all the data, it includes models for overall probability distribution of the data and groups and models describing the relationships between the variables. Some of the commonly used descriptive modeling techniques are clustering and data visualization.

Clustering is used to identify a finite set of categories or clusters to describe the data. It involves partitioning data according to natural classes present in it, assigning data points that are "more similar" to the same "cluster". In clustering, no data are tagged before being fed to a function. The goal of clustering is to sift/filter the data to produce a control of the input records. Different clustering functions will hence yield different sets of sorted data. It is up to the miner to determine what meaning, if any, to attach to the resulting clusters [25].

Visualization method is another powerful form of descriptive data mining [26]. It is a means for presenting data, both at the input and the output stages. Visualization techniques may help to discover relationships between features at the input stages, and explain the data mining results present them to the decision makers at the output stage. In the study, an extent experimental has been conduct using classification algorithms for designing a predictive model for train arrival time management.

2.3. Data Mining Process Models

To date many data mining and knowledge discovery process models have been developed. The most used in scientific research works, in industrial and academic projects are SEMMA(Sample, Explore, Modify, Model, and Assess), HYBRID,CRISP-DM(Cross-Industry Standard Process for Data Mining) and KDD(Knowledge Discovery in Databases).

2.3.1. SEMMA Process Models

SEMMA (Sample, Explore, Modify, Model, and Assess) was developed by the SAS Institute. The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. The SEMMA analysis cycle guides the analyst through the process of exploring the data using visual and statistical techniques, transforming data to uncover the most significant predictive variables, modeling the variables to predict outcomes, and assessing the model by testing it with new data [27]. It consists of five stages along with iterative experimentation cycle [27].

- **Sample:** During this stage, representative sample data are extracted from the portion of a large data, which is large enough to contain important data and yet small enough to manipulate quickly.
- **Explore:** This stage helps the user for better understanding of the data set through exploration of the data by searching for unanticipated trends and anomalies. It enhance to visualize the data for discovery process
- **Modify: The** aim of this stages is to undertake necessary adjustments to the data through creating, selecting, and transforming the variables for model construction purpose.
- **Model**: This stage involves construction models using appropriate modeling techniques that can explain patterns in the data.
- **Assess**: Finally, the usefulness and reliability of the models needs to be assessed and evaluated. It helps to estimate the performance of the models.

2.3.2. KDD Process Models

KDD (Knowledge Discovery in Databases) is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process [24]. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful

knowledge is derived from the data. Based on the aforementioned definition, they classify the stages of data mining in to five stages. These are stated below [24].

- **Selection**: This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
- **Preprocessing**: This stage consists on the target data cleaning and preprocessing in order to obtain consistent data.
- **Transformation**: This stage consists on the transformation of the data using dimensionality reduction or transformation methods.
- **Data Mining**: This stage consists on the searching for patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction)
- **Interpretation/Evaluation**: This stage consists on the interpretation and evaluation of the mined patterns.

Figure 1. Overview of the steps constituting the KDD process

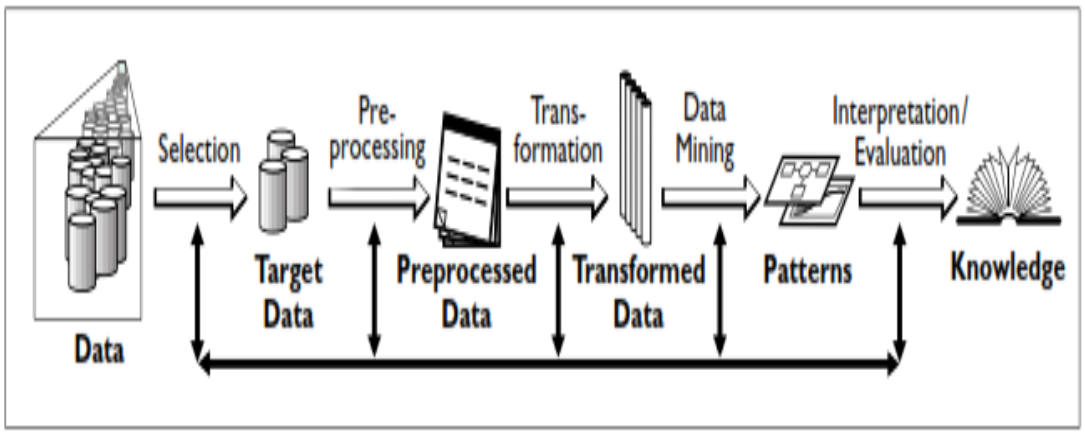


Figure 2.1 KDD process: "From Knowledge Discovery to Data Mining" [24]

2.3.3. CRISP-DM Process Models

CRISP-DM (Cross-Industry Standard Process for Data Mining) was developed in 1996 by analysts. It is applicable in typical data mining problems such as data description and summarization, segmentation, concept descriptions, classification, prediction, dependency analysis [28]. CRISP-DM is now being used in industry as the standard for a technology-neutral data mining process model [28]. The CRISP-DM process has six stages.

- **Business understanding:** This phase focuses on understanding the research objectives and requirements from a business perspective, then converting this Knowledge into a DM problem definition and a preliminary plan designed to Achieve the objectives.
- **Data understanding:** It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
- **Data preparation:** It covers all activities to construct the final dataset from the initial raw data.
- **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
- **Evaluation:** In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the DM results should be reached.
- **Deployment:** The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

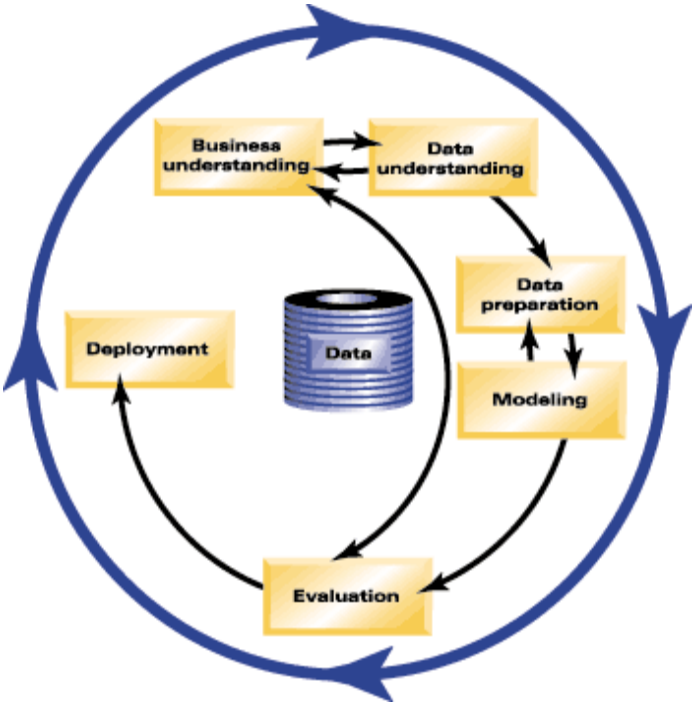


Figure 2.2 CRISP-DM process modeling (source: http://www.crisp-dm.org/).

2.3.5. Hybrid Models

According to Swiniarski and Kurgan [14], the hybrid models are enhanced the knowledge discovery process by combining the academic and industrial models in data mining projects. The development hybrid models was adopted from the CRISP-DM model as its can be used for academic research. Thus, these models are research-oriented, which present data mining step than the modeling step [14]. The six steps of hybrid models allow a number of feedback mechanisms. Moreover, the knowledge discovered in the final step for a specific domain may be applied in other domains. The following descriptions present the six steps of hybrid models.

- **Understanding of the problem domain:** The initial step involves task such as the problem definition and project goal determination, identification of key people and grasping the current solution to the problem through close consultation with the domain experts. Then the research goals are transformed to DM goal and the preliminary selection of DM tools to be used in the study is conducted.
- **Understanding of the data:** This step involves tasks such as the collection of data and choosing the size and format of the datasets. Furthermore to the quality of the data are assessed by checking the completeness, redundancy, missing values, plausibility of attribute values, etc. lastly, the usefulness of the data are verified with respect to the DM goals.
- **Preparation of the data:** In this step, the data going to be used are prepared to apply the DM methods. It consist of tasks such as sampling, testing the correlation and significance of the data, cleaning the data, checking the completeness of the tuples, handling noisy and missing values. Then, the dimensionality of the data is reduced by feature selection and extraction algorithms. This step also comprises, the derivation new attributes, summarization of the data. Finally, the datasets that meet the input requirements of DM tools stated in the first step are selected for modeling purpose.
- **Data mining:** This is another key step in the knowledge discovery process. Although it is the DM tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned DM tools and selection of the new ones. DM tools include many types of algorithms, such as neural networks, clustering, preprocessing techniques, Bayesian methods, machine learning, etc.

- **Evaluation of the discovered knowledge:** In this step the results of DM models are evaluated whether the discovered knowledge is novel and interesting and the results of the models are interpreted with respect to domain experts 'knowledge. In addition, the approved models are taken and the whole process is revised to pinpoint an alternative solution, in order to improve the results achieved. Finally, the errors arisen in the process are listed and arranged.
- **Use of the discovered knowledge:** The final step comprises planning the regarding the usage of the discovered knowledge. The knowledge discovered in the current domain may be applied in other domains. Also, a plan is created concerning the implementation of the knowledge discovered and the documentation of the whole project. Lastly, the deployment of the model takes place.

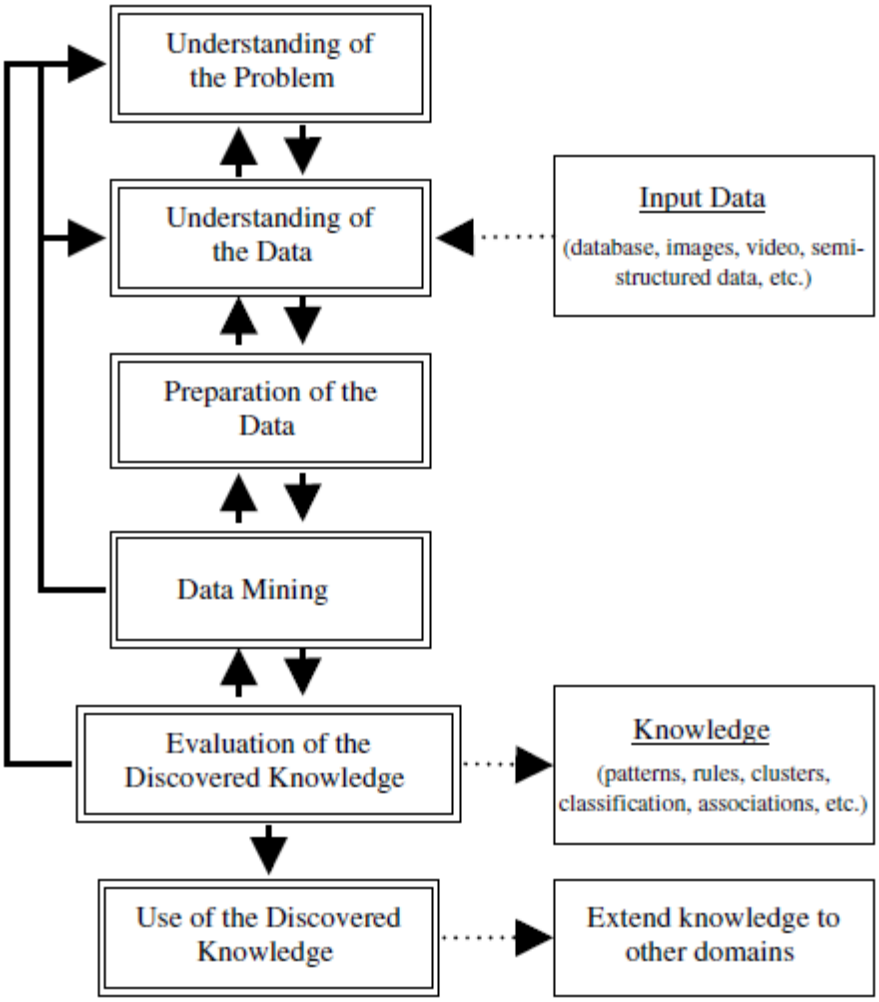


Figure 2.3 Hybrid model of Cios six-step methodology [29]

2.5. Classification algorithms

The classification task can be seen as a supervised technique where each instance belongs to a class, which is indicated by the value of a special goal attribute or simply the class attribute [30]. The goal attribute can take on categorical values, each of them corresponding to a class. Each example consists of two parts, namely a set of predictor attribute values and a goal attribute value. The former is used to predict the value of the latter. The predictor attributes should be relevant for identifying the class of an instance. In the classification task the set of examples being mined is divided into two mutually exclusive and exhaustive sets, called the training set and the test set [30].

The classification model is built from the training set, and then the model is evaluated on the test set. During training, the classification algorithm has access to the values of both predictor

attributes and the other attribute for all examples of the training set, and it uses that information to build a classification model. This model represents classification knowledge essentially, a relationship between predictor attribute values and classes that allows the prediction of the class of an example given its predictor attribute values. For testing, the test set the class values of the examples is not shown. In the testing phase, only after a prediction is made is the algorithm allowed to see the actual class of the just-classified example. One of the major goals of a classification algorithm is to maximize the predictive accuracy obtained by the classification model when classifying examples in the test set unseen during training [31].

There are different classification algorithms that are used for constructing a predictive model. Common classification algorithm includes Decision Tree, rule induction, K-Nearest Neighbor, Support Vector Machines, Naive Bayesian Classification and Neural Networks [31]. In this study decision tree and rule induction algorithms are used.

2.5.1. Decision Tree Classification

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items [32].

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree (see figure 2.4 below), meaning it is a directed tree with a node called “root” that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute’s value. In the case of numeric attributes, the condition refers to a range.

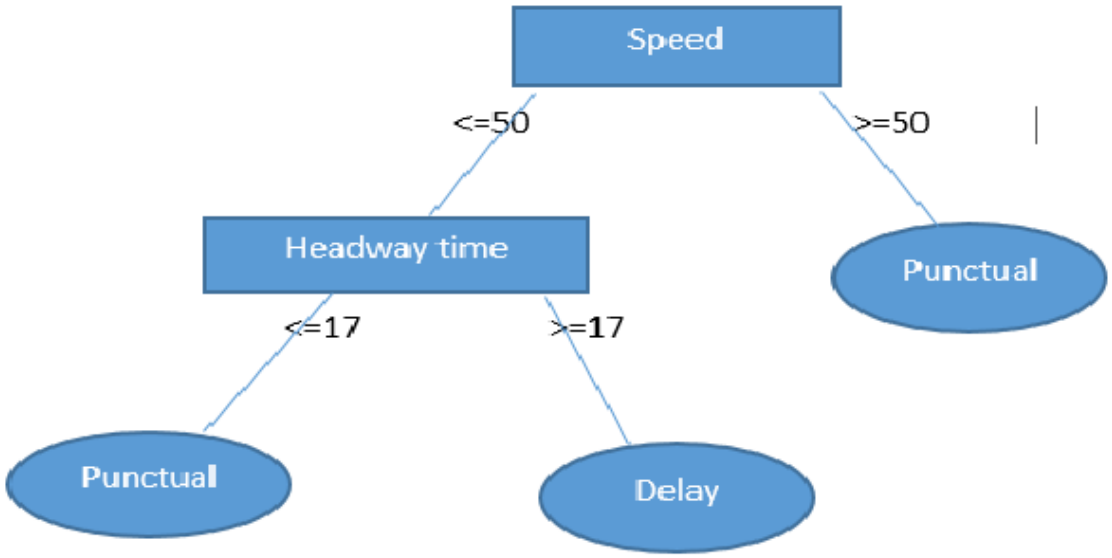


Figure 2.4 Simple decision tree train arrival time management

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path.

The many benefits in data mining that decision trees offer include the following [32]:

➢ Decision trees require very little data preparation whereas other techniques often require data normalization, the creation of dummy variables and removal of blank values.

- Uses a white box model i.e. the explanation for the condition can be explained easily by Boolean logic because there are mostly two outputs. For example, yes or no.
- Self-explanatory and easy to follow when compacted
- Able to handle a variety of input data: nominal, numeric and textual
- Able to process datasets that may have errors or missing values
- High predictive performance for a relatively small computational effort
- Useful for various tasks, such as classification, regression, clustering and feature selection

Some of the weaknesses of DT are [32]:

- Some DT can only deal with binary valued target classes, others are able to assign records to an arbitrary number of classes, but errors are prone when the number of training examples per class gets small. This can happen rather quickly in a tree with many levels and many branches per node.
- The process of growing a DT is computationally expensive. At each node, each candidate splitting field is examined before its best split can be found.
- Decision tree are less appropriate for estimation tasks where the goal is to predict the value of continuous such as income, blood pressure, or interest rate.
- Decision tree are also problematic for time-series data values a lot of effort is put into presenting the data in such a way that tends and sequential patterns are made visible

2.5.1.1. Decision tree Basic Principle (Hunt's method)

All DT induction algorithms follow the basic principle, known as CLS (Concept Learning system), given by Hunt. A CLS tries to mimic the human process of learning a concept, starting with examples from two classes and then inducing a rule to distinguish the two classes based on other attributes. Let the training dataset be T with class-labels {C1, C2... Ci}. the decision tree is built by repeatedly partitioning the training data using some splitting criterion till all the records in a partition belong to the same class. The steps to be followed are [28]:

I. If T contains no cases (T is trivial), the decision tree for T is a leaf, but the class to be associated with the leaf must be determined from information other than T.

II. If T contains cases all belonging to a single class Cj (homogeneous), corresponding tree is a leaf identifying class Cj.

III. If T is not homogeneous, a test is chosen, based on a single attribute, that has one or more mutually exclusive outcomes {O1, O2..., On}. T is partitioned into subsets T1, T2, T3... Tn, where Ti contains all those cases in T that have the outcome Oi of the chosen test.

The decision tree for T consists of a decision node identifying the test, and one branch for each possible outcome. The same tree building method is applied recursively to each subset of training cases.

2.5.1.2. Measures of Diversity in Decision tree

The diversity index is a well-developed topic with different names corresponding the various fields. To statistical biologist, it is Simpson diversity index. To cryptographers, it is one minus the repeat rate. To econometricians, it is the Gini index that is also used by the developers of the Classification and Regression Trees (CART) algorithm. A high index of diversity indicates that the set contains an even distribution of classes whereas a low index means that members of a single class predominate [33]. The best splitter is the one that decreases the diversity of the record sets by the greatest amount. The three common diversity functions are discussed here. Let there be a dataset S (training data) of C outcomes. Let P(I) denotes the proportion of S belong into a class I where I varies from 1 to C for the classification problem with C classes.

Simple Diversity index = Min(p(I))……(2.1)

Entropy provides an information theoretic approach to measure the goodness of a split. It measures the amount of information in an attribute.

$$\textbf{Entropy(S)} = \sum_{I=0}^{C}(-\mathbf{p(I)log_2 P(I)}) \quad \textbf{……..(2.2)}$$

Gain(S, A), the information gain of the example set S on an attribute A, defined as

$$\textbf{Gain(S, A) = Entropy(S)} - \sum\left(\left(\frac{|\mathbf{SV}|}{|\mathbf{S}|}\right) * \boldsymbol{Entropy}(\mathbf{S_V})\right) \quad \textbf{…….(2.3)}$$

Where Σ is over each value V of all the possible values of the attribute A,

SV = subset of S for which attribute A has value V, $|S_V|$ = number of elements in S_V, and

|S| = number of elements in S.

The above notion of gain tends to favor the attributes that have a larger number of values. To compensate this, it is suggested using the gain ratio instead of gain, as formulated below.

$$\textbf{Gain Ratio(S, A)} = \frac{\boldsymbol{Gain(S,A)}}{\boldsymbol{SplitInfo(S,A)}} \quad \textbf{……..(2.4)}$$

Where Split Info(S, A) is the information due to the split of S on the basis of the value of the categorical attribute A. Thus Split Info(S, A) is entropy due to the partition of S induced by the value of the attribute A.

Gini index measures the diversity of population using the formula

$$\textbf{Gini Index} = 1 - \sum(\mathbf{p(I)}^2) \quad \textbf{……(2.5)}$$

Where P(I) is the proportion of S belonging to class I and Σ is over C.

A number of different algorithms may be used for building decision trees including Chi-squared Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5, J48.

2.5.1.3. J48 decision tree algorithm

Decision tree models are constructed in a top-down recursive divide-and-conquer manner. J48 decision tree algorithms have adopted this approach. The training set is recursively partitioned into smaller subsets as the tree is being built [34].

According to Rokach [32]., J48 decision tree algorithm is a predictive machine learning model that decides the target value of a new sample based on various attribute values of the available data. Having the capability of generating simple rules and removing irrelevant attributes, the J48 decision tree can serve as a model for classification.

According to Hemalatha [35], J48 decision tree algorithm performs the following sequence of steps to accomplish its classification task.

J48 Decision Tree Classifier uses two phases [35]: tree construction and tree pruning.

Tree construction starts with the whole data set at the root. It then checks the attribute of the data set and partition them based on the following cases

Step 1: - If the attribute value is clear and has a target value, then it terminates the branch and assigns the value as Target value (classification)

Step 2: - If the attribute gives the highest information, then continue till we get a clear decision or run out of attributes.

Step 3: - If we run out of attributes or we are presented with ambiguous result, then assign the present branch as target value.

Step 4: - Ignore missing values.

The second phase Tree Pruning identifies and removes branches that reflect noise and outliers to reduce classification errors.

J48 algorithm has various advantages. Some of these advantages are the following [35]:

- Gains of balanced flexibility and accuracy
- Capability of limiting number of possible decision points
- Higher accuracy

2.5.2. Rule Induction system

According to Witten [36], the rule-based induction method is one of the most important machine learning techniques as it can express the regularities regarding rules that are frequently hidden in the data. It is the most fundamental tool in the data mining process. Generally, rules are expressions of the form: If (condition), then conclusion.

If (characteristic 1 is equal to value 1) and (characteristic 2 is equal to value 2) and (characteristic n is equal to value n), then (decision will be equal to the result).

Witten [36] stated that some rule induction systems provoke more complex rules, in which the characteristic values are expressed by the contradiction of some other values or by a value of the overall subset of the characteristic domain. It was further explained that the data by which the rules are provoked are generally presented in a form similar to a table that shows different cases (rows) against the variables (characteristics and decisions).

Rajput [37] said that the rule induction belongs to supervised learning, and all of it cases are pre-classified by experts. In simple words the decision values are assigned by the experts in this process. Anil further elaborated that the characteristics represent independent values, while the decisions represent the dependent variables. The covering method represents classification of knowledge in the form of a set of rules which represent or give a description of each class.

This procedure makes use of the following search process to produce the rules for each class in the training set T: While the stopping criterion is not satisfied:

- Form a new rule to cover examples belonging to a target class employing the Rule Forming Process;
- Add this rule to the Rule Set;
- Remove all examples from T which are covered by this new rule.
- Stop the procedure when there are no more classes to classify.

2.5.2.1. JRIP rule classifier

According to Rajput [37], JRIP is one of the basic and most popular rule induction algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRIP (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class and finding a set of rules that cover all the members of

that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is based in association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms [37].

2.5.3. Naïve Bayes Classifier

A Naive Bayesian classifier is a simple method particularly suited when the dimensionality of the inputs is high and classification based on the theory of probability i.e. Bayesian theorem (from Bayesian statistics). It is called naïve because it simplifies problems relying on two important assumptions: it assumes that the prognostic attributes are conditionally independent with familiar classification, and it supposes that there are no hidden attributes that could affect the process of prediction. This classifier represents the promising approach to the probabilistic discovery of knowledge, and it provides a very efficient algorithm for data classification [38].

Bayesian network model is constructed by explicitly determining all the direct dependencies between the features of the problem domain. And also there has been much interest in learning Bayesian networks from data. In this research study, the researcher made experiments based upon the Bayes approach defines the classification problem in terms of probabilities that formulated by the underneath proof. More specifically, the three main concepts required are conditional probability, Bayes Theorem, and the Bayes decision rule. The conditional probability P (A/B), which is used to define independent events [39] is defined by P (A/B) =p (AUB)/p(B').

Where P (A/B) is the probability that event A happens, given that B is observed. Similarly,

P (B/A) =p(AUB)/p(A') Where P (B/A) is the probability that event B happens, given that A is observed. It then follows (by substitution) that $(A\ n\ B) = P(A)P(B/A)$

Although, the premise of Bayes Theorem starts with an initial degree of belief that an event was Occur, and then with new information this degree of belief can be "updated" [39]. These two degrees are represented, respectively, by the prior probability P(A/B) and the posterior probability P (B/A), which are related by P (A/B) =p (A) p (B/A)/p (B).

Generally speaking, the Bayesian methodology for classification as well as prediction of the pattern to designing train arrival time management, particularly for the AALRT for OCC data warehouse follows these five steps [40]:

- Collect data, and estimate parameters such as mean and covariance for each class.

- Choose a set of features.
- Choose a model and derive a decision rule with these parameters.
- Train the classifier and apply the decision rule by using a discriminant function (a way to Represent a pattern classifier), and apply it to a test data set to classify each sample.
- Evaluate the decision rule. Measure the accuracy /error rate in order to improve the choice of features and the overall design of the classifier.

2.6. Weka Data Mining Tools

There are varieties of tool available for data mining like WEKA, Shogun, Orange, Scikit-learn, R and Rapid Miner and those tools comprise on the basis of operating system and file formats supported, general features and language bindings. This is useful for various users to select the tool best suitable for their application. All the tools do not support all the data mining operations. However WEKA and Shogun supports all the three so in this research the tools selected WEKA stands for Waikato Environment open source tools available for data mining from some of tools work is for Knowledge Analysis. It is developed in Java programming language. It contains tools for data preprocessing, classification, clustering, association rules and visualization. Data file can be used in any format like ARFF (attribute relation file format), CSV (comma separated values), C4.5 and binary and can be read form a URL or from SQL database as well by using JDBC. One additional feature is that data sources, classifiers etc. are called as beans and these can be connected graphically consulting studies conducted in the area of DM tools comparison.

No.	Tool Name	Release Date	License	Language	Operating System	Type
1.	RAPID MINER	2006	AGPL Proprietary	Language Independent	Cross platform	Statistical analysis, data mining, predictive analytics.
2.	ORANGE	2009	GNU General Public License	Python C++, C	Cross Platform	Machine learning, Data mining, Data visualization
3.	KNIME	2004	GNU General Public License	Java	Linux, OS X, Windows	Enterprise Reporting, Business Intelligence, Data mining
4.	WEKA	1993	GNU General Public License	Java and C#	Cross Platform	Machine Learning
5.	KEEL	2004	GNU GPL v3	Java	Cross Platform	Machine Learning
6.	R	1997	GNU General Public License	C, Fortran and R	Cross Platform	Statistical Computing

Table 2.1 Data Mining Tools Summary

2.6.1. WEKA (Waikato Environment for Knowledge Analysis)

Weka is a collection of machine learning algorithms for data mining tasks. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.

The Explorer for exploratory data analysis to support preprocessing, attribute selection, learning, visualization, the Experimenter that provides experimental environment for testing and evaluating machine learning algorithms, and the Knowledge Flow for new process model inspired interface for visual design of KDD process. A simple Command-line explorer which is a simple interface for typing commands is also provided by WEKA.

Weka loads data file in formats of ARFF, CSV, and C4.5, binary. Though it is open source, Free, Extensible, can be integrated into other.

2.6.2. WEKA Interfaces

WEKA (Waikato Environment for Knowledge Analysis) is a machine learning and data mining software tool written in Java and distributed under the GNU Public License. The goal of the WEKA project is to build a state-of-the-art facility for developing machine learning techniques

and to apply them to real-world data mining problems. It contains several standard data mining techniques, including data preprocessing, classification, regression, clustering, and association. Although most users of WEKA are researchers and industrial scientists, it is also widely used for academic purposes [41].

Figure 2.5 WEKA interface

WEKA version 3.8 has five interfaces, which start from the main GUI Chooser window, Whereas the Explorer and Knowledge Flow are tailored to beginning users, the experimenter, workbench and simple CLI target more advanced users. The buttons can be used to start the following applications:

• **Explorer:** An environment for exploring data with WEKA (the rest of this documentation deals with this application in more detail).

• **Experimenter:** An environment for performing experiments and conducting statistical tests between learning schemes.

• **Knowledge Flow:** This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.

Workbench: provides as a working area for creating a model same as experimenter.

• **Simple CLI:** Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

The Explorer is possibly the first interface that new users will run simulations in. It allows for data visualization and preprocessing. In this study we use Explorer environment to conduct the experiment.

2.7. Application of Data Mining in Train arrival time management Sector

The train arrival time industry generates and stores a tremendous amount of data. These data include rescheduled data, which describes the calls that traverse the train arrival time management, network data, which describes the state of the hardware and software components in the network, and time data, which describes the train arrival time management. The amount of data is so great that manual analysis of the data is difficult, if not impossible.

Globally, the development of train arrival time industry is one of the important indicators of social and economic development of a given country. In addition to this, the development of transport services plays a vital role in the overall development of all sectors related to social, political and economic affairs. This sector is very dynamic in its nature of innovation and dissemination [6].

In train arrival time management sector, data mining is applied for various purposes.

2.7.1. Customer relationship management (CRM)

CRM is an enterprise approach to understand and influence customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability [42]. According to Parvatiyar and Sheth [42] CRM can also be defined as a comprehensive strategy and process of acquiring, retaining, and partnering with selective customers for the purpose of creating superior value for both the company and the customer. To achieve a better efficiencies and effectiveness in delivering customer value, CRM involves the integration of marketing, sales, customer service, and the supply-chain functions of the organization.

Generally, CRM involves acquisition of customers, customer retention and customer segmentation.

The acquisition of new customers is an important business problem related to ratemaking.

Traditional approaches involve attempts to increase the customer base by simply expanding the efforts of the sales department. In contrast to traditional sales approach, DM strategies enable analysts to define the marketing focus. Analysts in the railway industry can utilize advanced DM

techniques that combine segmentations to group the high lifetime-value customers and produce predictive models to identify those in this group who are likely to respond to marketing campaign. Using a DM technique called association analysis, insurance firms can more accurately select which policies and services to offer to which customers.

2.7.2. Time management

Improving the performance of railway infrastructure and train services is the core business of railway infrastructure managers and railway undertakings. Train delays decrease capacity, punctuality, reliability and safety, and should be prevented as much as possible [43].

Time management predictive analytics uses data mining and statistical analysis to provide actionable predictions and help drive the decision-making process. Traditionally, time management deals with time recording and time reporting. Reporting is helpful in answering questions about the past and to a certain extent in explaining current events, but doesn't provide insight into the future, nor does it provide actionable recommendations for the managers. Reports help managers answer questions such as what happened, why did it happen, and what was the problem. Predictive analytics goes beyond this and helps the managers understand what is happening in real-time and what will happen next.

Time management predictive analytics provides actionable predictions based on trends, patterns, relations and correlations in time records. It helps the decision-making process by making the intelligent prediction available to the managers in a user friendly format. Time management predictive analytics applies a mathematical modeling and statistical analysis approach to the time records to develop knowledge that can be used to predict future trends of employee behavior.

Not every manager is an experienced data modeler, so predictive analytics needs to auto-suggest a best analysis model and key performance indicators. The system needs to provide the analysis in simple English with actionable recommendations for the manager.

No one works in isolation, so the predictive analytics needs to be able to connect to other organizational functionality like task management, expense management, and process management. By assessing the degree of change required, management can calculate the impact on cost, time and resources. Only through the analysis of the time records can you begin to understand employee behavior, identify working trends and discover where work can be made more efficient. Predicting employee behavior lays a foundation for improving business performance.

Time management predictive analytics improves strategic business planning by eliminating the reliance on averages or guesswork. Embedding the ability to intelligently predict the uncertain future and the ability to measure the impact of this uncertainty in strategic planning goes a long way in helping companies achieve their business objectives.

The success of predictive analytics relies on the number of time records being fed into the system. The more employees use it - the better the analysis. The more information the employees provide to the system - the better the analysis.

Predictive analytics helps identify possible reasons for employees not achieving the expected results: incorrect time estimates, working on low importance tasks, too many urgent tasks causing distraction, fatigue and rework. Predictive analytics uses a wide range of information to analyze factors such as morning/evening effort, repetitive tasks, employee breaks, rework, and churn. This enables managers to identify trends and provides an opportunity to correct the employee's schedule.

2.8. Related Works

Modeling urban behavior by mining geo-tagged data is a popular topic for research [2, 3]. A major difference between timetable planning and rescheduling is that in the latter case models and algorithms must be compliant with the actual state of the network and with the current position of each train. Moreover, rescheduling algorithms must deliver good solutions within short computation time. We classify the problems addressed in the literature in distributed rescheduling, centralized rescheduling and coordinated rescheduling. The complexity of the railway network considered ranges from a simple junction to a set of dispatching areas.

Among distributed approaches based on negotiations at the level of junctions, Vernazza and Zunino [27] propose an approach to solve train conflicts locally by enabling a negotiation between the trains and the local infrastructure administrator. The control problem is modeled in terms of resource allocation tasks and priority rules are adopted for each local controller. The system simulates a realistic network and train conflicts at simple railway junctions are solved by local decision rules depending on the traffic intensity. Parodi et al.[20] study an advanced resource-allocation task for train scheduling based on local decision rules, and study how to detect and solve in advance possible deadlocks for different network configurations.

In Iyer and Gosh [14] every train is equipped with an onboard processor that claims the setup of train routes, dynamically and progressively, through explicit processor to processor communication primitives. Each train negotiates to get access to block sections while minimizing its total travel time. The decision process of each station is executed by a dedicated processor that, in addition, maintains absolute control over a given set of track segments and participates in the negotiation with the trains. Their experiments, carried on an artificial test case with up to 12 stations, 17 track segments and 48 trains, show that the computation time increases rapidly. Other distributed approaches use Petri Nets to model the train arrival time flow. Fay [12] describes an expert system and suggests a fuzzy rule-base, Fuzzy Petri Net, for train traffic control during disturbances. Experiments are performed on fictional data with some trains and one station. Zhu [29] introduces a simulation model based on stochastic Petrinets in order to assess the impact of incidents on the quality of operations. The latter approach focuses on how to determine train traffic delays caused by primary stochastic disturbances, especially technical failures. Cheng and Yang [1] include further factors, such as train connections and passenger trip types, in a similar fuzzy Petri Net approach for managing the dispatching process. The dispatching local decision rules are collected via interviews to experts. However, Petri Net approaches still seem to be far from producing near-optimal solutions to practical problem instances.

Among the centralized approaches for managing a dispatching area, ahin[23] formulates a meet and pass problem as a job shop scheduling problem. Conflicts between up to 20 trains are solved in the order they appear for 19 meet points. An algorithm based on look-ahead measures detects potential delays and takes ordering decisions at merging or crossing points in order to minimize the average delays.

Wegele et al. [28] use genetic algorithms to reschedule trains with the objective of minimizing passenger annoyance, e.g. delays, change of platform stops and missed connections. The adopted dispatching strategies are dwell time modifications, adaptation of train speeds on corridors and local rerouting in-side stations. Examples of application on a

Large part of the German railway network are reported for a single delayed train.

Rodriguez [22] focuses on a real-time train conflict resolution problem and proposes a train routing and scheduling system based on constraint programming. The experiments show that a truncated branch and bound algorithm can find satisfactory solutions within a short time for a

railway junction of a few kilometers traversed by up to 24 trains. Chou et al. [2] propose a distributed control system and study a number of railway areas that are mutually influenced. A novel time-shift coordination strategy between neighboring traffic control areas is proposed for collaborative train rescheduling. Distributed control techniques in neighboring regions are applied in a fictitious network and evaluated in terms of delay cost. Recently, Chou et al. [3] demonstrate a significantly lower delay cost compared to a first come first served rule for a realistic railway junction with around 12 trains per peak hour traveling in a single direction.

As a general remark about the existing literature, we observe that most of the existing approaches lack of a thorough computational assessment and limit the analysis to simple networks or simple perturbation patterns. In fact, the analyzed delay patterns are often quite specific, e.g. only one train is delayed or the problems limited to a single junction or to a straight line. Moreover, the models used in the literature for the assessment are often simplified and do not capture entirely the consequences of delays and other disturbances.

In order to solve practical problems, detailed models are necessary that capture the real problem complexity. For instance, when dealing with large networks and dense traffic, possibly with disruption, the risk of deadlock is relevant and should not be ignored by real-time models. The alternative graph model is among the few models that incorporate the level of detail necessary to generate deadlock-free schedules within an optimization framework.

In this research an attempt made to explore the limits for practical applicability and coordination algorithms based on the alternative graph to support railway operators in the management of disturbed traffic situations on an increasing number of trains and levels of disturbance. *To the end* this study aims to design a predictive model for train arrival time management of Addis Abeba railway transit by using data mining techniques over the dispatching areas, with various time horizons of traffic predictions and different network decompositions. The disturbances include multiple delayed trains and a serious and permanent disruption in the network, which requires the rerouting of several trains and the management of complex traffic situations with the risk of deadlocks. Several approaches to traffic state prediction can be found in the current practice or academic literature. Macroscopic models [5] focus on predicting only the event times in stations (departures, arrivals and through rides).The work of [6] uses a heuristic algorithm to reschedule trains. It compares the planned arrival times of trains with the current expected arrival times for waiting detection. Then, it uses a discrete event simulator to evaluate the alternative

choices to solve the first time waiting found. Dispatching rules solve conflicts by means of a local decision criterion. Two of the most common rules are the first-in-first-out (FIFO) rule and the first-out-first in (FOFI) rule. The research uses classification data mining algorithms to extract hidden patterns from reschedule' data.

CHAPTER THREE

Understanding of the Problem and Data Preparation

3.1. Rail Transport services

Rail transport is also known as train transport. It is a means of transport, on vehicles which run on tracks (rails or railroads). It is one of the most important, commonly used and very cost effective modes of commuting and goods carriage over long, as well as, short distances. Since this system runs on metal rails and wheels, it has an inherent benefit of lesser frictional resistance which helps attach more loads in terms of wagons or carriages. This system is known as a train. Usually, trains are powered by an engine locomotive running on electricity or on diesel. Complex signaling systems are utilized if there are multiple route networks. Rail transport is also one of the fastest modes of land transport [44].

Rail transport has emerged as one of the most dependable modes of transport in terms of safety. Trains are fast and the least affected by usual weather turbulences like rain or fog, compared to other transport mechanisms. Rail transport is better organized than any other medium of transport. It has fixed routes and schedules. Its services are more certain, uniform and regular compared to other modes of transport. Now it has evolved into a modern, complex and sophisticated system used both in urban and cross-country (and continent) networks over long distances [45]. Rail transport is an enabler of economic progress, used to mobilize goods as well as people. Adaptations include passenger railways, underground (or over ground) urban metro railways and goods carriages. Rail transport has some constraints and limitations also. One of the biggest constraints of rail transport is heavy cost. Trains need high capital to build and maintain and the cost is magnified when a whole rail network is to be built. The cost of construction, maintenance and overhead expenses are very high compared to other modes of transport. Also, rail transport cannot provide door-to-door service as it is tied to a particular track. Intermediate loading or unloading involves greater cost, more wear and tear and, also wastage of time [46].

3.1.1. Train arrival time Management system

Train control systems pose high demands on positioning with respect to availability, reliability and integrity. These requirements can only be fulfilled by means of integrated positioning systems, which combine GNSS(Global Navigation Satellite System) with other sensors.

The use of GNSS in railway systems presents many advantages, in particular the monitoring of train's exact location, logistic information management, enhanced train signaling (which improves safety, but also enables for example. reduced distances between trains and therefore increased train frequencies), and the possibility to map the transport infrastructure [47].

Thus, while the number of applications based on GNSS is considerably behind the number of those used in other domains, such as road transport, incorporating GNSS receivers into modern signaling, train control and other railway systems has become common [47].

3.1.2. Ways for Train arrival time Management system

Train control, signaling, passenger traffic or transportation of dangerous goods are safety-critical applications, which show very demanding requirements in terms of availability, continuity and integrity. In order to fulfill these high performance demands complementary positioning sensors such as accelerometers or digital track maps and alternative communication components have to be grouped around the receiver/communication core and revised to time schedules.

The Global Navigation Satellite System (GNSS) receivers for Traffic Management and Signaling are considered safety critical applications.

Nowadays, modern railways in many countries are adopting Positive Train Control (PTC) systems to prevent collisions, derailments, work zone incursions, and passage through switches in the wrong position.

The PTC systems are integrated command, control, communications, and information systems for controlling train movements with safety, security, precision, and efficiency. PTC systems will improve railroad safety by significantly reducing the probability of collisions between trains, casualties to roadway workers and damage to their equipment, and over speed accidents.

A PTC system can automatically vary train speeds, re-route traffic, rescheduling and safely direct maintenance crews onto and off tracks. In addition to enhancing safety, PTC increases track capacity by maintaining a constantly updated operating plan that optimizes rail use and flow and Give dispatchers and passengers more accurate information on train arrivals times [48]. In the next Section we provide details about the current railway system in Ethiopia specifically the Addis Ababa Light Railway Transport (LRT), and the techniques to be used in order to control the railway system.

3.1.3. Factors Affecting Train arrival time Management System

There are different factors that should be taken in to amount for selecting train arrival time management.

3.1.3.1. Train

A train is a form of transport consisting of a series of connected vehicles that generally runs along a rail track to transport cargo or passengers [49]. Train tracks usually consist of two running rails, sometimes supplemented by additional rails such as electric conducting rails and rack rails. Monorails and maglev guide ways are also used occasionally [49].

A passenger train includes passenger-carrying vehicles and can often be very long and fast. One notable and growing long-distance train category is high-speed rail. In order to achieve much faster operation at speeds of over 500 km/h (310 mph), innovative maglev technology has been the subject of research for many years. The term "light rail" is sometimes used to refer to a modern tram system, but it may also mean an intermediate form between a tram and a train, similar to a heavy rail rapid transit system. In most countries, the distinction between a tramway and a railway is precise and defined in law.

A freight train (or goods train) uses freight cars (or wagons/trucks) to transport goods or materials (cargo). It is possible to carry passengers and freight in the same train using a *mixed consists*.

- **Train Speed**

Speed is one of the key parameters for designing a timetable. Because the train speed has a lot of effects in most of the other parameters to be considered. For instance, the train speed affects the headway consequently; this headway has a huge effect on the train capacity. That is if the train

runs slowly it decrease the capacity because it needs more travel time and hence they need more headway. In contrast, if the train speed increase, the capacity of the line increases because it needs less headway time to the train which follows the train ahead. Also, the braking distance mostly depends on the train speed. In the AA-LRT the speed designed ranges from 20-70km/h.

- **Train model**

Chinese train manufacturer, **CNR Corporation** (CNR) has signed contracts with **Ethiopia** to provide a fllet of 41 modern **tramcars**, (LRVs) for the 37.4km light rail network which is currently under construction in the Ethiopian capital Addis Ababa.

The tramcars will be customized for use in Ethiopia's capital of **Addis Ababa**, where the altitude is 2,400 meters , according to CNR's statement.

According to CNR, the tramcars are the world's most sunlight-resistant and will use special components in the glass, rubber, paint and cable.

The three-section 70% low-floor vehicles will have a maximum speed of 70km/h and the first units are due to be delivered to Ethiopia at the end of 2014.

China is financing 60% of the $US 400m light rail project, with the remainder coming from the Ethiopian government. The network will have three lines: Defence Forces Hospital – Ayat Village (17.3km), Meskel Square – Kality (16.2km) and Lideta – Menilik Square (3.9km).

- **Manufactured date**

CRRC Changchun Railway Vehicles Co., Ltd is a Chinese rolling stock manufacturer and a division of the CRRC. While the CRV emerged in 2002, the company's roots date back to the establishment of the Changchun Car Company in 1954. The company became a division of CNR Corporation before its merger with CSR to form the present CRRC. It has produced a variety of rolling stock for customers in China and abroad, including locomotives, passenger cars, multiple units, rapid transit and light rail vehicles. It has established technology transfer partnerships with several foreign railcar manufacturers, including Bombardier Transportation, Alstom, and Mobility. To serves LRV for as AALRT at 2012 to 2042.

- **Train types**

To guarantee a comfortable and fast train ride, high-speed trains are the top options for customer. Those trains named after single, couple or multi trains.

- **Rolling stock**

RSR has as input a timetable and a rolling stock circulation where the allocation of the rolling stock among the stations at the start or at the end of a certain planning period does not match with the allocation before or after that planning period. The problem is then to modify the input rolling stock circulation in such a way that the number of remaining off-balances is minimal. If all off-balances have been solved, then they obtained rolling stock circulation can be implemented in practice [11].

Delay at the origin. It is the difference between the actual train's departure time and the scheduled train's departure time.

• **Incidence with another passenger train.** This case happens when trains running in opposing directions pass each other at places where loops or sidings are available. It is the essentials of train waiting time for a line clearance.

• **Unscheduled waiting time at overtaking points**. It is the train waiting time for the arrival and passing of another train with common path according to its priority.

• **The other trains' engine breakdowns.** The other trains' engine damages, which have some negative effects on travel time of this train.

Accordingly, describe the train involves the following attributes with its description presented the table 3.1

No	Attributes Name	Description
1	Train Speed	Standard design speed and operational speed
2	Train Model	Old Model or new model
3	Train Types	Comfortable and fast train ride and slowest train
4	Rolling Stock	Number of breakdowns in rolling stock
5	Manufacture Date	Manufacture to the train and last service date

Table 3.1 List of train attributes

3.1.3.2. Track

Track is the base upon which the railway runs. To give a train a good ride, the track

Alignment must be set to within a millimeter of the design. Track design and construction is part of a complex and multi-disciplinary engineering science involving earthworks, steelwork, timber and suspension systems - the infrastructure of the railway. Many different systems exist throughout the world and there are many variations in their performance and maintenance. This page looks at the basics of infrastructure and track design and construction with drawings, photos and examples from around the world. Some information was contributed by Dan McNaughton, SimonLowe and Mike Brotzman.

Track is the most obvious part of a railway route but there is a sub-structure supporting the track which is equally as important in ensuring a safe and comfortable ride for the train and its passengers or freight. The infrastructure shows the principal parts of an electrified, double-track line. The total width across the two-track alignment will be about 15 m (50 ft) for a modern formation. The "chess" shown each side of the alignment is the area available for a walkway or refuge for staff working on the track.

- **Track line**

In the study of the Addis Ababa light rail transit service, there are two lines which are East-west(EW) line and North-south(NS) line contain 39 stations in total, including 5 common station and 2 depots which are Kality and Ayat. The east-west line extends 16.998 kilometers, stretching from Ayat Village to Torhailoch, and passing through Megenagna, Meskel Square, Legehar and Mexico Square.

The north-south line, which is 16.689 kilometers in length, passes through Menelik II Square, Merkato, Lideta, Legehar, Meskel Square, Gotera and Kaliti. However, two lines have a common track of about 2.662 km. The common track is the elevated section which runs east to west across the southern edge of the CBD from Meskel Square to Mexico Square, and onwards to Lideta. Trains on the north south line are blue and white, whilst on the east west line they are green and white.

Section, the area between two adjacent stations, is different. In EW line the longest section is 1260m from EW2 to EW1, the shortest is 445m from EW17 to EW16. In NS line the longest section is 1971.66m from NS12 to NS11, the shortest is 445m from EW17 to EW16.

- **Ballast**

Ballast is provided to give support, load transfer and drainage to the track and thereby keep water away from the rails and sleepers. Ballast must support the weight of the track and the considerable cyclic loading of passing trains. Individual loads on rails can be as high as 50 tons (55 US or short tons) and around 80 short tons on a heavy haul freight line. Ballast is made up of stones of granite or a similar material and should be rough in shape to improve the locking of stones. In this way they will better resist movement. Ballast stones with smooth edges do not work so well. Ballast will be laid to a depth of 9 to 12 inches (up to 300 mm on a high speed track). Ballast weighs about 1,600 to 1,800 kg/cu/m. See also Ballasted vs Non-Ballasted Track below.

- **Track**

The usual track form consists of the two steel rails, secured on sleepers (or crossties, shortened to ties, in the US) so as to keep the rails at the correct distance apart (the gauge) and capable of supporting the weight of trains. There are various types of sleepers and methods of securing the rails to them. Sleepers are normally spaced at 650 mm (25 ins) to 760 mm (30 ins) intervals, depending on the particular railway's standard requirements.

- **Sleepers (Ties)**

Traditionally, sleepers (known as ties in the US) are wooden. They can be softwood or hardwood. Most in the UK are softwood, although London Underground uses a hardwood called Jarrah wood. Sleepers are normally impregnated with preservative and, under good conditions, will last up to 25 years. They are easy to cut and drill and used to be cheap and plentiful. Nowadays, they are becoming more expensive and other types of materials have appeared, notably concrete and steel.

Concrete is the most popular of the new types(left). Concrete sleepers are much heavier than wooden ones, so they resist movement better. They work well under most conditions but there are some railways which have found that they do not perform well under the loads of heavy haul freight trains. They offer less flexibility and are alleged to crack more easily under heavy loads with stiff ballast. They also have the disadvantage that they cannot be cut to size for turnouts and special track work. A concrete sleeper can weighs up to 320 kg (700 lbs) compared with a

wooden sleeper which weighs about 100 kg or 225 lbs. The spacing of concrete sleepers is about 25% greater than wooden sleepers.

- **Rail**

The standard form of rail used around the world is the "flat bottom" rail. It has a wide base or "foot" and narrower top or "head". The UK introduced a type of rail which was not used elsewhere - apart from a few UK designed railways.

- **Rail Welding**

Modern track work uses long welded rail lengths to provide a better ride, reduce wear, reduce damage to trains and eliminate the noise associated with rail joints. Rail welding is a complex art (or science) depending on how you feel about it. There are two main types of welding used for rails: Thermite welding and Flash Butt welding.

- **Gauge**

The standard track gauge - the distance between the two rails - is 4 ft. 8½ in or 1435 mm. but many other gauges, wider and narrower than this, are in use around the world. Gauge is often intentionally widened slightly on curved track. There is some additional information on Track Gauges at the Pacific Southwest Railroad Museum site.

- **Curves**

Curves in the track are almost a science on their own. Careful calculations are required to ensure that curves are designed and maintained properly and that train speeds are allowed to reach a reasonable level without causing too much lateral stress on the track or inducing a derailment. There are both vertical curves and horizontal curves. There is also a section of track on either side of a curve known as the transition, where the track is changing from straight to a curve or from a curve of one radius to one of another radius.

Minimum Curve Radius: 50m for mainlines, 30m or parking garage

Minimum vertical curve Radius: 1000m

- **Turnouts**

I have used the word "turnout" to describe the junctions in track work where lines diverge or converge so as to avoid "points" (UK) or "switches" (US), both of which terms can be confusing. In the railway "trade", turnouts are referred to as "switch and crossing work". The moving part of the turnout is the switch "blade" or "point", one for each route. The two blades are fixed to each other by a tie bar to ensure that when one is against its stock rail, the other is fully clear and will

provide room for the wheel flange to pass through cleanly. Either side of the crossing area, wing and check rails are provided to assist the guidance of the wheelsets through the crossing.

There are a number of standard layouts or types of turnouts as: left hand turn out,Y turn out, diamond crossing ,single slip and double slip

- **Rout station**

The platform of the AA-LRT stations has a length of 60m. It has a total of 22 stations in the east-west (EW) line and similarly 22 stations in the North-south (NS) line but, it has a total of 39 stations in both routes with five stations in common. The name of the stations is written in two ways, the first one is they use the local name of the city. The other and systematic name (i.e. easy to handle) uses the abbreviation letters EW following a number of the east-west direction and the letters NS following some number of the stations of north south directions. The naming of stations for the EW direction starts from EW1, EW2, EW3, ..., EW22, to mean'Ayat', 'Meri', 'CMC', …, 'Torhailoch', and for the NS line is NS6, NS7, NS8, …, NS27 to mean 'Kality','Abo Mazoria', 'Saries', …, 'Menelik II square'. The common lines are represented either as EW16, EW17,EW18, EW19 and EW20 or NS16, NS17, NS18, NS19 and NS20. These names can be used interchangeably but most of the time the name EW is used.

- **In Signaling**

Safety and signaling systems are an essential part of modern railways. Their main Purpose is to ensure safe train runs by preventing derailments and collisions between trains that share the same infrastructure elements, and accidents between trains and other vehicles and objects. This overview is focused on the fixed-block signaling system that gives the movement authority for all trains on open tracks and in interlocking areas. A comprehensive description of components and functions of safety and signaling systems is given by [50].

The safety principles required for route setting need to hold as long as the route is being used by a train or until it is cancelled by the controller. A route is released only after the train has cleared it. In order to increase the capacity of interlocking areas, especially in complex and busy stations, the modern interlocking systems employ the sectional-release route setting principle. Each section of the route becomes available for another route as soon as it is released by the last axle on the rear of the train. Route holding ensures that the occupied and non-traversed sections still stay locked in the route.

- **Power**

DC systems (especially third rail systems) are limited to relatively low voltages and this can limit the size and speed of trains and cannot use low-level platform and also limit the amount of air-conditioning that the trains can provide. This may be a factor favoring overhead wires and high voltage AC, even for urban usage. In practice, the top speed of trains on third-rail systems is limited to 100 mph (160 km/h) because above that speed reliable contact between the shoe and the rail cannot be maintained [51].

For describing 'track' select different attributes are identified table 3.2 shows list of attributes and description.

No	Attributes Name	Description
1	Ballast	Support the weight of the Track
2	Track	Keep the rails at the correct distance apart
3	Sleepers	Sleepers are normally impregnated with preservative and, under good conditions
4	Rail	Wide base and Narrower top
5	Curves	Both vertical curves and horizontal curves
6	Gauge	The distance between the two rails
7	Turnout	Describe the junctions in track work where lines diverge
8	Rail Welding	Reduce damage to trains and eliminate the noise associated with rail joints
9	Signal	How much station for the root 。
10	Route Station	Akaki to mnilik and hayat to torhayeloch
11	Power	It is the total of breakdowns in AALAT power facilities recorded.

Table 3.2 List of track attributes

3.1.3.3. Passenger

a passenger train is one which includes passenger-carrying vehicles and can often be very long and fast. Passenger trains travel between stations or depots where passengers board and disembark. In most cases, they operate on a fixed schedule and Due to the large passenger flow, the services to passengers getting on or off slowly. This result to Cancellation of passenger connections is one of the main sources of passenger dissatisfaction, especially for long distance

travels. Cancellation of some scheduled connection reduces delay propagation at the expenses of the delay of passengers affected by the missed connection.

- **Discount fee user passenger**

Children, student ticket users (below high school), elders, people with disability, Victorian and some metro also including military passengers

- **Normal Fare user passenger**

Passengers had different levels of satisfaction for the service attributes of the light rail transit. Because commuters satisfied with the staff behavior, affordability, and ticket payment system and also somewhat satisfied with reliability, comfort, safety and security, accessibility and availability. Passengers outside of discount fee user, one-way ride and repotted ride.

For describing 'passenger' select two attributes are identified table 3.3 shows list of attributes and description.

No	Attributes Name	Description
1	Discount Fee Passenger	Beneficial passengers
2	Fare Passenger	Passengers outside of discount fee user

Table 3.3 List of passenger attributes

3.1.3.4. Driver

Train drivers are in charge of, and responsible for, driving the locomotives, as well as the mechanical operation of the train, train speed and all train handling. They may also inspect trains, report defects and carry out adjustments, shunt rolling stock in marshalling yards and sidings along the line, and refuel diesel trains. In some organizations, they may make announcements and work with on-board staff (including guards) and routinely exchange information with them using radio or other communication systems. Team work is important as you work closely with others including shutters and signalmen.

A train driver needs to be punctual and reliable. Maintaining concentration is of critical importance in this role as is stamina and the ability to work both independently, with little social interaction, and as part of a team. You need good problem-solving and decision-making skills especially in emergencies, good eyesight, quick reflexes, and strong communication skills. You must have a positive customer focus and a very strong regard for safety. Sometimes rules and rail

gauges vary between states and the ability to learn and apply new knowledge and rules is critical in succeeding in this career (Due to train masters unfamiliar with trains operation resulted in trains delay).

- **LRV driver**

Be responsible for completing driving works, such as reorganization and outfitting of trains, main-track operation of electric passenger trains and depot dispatching and commissioning according to running regulations, and safely providing good-quality services on time, ensuring safe, stable, well-organized and controllable operation of light railways.

- **Engineering driver**

Correctly operate, reorganize and outfit trains according to department working requirements, ensuring safe shunting operation and running safety, and provide traction power guarantee for light railway construction, maintenance, engineering transportation and rescue & repair. For describing 'dirver' select two attributes are identified table 3.4 shows list of attributes and description.

No	Attributes Name	Description
1	Lrv Driver	controllable operation
2	Engineering Driver	safe shunting operation and running safety

Table 3.4 List of driver attributes

3.1.3.5. Timetable

Railway timetable is a program for space and time-wise running of railway passenger and/or freight traffic on a railway line. A timetable for a railway line or railway network, at least it contains a list of stations per railway line with the arrival and departure times for trains. Operating economy wise the timetable is the result of the traffic production planning for a given time period i.e. the validity period for the timetable [8].

To have an effective railway transport service it involves many procedures for railway operators [9]. Obviously, the timetable is not the only plan that needs to be composed in order to operate a railway system but also areas like demand estimation, rail line planning, rolling stock scheduling and crew scheduling too. This also indicates the dependencies between the timetabling process and other railway planning processes.

- **Block time**

The blocking time can be defined as the time during which a block between two signals is reserved exclusively to one train and therefore blocked for all other trains. It consists of the sight and reaction time of the train driver, approaching time, which is equivalent to the running time over the preceding block, the running time, clearing time needed for the full train length to leave the block, and setup and release time of the signaling system [52]. Using the blocking time, a route conflict between two trains corresponds to an overlap of their blocking times. The second train is within the sight distance of approaching signal and the first train has still not left the block.

- **Dwell Time**

Dwell time is the duration between the stopping time of the train at a station and the departure time from the station. It is measured between the instance the train wheel stops and the instance it starts to move again. The minimum dwell time is the necessary time for passengers to alight and board the train and sometimes it includes the door opening time. The first lost time can be obtained by calculating the time difference between train stop and door opened. Similarly, the second lost time can be obtained by computing the difference between doors closed and train start traveling [53]. In the Current AA-LRT situation the lost time in the third interval is longer than the lost time in the first interval. The main reason for this difference is that when the door is closed the train driver will come to his side's door as a rule for a matter of checking as the doors are closed and for indicating as the train is immediate to start traveling and of course to give the drivers code of greetings. Hence, during this activity, there is some lost time which leads to more time lost in the third interval than the first interval although they use the same mechanism (via electrically sliding) to open and close the doors.

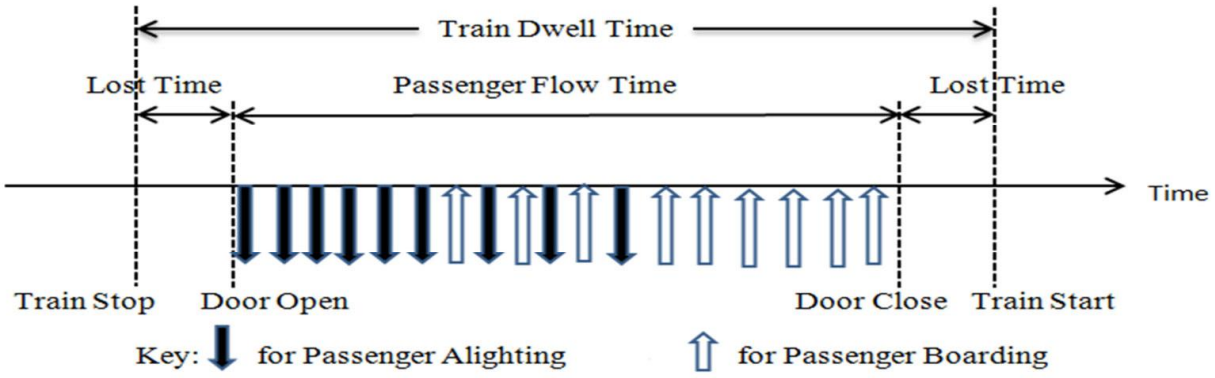


Figure 3.1 Activates during train dwell time

The duration of the second interval depends on the number of passengers flow. In some stations, the number of passengers can be fewer than other stations and hence the duration varies from one station to other stations [54].

- **Actual delay**

The identification of causes and prediction of delays that repeatedly occur on the network-wide level is a complex task that involves applying advanced data mining techniques for analyzing historical traffic realization data. Delay dependencies and identified structural errors in a timetable that result in systematic delays, can be effectively used not only for timetable improvement but also for real-time predictions.

Kecman [55]**,** developed a data mining tool for identifying delay dependencies in large networks. In this approach they distinguish between secondary delay due to capacity constraints and due to synchronization constraints. A separate model has been developed for each type of dependencies. These models are further used the aggregated traffic realization data in order to identify dependencies between delays. Further extensions include identifying

Multiple (capacity and connection) delay dependencies and improving robustness of the approach to measurement errors and outliers. The model was applied to a set of large-scale traffic realization data. Important dependencies due to capacity and synchronization constraints that were difficult to identify using correlation were discovered. And deferent interruptions(passenger over flow, power fluctuation, signal machine failures, poor outlook line, communicating gap and none master to driver).

- **Headway Time**

In its simplest case, headway time is the time gap between two consecutive trains running to the same direction through the same rail line. Hence to protect any collision, theoretically, these two trains must be separated by at least the braking time plus the length of the train ahead.

- **Number of Trips**

Number of trains operated during peak hours and flat hours according to the time table It is the result of the operation.

- **Daily Operation**

This deals with the assignment of train units to the rail lines in the timetable. When allocating the trains it is with the consideration of peak and off-peak hours since train canceled from or added to the timetable service regularly happens.

The following attributes are identified for describing time table and presented in table 3.5 below

No	Attributes Name	Description
1	Block time	block time between two signals
2	Dwell time	the duration between the stopping time of the train at a station
3	Actual delay	identified structural errors in a timetable
4	Number Of Trips	Transport service trip order
5	Headway Times	Time gap between two consecutive trains running
6	Daily Operation	Number of services to be operated in peak and flat hours

Table 3.5 List of time table attributes

3.1.3.6. Summary of attributes

Based on the relevance criteria the data, were selected from the above tables.

No	Attributes Name	Description
1	Train Speed	Standard design speed and operational speed
2	Train Model	Old Model or new model
3	Train Types	Comfortable and fast train ride and slowest train
4	Rolling Stock	Number of breakdowns in rolling stock
5	Manufacture Date	Manufacture to the train and last service date
6	Ballast	Support the weight of the Track
7	Track	Keep the rails at the correct distance apart
8	Sleepers	Sleepers are normally impregnated with preservative and, under good conditions
9	Rail	Wide base and Narrower top
10	Curves	Both vertical curves and horizontal curves
11	Gauge	The distance between the two rails
12	Turnout	Describe the junctions in track work where lines diverge
13	Rail Welding	Reduce damage to trains and eliminate the noise associated with rail joints
14	Discount Fee Passenger	Beneficial passengers
15	Fare Passenger	Passengers outside of discount fee user
16	Lrv Driver	controllable operation
17	Engineering Driver	safe shunting operation and running safety
18	Block time	block time between two signals
19	Dwell time	the duration between the stopping time of the train at a station
19	Actual delay	identified structural errors in a timetable
20	Number Of Trips	Transport service trip order
21	Daily Operation	Number of services to be operated in peak and flat hours
22	Headway Times	Time gap between two consecutive trains running

Table 3.6 summary of attributes

3.2. Data understanding

Next to identifying the problem and building a simple plan for solving the problem. the central item in data mining process is understanding. This includes listing out attribute with their respective values evaluation of their importance for this research and careful analysis of the data and structure is done domain expert by evaluating the relationships of the data with the problem at hand and the particular DM tasks to perform.

Finally, we verify the usefulness of the data with respect to the DM goals.

3.2.1. Data collection

The huge amount of data should be handling properly for different purpose like traffic control, network traffic performance analysis, report to higher official, network planning and optimization and to support decision making [8]

From the above AALRT data, train arrival time data is used for this research. The database of this train arrival time data was manipulated by oracle software system. The first and the major source of the data was network alarm data's which are sent from each network elements to the Operating Controlling Centers (OCC) during failures. During this research we used the data stored in Operating controlling centers database in the period between 2016 to 2018. During the interviews, domain expert of AALRT department explained that the operating controlling center database handles more than five different tables. Ballast, Sleepers, Rail, Curves, Gauge, Turnout, Rail Welding, Discount Fee Passenger, Lrv Driver, Engineering Driver, and Block time data have no data and missing some of attributes of track data underwriting and claim data separately, and manufactured date, train model have constant values. In AALRT Railways, the data of passenger train delays are being registered and recorded every day. At the weekend, these files are merged and from these weekly delay data are created. Finally, at the end of each month, these weekly files are merged to create monthly delay data. In this research, monthly files from 2016 to end of 2018 were used. In each year, the number of patterns represents the number of dispatched trains of that year. This data can be extracted to various file formats like Excel, PDF and CSV file type; for this research we used the data in Excel and CSV formats. These formats can be used directly or converting the format to WEKA tool for analysis.

The data from the OCC server is only extracted by authorized domain expert. So to get the official permission is needed. Managing the huge OCC data was one of the challenges and time consuming because the size of data increases, preprocess data become more complex. After

eliminating irrelevant and unnecessary data, a lot of 15040 datasets are used for the purpose of conducting this study. We select 22 attributes for this study based on their relevant for this research.

3.2.2. **Description of the collected data**

Description of the data is very important in data mining processing in order to clearly understand the data. Without such an understanding, useful application cannot be developed. as indicated before this research is done by collecting the data from AALRT OCC of 2016 up to 2018 records.

In this section, from the source described the above, the attributes with their data types and description are shown in the following table 3.6 below. The attributes have different data types like date,string,nominal and numeric data type.

NO	ATTRIBUTES NAME	DESCRIPTION	DATA TYPE	MISSING VALUE
1	Track Line	From x-station to y-station or From y-station to x-station.	Nominal	0%
2	Number Of Trips	Transport service trip order	Numeric	0%
3	Route Station	Akaki to mnilik and hayat to torhayeloch	Nominal	0%
4	Daily Operation	Date of transportation	Date	0%
5	Signal	How much station for the root 。	Numeric	15%
6	Power	It is the total breakdowns in AALAT power facilities recorded.	Nominal	0%
7	Rolling Stack	Number of breakdowns in rolling stock	Numeric	7%
8	Speed	Standard design speed and operational speed	Numeric	10%
9	Dwell time	Passengers getting on or off	Date	0%
10	Train Types	Comfortable and fast train ride and slowest train	Nominal	0%

Table 3.7 all attributes with their description of missing value of train arrival time managements

3.2.3. Derived attributes

In this research, based on the assessment of existing situation and discussion with domain experts, we derived two attributes (Passenger loading and Headway time) which are necessary and have direct relationship with train arrival time management.

The attribute passenger loading are derived from dwell time attributes in a concept that if there are dwell time increases where there is high number of passenger loading, there must be sufficient amount of train for this dwell time otherwise the higher passenger loading. So, amount of passenger loading has direct relationship with train arrival time management.

The other derived attribute was headway time. It also derived from train speed and this concept that if there is high train speed and less than head way time, and then became punctual. So, headway time has also direct relationship with train arrival time management.

No	Original Attribute	Derived Attribute	Data type	Value
1	Dwell time	Passenger Loading	Nominal	High, low
2	Train speed	Headway Times	Numeric	Hour/Minute/ Second

Table 3.8 Original attributes and derived attributes with their value type

3.3. Preparation of the Data

Currently real world database are highly susceptible to noisy, missing and inconsistent data due to their typically huge size and their likely the origin from multiple, heterogeneous resources. Low quality data will lead to low quality mining results [56]. Hence, Data preprocessing is required to have a data set which is suitable for analysis.

Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes, and data transformation, such as generalizing the data to higher-level concept or normalizing the data [57].

The purpose of data preprocessing is to clean select data for better quality. Data quality is multifaceted issue that represents one of the biggest challenges for data mining. It refers to the

accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presences of duplicated records, the lack of data standard, and the timelines of updates human error can significantly impact of the effectiveness of the complex data mining techniques, which are sensitive to understated differences that may exist in the data. To improve data quality, it is sometimes necessary to clean the data, which can involve the removal of duplicate records, normalizing the values used to represents information in the database [57].

Select data may be different formats, and then order to use the data needs to convert in to suitable format.

3.3.1. Data cleaning

Data cleaning is a process that attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data [58]. Described data cleaning as a time-consuming and procedure but it is absolutely necessary for successful data mining. Some of the data cleaning tasks that are applied in this study are removing outliers and handling missing value.

3.3.1.1. Handling Missing Values

For many real-world applications of data mining, even when there are huge amounts of data, the Subset of cases with complete data may be relatively small. A number of problems are faced while bringing the data into proper format. Missing data is the most common problem that comes up during the data analysis process. Missing values minimizing the accuracy of classification and rules generated by the selected data mining algorithm. Missing values lead to the difficulty of extracting useful information from that data set. Solving the problem of missing data is of a high Priority in the field of data mining and knowledge discovery. Handling missing values by appropriate methods does not affect the quality of the data. In this thesis the two widely used methods are applied. One is avoiding the missing data and other is data Imputation [59].

Avoiding the missing data is not time consuming and same time it is very easy to follow. But there are many drawbacks associated with this method. Deleting records may result in losing some information. If the sample data size is large avoiding some records or attributes may not affect the results, but still we need to keep in mind we are losing something.

Missing data is a problem that continues to plague data analysis methods. Even as our analysis Methods gain sophistication, the researcher has to continue to encounter missing values in fields,

especially in databases with a large number of fields. The absence of information is rarely beneficial. All things being equal, more data is almost always better. Therefore, the researcher considered carefully about how to handle the thorny issue of missing data [60]. Having efficient methods to fill up missing values extends the applicability in terms of accuracy for many DM methods. The accuracy of the tool is increased and with a larger training set better rules and decision trees can be developed which contributes towards better classification of the data to predict the train arrival time management, particularly in AALRT areas. A common method of handling missing values is simply to omit from the analysis the records or fields with missing values. However, this may be dangerous, since the pattern of missing values may in fact be systematic, and simply deleting records with missing values would lead to a biased subset of the field value is missing. Replace the missing value with the field mean (for numerical variables) or the mode (for categorical variables) [60]. Therefore, in this research study the investigator tried to handle the missing values by replacing missing value with the field mean, since they are numerical attributes. Table 3.10 summarizes attributes and percentage (%) of missing values associated with each other.

Attribute Name	No. of missing values	% of missing values	Mean value of missing values
Signal	562	15%	2
Speed	720	10%	6
Rolling Stack	300	7%	2

Table 3.10 Missing values and their percentage

As a result, the missing values of the dataset were handled in accordance with the above Suggestion. The missing value of Signal, Speed and rolling stack attributes were filled by their mean values since they are numeric value type.

3.3.1.2. Discretize Numeric Attributes

Discretization transforms numeric (continuous) attributes to nominal (categorical or discrete) attributes. The range of a numeric attribute is divided into intervals and each interval is given a label. Attribute values are replaced by the labels of the intervals into which they fall. Using discretization method can give generalized information which is easier and meaningful to

Interpret data mining results conducted on different data mining tasks. As a result, the experiment conducted on different data mining techniques and algorithms will have consistent representation of dataset. Generally, using reduced number dataset that prepared through discretization (interval labels) over large dataset (un generalized dataset) advances the mining results more efficient, consistent, simplified and easy to interpret and represent [61]. Here, passenger loading value and number of trips value attributes are transformed into categorical or discreet values. In this the study, to discrete the aforementioned attributes, Equal width discretization was used to divide the ranges of a numeric attribute into a specified number of intervals of equal width. This method considers the class information accordingly data discretization table 3.10 below.

Attributes	**Previous value**	**New value**
Passenger loading	0-14,14-120	Low, high
Number of trips	1- 8, 8- 16	minor, critical

Table 3.10 list of data discretization attributes

3.3.3. Data formatting

The datasets provided to this software were prepared in a format that is acceptable for Weka tools. it accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation File Format) file format (a file name with an extension of ARFF i.e. FileName.arff).

At first the integrated dataset was in an excel file format. To feed the final dataset into the Weka DM software the file is changed into other file format. The excel file was first changed into a comma delimited (CSV) file format. After changing the dataset into a CSV format the next step was opening the file with the Weka DM software. Then this file was saved with ARFF (Attribute Relation File Format) file extension. Now the dataset, which is in ARFF file format, is ready to be used in the WEKA software.

@attribute 'Track line' {'up line','down line'}
@attribute trips {critical,minor}
@attribute 'Train type' {'single train','couple train'}
@attribute station {EW,NS}
@attribute Speed numeric
@attribute 'Head way time' numeric
@attribute 'passenger load' {low,high}
@attribute power {on,off}
@attribute 'rolling stack' numeric
@attribute signal numeric
@attribute 'Daily operation ' {'flat hours','peak hours'}
@attribute 'punctual rate' {Punctual,Delay}

sample arff format data sets prepared for WEKA

Chapter Four

Experimentation and modeling

In this chapter, the researcher describes the techniques that have been used in developing a model to predict train arrival time management techniques. In addition to incorporated typical stages that characterize a data mining process. This study has been organized according to hybrid data mining process model, which is described and discussed methodology section in chapter one. Here the researcher discuss the experimentation process by relating the steps followed, the choice made , the task accomplished , the result obtained, evaluation of the model and results , and it present a way that the organization can easily understand and use it.

4.1. Model building

Modeling is one of the major tasks which is undertaken under the data mining in hybrid process model. In this phase different techniques can be employed for similar data mining problems. Some of the tasks include, selecting the modeling technique, experimental setup or design, building a model and evaluating the model. The output of experiments of classification models are analyzed and evaluation in terms of the details of the confusion matrix of the model. Furthermore, models of the different classification algorithms, such as decision tree and rule induction were compared with the respect to their performance measure such as Precision, Recall, F-measure and accuracy.

4.1.1. Selecting modeling technique

In this research, the supervised classification techniques are adopted. Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of these research three classification algorithms has been selected for model building. The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, J48, JRIP and *Naïve Bayes* were used for experimentation of this study. In this work attempt was done to be a model using select algorithms for classification of train arrival time management technique.

Firstly, the J48 decision tree algorithms is chosen because it is one of the most common decision tree algorithms that are used today to implement classification techniques WEKA.

Finally JRIP rule induction techniques was application it is one of the basic and the most popular rule induction algorithms. Classes are examined increasing size and an initial set of rules for the class is generated using incremental reduced error.

4.2. Experimental design

The model was built based on the default 66% percentage split and 10-fold cross validation. The default ratio is 66% for training and 34% for testing. In 10-fold cross validation, the initial data are randomly partitioned into 10 mutually exclusive subsets or folds, 1,2,3,……10 ,each approximately equal size. The training and testing are performed 10 times. In the first iteration, the first fold is reserved as a test set, and the remaining 9 folds are collectively used to train the classifier [62].

The accuracy estimate is the overall number of correct classifications from the 10 iterations divided by the total number of sample in the initial dataset [19]. Generally, a procedure or mechanism was used to the test model's quality and validity is needed to beset before the model is actually built. In order to perform the model building process of this stately. we use 15040 dataset with 12 attributes to investigate the model. The training and testing dataset were prepared by purposive sampling technique from the original dataset.

In this study, we perform six experiments with Bayes, decision tree and rule induction algorithms. Since this research is experimental research design, we use J48 algorithms from decision tree, JRIP algorithms from rule induction and Naïve Bayes from Bayes. In order to validate and compare the classification performance of the techniques the 10-fold cross validation and percentage split are used. Both methods are tested with default value by using 10-fold cross validation and 66% percentage split. List of experiments conducted table 4.1 below.

Experiment	Algorithms	Test mode
1	J48 decision tree	10-fold cross-validation
2	J48 decision tree	Percentage split (66%)
3	JRIP rule induction	10-fold cross-validation
4	JRIP rule induction	Percentage split (66%)
5	Naïve Bayes	10-fold cross-validation
6	Naïve Bayes	Percentage split (66%)

4.3. J48 decision tree based model building

A decision tree is a classifier to express as a recursive partition of the instance space. Decision tree consist of nodes that form a rooted tree, meaning it is a directed a tree with a node called root that the incoming edges. All the nodes have exactly one incoming edge. Node with outgoing edges is called an internal or test node. All other nodes are called leaves. In a decision tree each internal nodes splits the instances spaces in to two or more sub-spaces according to a certain discreet function of the input attribute values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute values.

Using J48 decision tree two experiments were conducted by partioning the data by 10-fold cross validation and 66% percentage split. Experiment result is shown in table 4.2

Experiments	Algorithm	Test option	Accuracy	Recall	Precision	F-Measure	Class
1	J48	10-fold cross-validation	95.4189 %	96.1%	91.8%	93.9%	Punctual
				95%	97.7%	96.3%	Delay
	Weighted Avg			95.4%	95.5%	95.4%	
2	J48	Percentage split(66%)	95.5612 %	93.5%	94.3%	93.9%	punctual
				96.8%	96.3%	96.5%	Delay
	Weighted Avg			95.6%	95.6%	95.6%	

Table 4.2 performance results for J48 algorithm with 10-fold cross validation and percentage split (66%)

With 10-fold cross validation in table 4.2, the J48 learning algorithm scored an accuracy of 95.4189 %. This result shows that out of the total training datasets 14351(95.4189 %) records are correctly classified instances, while 689(4.581%) of the record are incorrectly classified. On the other hand, the experiment conducted using 66% percentage split results in 95.6% correct classified instances and 4.4% incorrect classified instances. Generally, from the two experiment conducted before, the model developed with the Percentage split(66%) test option given better classification performance of identifying the train arrival time management techniques to their

respectively class category. Therefore, among the two decision tree models built in the forgoing experimentations, J48 decision tree with 66% percentage spilt is selected

4.4. JRIP Rule induction model building

JRIP is one of the basic most popular rule induction algorithms. To building rule model 15016 dataset was used as an input to the system for the experiments of 10-fold and default percentage splits 66%.

Using JRIP rule induction two experiments were conducted by partioning the data by 10-fold cross validation and 66% percentage split. Experiment result is shown in table 4.3 blow

Experiments	Algorithm	Test option	Accuracy	Recall	Precision	F-Measure	Class
3	*JRIP*	10-fold cross-validation	95.3391 %	96.3%	91.5%	93.8%	Punctual
				94.8%	97.8%	96.3%	Delay
	Weighted Avg			95.3%	95.5%	95.4%	
4	*JRIP*	Percentage split (66%)	95.4439%	93.5%	94.0%	93.8%	punctual
				96.6%	96.2%	96.4%	Delay
	Weighted Avg			95.4%	95.4%	95.4%	

Table 4.3 performance results for JRIP rule induction algorithm with 10-fold cross validation and percentage split (66%)

With 10-fold cross validation in table 4.3, the JRIP rule induction learning algorithm scored an accuracy of 95.3391 %. This result shows that out of the total training datasets 14339 records are correctly classified instances, while 701 (4.6609%) of the record are incorrectly classified. On the other hand, the experiment conducted using 66% percentage split results in 95.4439% correct classified instances and 4.5561% incorrect classified instances.

Generally, from the two experiment conducted before, the model developed with the Percentage split(66%) test option given better classification performance of identifying the train arrival time management techniques to their respectively class category. Therefore, among the two rule

induction models built in the forgoing experimentations, JRIP rule induction with 66% percentage spilt is selected

4.5. Naïve Bayes Classifier Model Building using WEKA Software

It is method of classification that does not use rules, a decision tree or any other explicit representation of the classifier. Rather, it uses the branch of Mathematics known as probability theory to find the most likely of the possible classifications. The Naïve Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which the researcher used to calculate the probability of each of the possible classifications in turn. Having done this the researcher chooses the classification with the largest values.

Using *Naïve Bayes* two experiments were conducted by partioning the data by 10-fold cross validation and 66% percentage split. Experiment result is shown in table 4.4 blow

Experiments	Algorithm	Test option	Accuracy	Recall	Precision	F-Measure	Class
5	*Naïve Bayes*	10-fold cross validation	93.9894%	100%	86%	92.5%	Punctual
				90.5%	100%	95%	Delay
	Weighted Avg			94%	94.8%	94.1%	
6	*Naïve Bayes*	Percentage split(66%)	94.2902%	100%	86.5%	92.8%	punctual
				91%	100%	95.3%	Delay
	Weighted Avg			94.3%	95.1%	94.4%	

Table 4.4 performance results for *Naïve Bayes* algorithm with 10-fold cross validation and percentage split (66%)

With 10-fold cross validation in table 4.3, the *Naïve Bayes* learning algorithm scored an accuracy of 93.9894%. This result shows that out of the total training datasets 14136 records are correctly classified instances, while 904 (6.0106%) of the record are incorrectly classified. On the other hand, the experiment conducted using 66% percentage split results in 94.2902% correct classified instances and 5.7098 % incorrect classified instances.

Generally, from the two experiment conducted before, the model developed with the Percentage split(66%) test option given better classification performance of identifying the train arrival time management techniques to their respectively class category. Therefore, among the two *Naïve Bayes* models built in the forgoing experimentations, *Naïve Byes* with 66% percentage spilt is selected.

4.6. Comparison of J48, Naïve Bayes and JRIP

Conducting a better classification technique for building model, which perform the prediction of train arrival time management techniques are one of the aims of this research. For this performing classification models are compared. Table 4.5 present algorithms are greatest performance.

Types of algorithm	Accuracy	Recall	Precision	F-measure
J48 decision tree	95.5612 %	95.6%	95.6%	95.6%
Naïve Bayes	94.2902%	94.3%	95.1%	94.4%
JRIP rule induction	95.4439%	95.4%	95.4%	95.4%

Table 4.5 performance comparison of the selected models

As shown in table 4.5, all algorithms are performing well, with accuracy of more than 90%. However, J48 decision tree algorithm registered the highest accuracy of 95.5612%. hence the model conducted J48 decision tree is selected for determining train arrival time management.

4.6.1. Confusion matrix classifier

The J48 classifier one of selected algorithms test train arrival time project. The confusion

Matrix for the J48 classifier shown in table 4.6 demonstrate the total 5114 records 1752 records are correctly classified as category “punctual” and 3135 records are correctly classified as category “delay”. The classifier incorrectly classified 122 records categorize as “delay” and 105 records categorize as “punctual”. It has totally 227 attributes are misclassified both category of “punctual” or “delay”. While the accuracy of the classifier is correctly predict the class value as “punctual” and “delay” is 95.6% which records J48 classifier is best result of the JRIP rule inaction and Naïve Bayes algorithm.

The confusion matrix of the selected J48 algorithm is also shown in the table 4.6 below.

Confusion matrix		
A	B	Classified as
1752	122	A= punctual
105	3135	B=Delay

As shown in table 4.6 there is misclassification happens between Punctual and Delay classes. This is because as shown in chapter three section 3.1 speed characteristics of the attributes, these second classes have high number of instances than the first class and due to this variation, the prediction model was highly influenced by these second classes. Then in order to avoid this variation between the data, we use resampling techniques.

As discussed with domain experts, the other reason for misclassification between these two classes is there is a relationship between these classes in that if Punctual occur, there is also a possibility that Delay to be occurred.

4.7. Rule generated by selected algorithms

As discussed before those experiments conducted in supervised approach, the J48 algorithm with Percentage split (66%) gives a better classification experiment of identifying the newly arrive train arrival time management techniques and high model building time the other algorithms used in this research. The rules indicate that the possible conditions in which the OCC records could be classified in each of the classes. Form this model a set of rules are extracted by traversing the decision tree and generating a rules for each leaf and making a combination of all the tests found on the path from the root to the leaf node.

One of the interesting rules detected is how much speed is critical in order to predict the train arrival time management techniques. The following are some of the interesting rules extracted from the decision tree. Therefore those cover more cases and have better accuracy are chosen. The following rules indicate the possible conditions in which a train arrival time management techniques could be classified in each of Delay and Punctual classes.

Rule 1

If Speed <= 50, Head way time <= 17, Punctual (974.94/1.94)

Rule 2

If Speed <= 50, Head way time > 17 and passenger load = high: Delay (1243.48)

Rule 3

If Speed <= 50, Head way time > 17, passenger load = low and signal >= 0: Delay (699)

Rule 4

If Speed <= 50, Head way time > 17, passenger load = low, signal <= 0 and power = off: Delay (87.54)

Rule 5

If Speed <= 50, Head way time <= 17, passenger load = low, signal <= 0, power = on and rolling stack >= 0: Punctual (77.0)

Rule 6

If Speed <= 50, Head way time <= 17, passenger load = low, signal <= 0, power = on, rolling stack <= 0 and Daily operation = flat hours: Punctual (952.0/339.0)

Rule 7

If Speed <= 50, Head way time > 17, passenger load = low, signal <= 0, power = on, rolling stack <= 0, Daily operation = peak hours and Train type = single train: Delay (608.69/197.0)

Rule 8

If Speed <= 50, Head way time <= 17, passenger load = low, signal <= 0, power = on, rolling stack <= 0 and Daily operation = peak hours, Train type = couple train and trips = critical: Punctual (288.9/123.9)

4.8. Use of Knowledge

After evaluating discovered knowledge, the last step is using this knowledge for the industrial purposes. In this step the knowledge discovered is incorporated in to performances system and take this action based on the discovered knowledge.

In this research the discovered knowledge is used by integrating the user interface which is designed by Visual C# with a Weka system in order to show the prediction of train arrival time management techniques.

In order to predict the feature train arrival time management techniques, we analyzed the current train arrival time management techniques based on the available data by generating rule selection algorithm. Then we used the generated rules for implementing their using visual C# programming language to predict the feature train arrival time management techniques.

4.9. Discussion of the results with domain experts

The experts explained that more attention was given to passenger loading where there is higher arrival times because when the number of user in specific are increase the passenger loading. Due to this the train arrival time is delayed. The delay status has also direct impacts on train arrival time management techniques. In AALRT OCC there are passengers loading when there is high amount of effects like passenger flow and rolling stack problems. So, because of these passenger loading effects increases train arrival time management problems. The domain experts argued that, the discovered rules are acceptable, and the passenger loading has directed relationship and great impact on train arrival time management techniques.

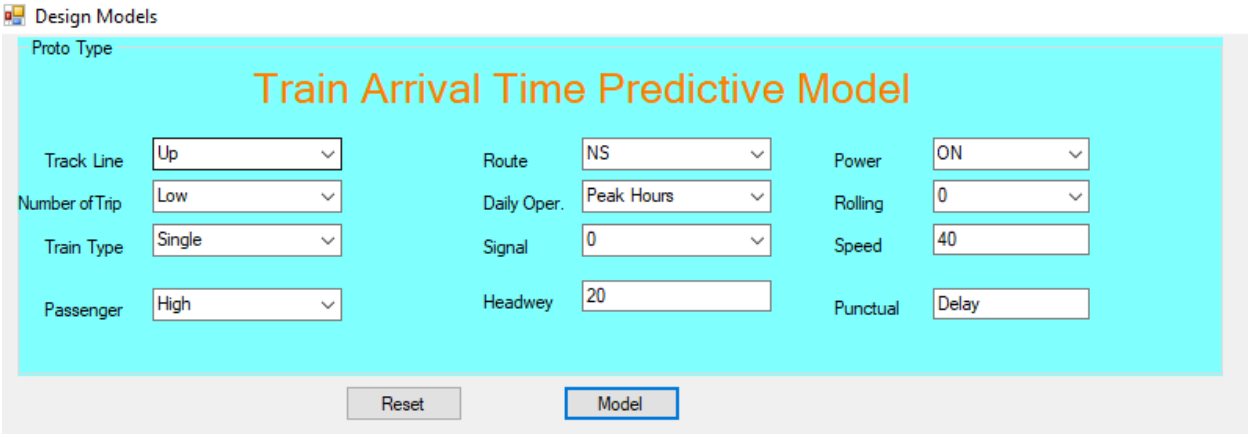


Figure 4.1 Train arrival time management prediction model sample prediction outputs

4.8.1. User Acceptance Testing

According to Luo [63], User Acceptance Testing is used to conduct operational readiness of a product, service or system as part of a quality management system. It is a common type of non-functional software testing, used mainly in software development and software maintenance projects. This type of testing focuses on the operational readiness of the system to be supported. It is done when the completed system is handed over from the developers to the customers or users. The purpose of user acceptance testing is rather to give confidence that the system is working than to find errors.

User Acceptance Testing verifies the system's behavior is consistent with the requirements. These tests will reveal defects within the system. The work associated with it begins after requirements are written and continues through the final stage of testing before the user accepts the new system [63].

The goal of User Acceptance Testing is to assess if the system can support day-to-day business and user processes and ensure the system is sufficient and correct for business usage. The primary objective of User Acceptance Testing is to demonstrate that you can run your business using the system if it is fit for the purpose [63].

In this research we perform User Acceptance Testing by presenting and discussing with the organization's domain experts. The following areas are discussed in detailed with domain experts and discussions are presented below.

4.8.2. Efficiency

Efficiency is the ability to avoid wasting materials, energy, efforts, money, and time in doing something or in producing a desired result. In a more general sense, it is the ability to do things well, successfully, and without waste times. In scientific terms, it is a measure of the extent to which input is well used for an intended task or function (output). It often specifically comprises the capability of a specific application of effort to produce a specific outcome with a minimum amount or quantity of waste, expense, or unnecessary effort.

In this research efficiency is considered as a time taken to predict the train arrival time management by taking inputs from the user. As discussed in chapter one, currently in Addis Ababa light railway transit analysis is done by traditional simple statistical methods which need more time to operate because every operation is done manually.

During our presentation and discussion with domain experts, we conduct sample experiments and compare the efficiency between the current statistical method and our new prediction method. From these sample experiments, our new train arrival time management model prediction method become more efficient and every domain expert agreed up on this. They argued that due to this efficiency improvement they can minimize more than half a time taken before.

4.8.3. Effectiveness

Effectiveness is the capability of producing a desired result or the ability to produce desired output. When something is deemed effective, it means that it has an intended or expected outcome or produces a deep impression.

In this research effectiveness is considered as the accuracy to predict right train arrival time management. As discussed in chapter four, we perform experiments with J48 tree algorithm,

JRIP from rule induction algorithm and *Naïve Bayes*. As a result, J48 algorithm registers better performance of 95.56% accuracy.

During our presentation and discussion with domain experts, our experimental results are discussed, and they give a recommendation to improve this performance.

4.8.4. Easy to learn and Easy to remember

Making a product easy to use is one of the non-functional requirements for any product. In order to make this research outputs easy to use, we prepare sample screen shots on the document.

In this study, we perform user acceptance testing to evaluate systems efficiency, effectiveness, easy to learn and easy to remember point of view.

In this study, a total of 8 domain experts from train arrival time management specifically from railway performance and quality analysis sections participated to evaluate the systems acceptance. Each of the study participants are asked to give feedback on the acceptability of the prediction and to rate it on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree). Summary of the result is presented in table 4.9 below.

Questionnaires		Strongly Agree (5)	Agree (4)	Undecided (3)	Disagree (2)	Strongly Disagree (1)
Efficiency	The prediction response is fast.	90%	10%	-	-	-
	The prediction saves energy & materials.	90%	10%	-	-	-
Effectiveness	The prediction is Reliable.	80%	10%	10%	-	-
	The prediction produces a desired result	70%	20%	5%	5%	
Easy to Learn:	The prediction system is Easy to learn.	80%	10%	10%	-	-
	The prediction system is User friendly	70%	20%	10%	-	-
Easy to Remember	The prediction system is easy to remember	75%	20%	5%	-	-
	The prediction model is explicit.	65%	25%	10%	-	-

Table 4.2 Experts response summary on the proposed train arrival time management prediction model.

This study revealed that from 8 domain experts 7 of them confirm that this train arrival time management prediction model was much efficient, and it saves their energy and materials while comparing with the way they perform currently which is the simple statistical method. In the

case of effectiveness, the domain experts revealed that this prediction model produces a desired result. In order to make the prediction near perfect, there is a need to enhance the performance of the model near 100%. Some (5%) of the domain experts were disagree with the effectiveness result because they stated that in case of train arrival time management analysis the reason must have to be perfect because it has a significant impact on the organizations revenue.

According to this study, the performance result of the prediction model scored an accuracy of 95.56% which is good. Most of the domain experts satisfied with the prediction results but some of them are strongly disagree with the prediction because in some cases the model doesn't tolerate errors. In order to make the prediction more accurate and error tolerable, we advise integration of the discovered classification rules with knowledge-based system.

Concerning the extent to which the prediction model was easy to learn and easy to remember, the respondents reply shows that the prediction model was user friendly and it is more explicit than before. However, there are up to 10% respondents who are undecided because they are agree to some extent on the systems user friendless and to some extent they are not agree.

Overall user acceptance criteria, 90% of the domain experts agreed that this train arrival time management prediction system is much efficient, it saves energy and materials, the prediction is reliable, the prediction produces a desirable result, the prediction system is easy to use and remember, it is user friendly and the prediction model is more explicit than before. The domain experts also suggested that there need to enhance the performance of the model to make the prediction near to 100%.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

The traditional method of turning data into knowledge relies on manual analysis and interpretation. In train arrival time management sector, different analysis was made manually, and different reports are generated. The report becomes the basis for future decision making and planning for train arrival time management expansion, performance and quality of service (QoS) evaluation. Data analysis using traditional simple statistical tools is slow, expensive, and highly subjective. Huge amount of data is generated from time planning elements and stored in train arrival time databases for different purposes. Hence, manual data analysis and interpretation is impractical.

The objective of this research was to develop a predictive model for train arrival time management of Addis Abeba railway transit by using data mining techniques. To achieve this objective, we use Operating Controlling Center (OCC) data because the OCC data includes enough information about train arrival time management. Due to the fact that the OCC data is very huge and requires more space, the train arrival time servers stored no more than 4 years data. Therefore, we took 3 years data of the train arrival time company.

This research proposes data mining to overcome the problem of manual data analysis. Hybrid process model was used while undertaking the experimentation. The study was conducted using WEKA software version 3.8 and three data mining algorithms of classification techniques was used, namely J48, *Naïve Bayes* and JRIP.

We use the purposive sampling techniques to extract the data. In ordered to extract the huge damp file from the train arrival time database, we use oracle database software. In ordered to manage the data in application software (in MS-Access and MS-Excel), file splinter software is used. After eliminating irrelevant and unnecessary data, a total of 15040 datasets are selected from OCC and used for the purpose of conducting this study. Two derived attributes and out of 22 attributes ten relevant attributes from OCC network database server are selected to conduct this research. It has been preprocessed and prepared in ARFF format which is suitable for the DM tasks.

The J48 decision tree algorithm registered better performance of 95.5612 % accuracy and processing speed of 0.03 sec running with Percentage split using 12 attributes than any experimentation done for this research.

One of the basic targets of data mining is to compare different models and to select the best classification accuracy accordingly. Therefore, detailed experimentation for different models has been conducted. Among the four models, the J48 decision tree algorithm with Percentage split registers better performance and processing time than other experimentations done in this research. It registers an accuracy of 95.5612% and processing time of 0.03 sec.

The finding of this study shows that identify for train arrival time management. The mining result identified that speed is the major factor, followed by generator status and availability of resources. Headway time, passenger loading and signal used also have a great contribution. The analysis which was closely undertaken with domain experts are achieved a good result.

The result of this research shows that applying data mining to analyze traffic network data helps train arrival time management to improve QoS and make decision based on the information discovered from the analysis. Providing QoS to its customers will lead the organization to satisfy its customers and revenue augmentation for Railway Company.

The main challenges that the algorithm encountered is inability to classify the Punctual and Delay classes. This misclassification between these two classes is happened due to the reason that there are unbalanced data instances between these classes; the prediction model was highly influenced by these two classes. Then in order to avoid this variation between the data, we use resembling techniques.

As discussed with domain experts, the other reason for misclassification between these two classes was there is a relationship between these classes in that if Punctual occur, there is also a possibility that Delay to be occurred.

5.2. Recommendations

This research is mainly conducted for an academic purpose. This research has proven the applicability of different DM classification techniques namely, J48, JRIP and *Naïve Bayes* algorithms which automatically discover hidden knowledge that are interesting and accepted by domain experts. Based on the investigations of the study, the following areas are given as a recommendation for the future.

• In this research, we use only Addis Ababa light railway transit data however further investigation is needed by including other regional train arrival time data so as to comprehensively see the cause for train arrival time management determines.

In order to design an intelligent train arrival time management system, there is need to conduct a study on the integration of the discover classification rules with knowledge-based system.

• In this research, we use one of data mining technique, classification. The association rule-based data mining techniques can be used in future:

- •To study the relationship or extract interesting correlations between attributes
- •To study the cause and effect or associations among sets of items

• This study has attempted to apply DM techniques on Operating controlling center data, but it could also be applied in other Train arrival time data like Automatic Train Supervision (ATS), Safety critical application and other purposes.

Reference

[1] F. Cordeau, "A Survey Of Optimization Models For Train Routing And Scheduling," *Transp. Sci,* Vol. 32, Pp. 380-404, 1988.

[2] I.Sahin, "Railway Traffic Control And Train Scheduling Baed On Inter-Rail Conflict Management," *Transp. Res. – Part B,* Vol. 33, Pp. 511-534, 1999.

[3] B. Schittenhelm, Investigation On The Performance Of Train Timetable For The Case, Addis Abeba: Bernd Schittenhelm , 2013.

[4] V. Aalst, "Models For Predictive Railway Traffic," *Belgrade,* 2011.

[5] W. Rygielski, "Data Mining Techniques For Customer Relationship Management," *Technology In Society,* Vol. 24, Pp. 483-502, 2002.

[6] T.Adane, "Mining Insurance Data For Fraud," *Addis Ababa, Ethiopia,* 2011.

[7] Y. Asnake, "Railway Infrastructure Health Monitoring Using Wireless Sensor," *Addis Abeba,Ethiopia,* 2015.

[8] D. Pacciarelli, "Models And Algorithms For Traffic Management Of Rail Networks, Technical Report," *Universit`A Roma Tre,* 2002.

[9] T. Haylekiros, "Investigation On The Performance Of Train Timetable For The Case," *School Of Electrical And Computer Engineering,* Vol. Vol.7, P. 49, 2017.

[10] Stefan, "The Multi-Objective Railway Timetable Rescheduling," Eurailpress, Hamburg, 2016.

[11] P. Hansen, "Models For Predictive Railway Traffic," *University Of Belgrade,* 2007.

[12] K. Corvo, "Why Is Methodology Important In Research," In *Etherland*, 2018.

[13] E. Frank, "Data Mining: Practical Machine Learning Tools And," San Fransico, 2ND , 2000.

[14] S. Kurgan, Knowledge Discovery Approach, New York, 2007.

[15] " Codeproject.Com/Kb/Cross-Platform/Benchmarkcppvsdotnet.Aspx .".

[16] Gonz´Alez, "On-Line Timetable Rescheduling In Regional Train Services," *European Journal Of Operational Research,,* Vol. 33, Pp. 387- 398.

[17] Witten, "Data Mining: Practical Machine Learning Tools And Techniques," *Elsevier: Morgan Kaufmann,* 2005.

[18] D. Olson, "Advanced Data Mining Techniques," *Springer,* 2008.

[19] U. Fayyad, "Knowledge Discovery And Data Mining: Towards A Unifying Framework," *Gregory ,*

1996.

[20] J. Seifert, "Data Mining: An Overview Of National Security Issues," Pp. 201-217, 2004.

[21] K.Prabha, A Hybrid Approach For Data Clustering Using Data Mining Techniques, India, 2014.

[22] K. Tsiptsis, "Data Mining Techniques In Crm," John Wiley And Sons, Ltd, 2009.

[23] S. Danso, "An Exploration Of Classification Prediction Techniques In Data Mining:The Insurance Domain," Bournemouth University, 2009, P. September.

[24] M. Dandi, "Application Of Data Mining Techniques For Customer Segmentation In Insurance Business: The Case Of Ethiopian Insurance Corporation," *Addis Abeba, Ethiopia,* July 2016.

[25] Gordon, S., "Data Mining Techniques For Marketing, Sales, And Customer Relationship Management," London, Second Edition, 2004.

[26] J. Han, "Data Mining Concepts And Techniques," Second Edition, 2006.

[27] O. Delen, "Advanced Data Miningtechniques," ,Berlin Heidelberg: Springer-Verlag, 2008.

[28] M. Frehiwot, "Predictive Model For Ecx Coffee Contracts," October 2014.

[29] K. Klas, "Utilization Of Data Mining Techniques For Prediction Of Diabetes Disease," Vol. 6, 2013.

[30] E. Frank, Data Mining Concepts And Techniques, Elsevier Inc: Morgan-Kaufmann, 2012.

[31] D. Olson, Advanced Data Mining Techniques, Springer, 2008.

[32] L. Rokach, "Top-Down Induction Of Decision Trees Classifiers-A Survey," *Ieee Transactions On Systems, Man, And Cybernetics,* Pp. 476-487, 2005.

[33] R. Haryana, "Decision Tree: Data Mining Techniques Department Of Computer Science," India, 2010.

[34] J. Jackson, "Data Mining:A Conceptual Overview," 2002.

[35] M. Hemalatha, "A Perspective Analysis Of Traffic Accident Using Data Mining Techniques," *International Journal Of Computer Applications,* 2011.

[36] R. Haryana, Decision Tree: Data Mining Techniques Department Of Computer Science, India, 2010.

[37] A. Rajput, "J48, Part And Jrip Rules For E-Governance Data," 2011.

[38] S. Edin, "Data Mining Aproach For Pridicting Stunents Performance," P. 10, 2012.

[39] M. Kamber, Data Mining: Concepts And Techniques. 2ND Ed., San Francisco, Usa: Morgan, 2006.

[40] P. Langley, "Estimating Continuous Distributions In Bayesian," Canada, 1995.

[41] U. Fayyad, "From Data Mining To Knowledge Discovery In Databases," 1996.

[42] N. Sheth, "Customer Relationship Management," *Journal Of Economic And Social Research*, Vol. 3, Pp. 1-34.

[43] L. Meng, "Advanced Monitoring And Management," *Journal Of Rail Transport Planning ,* 2012.

[44] H. Assefa, "Investigation On The Performance Of Train Timetable For The Case," Vol. 7, 2017.

[45] N. Ouburg, "Prorail Internal Specification," 2005.

[46] G. Linoff, "Mastering Data Mining: The Art And Science Of Customer Relationship Management," 1999.

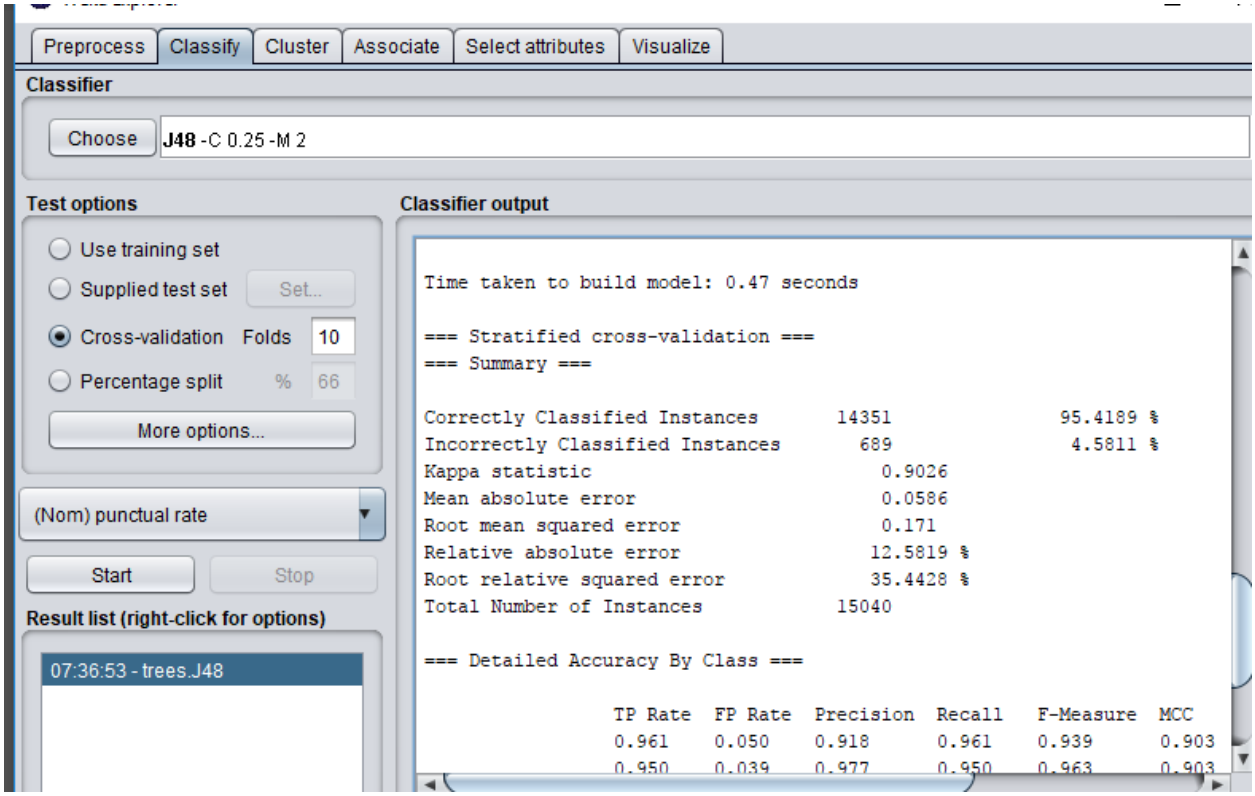
[50] T. A. Vlasenko, "Models For Predictive Railway Traffic Management," P. 21, 2014.

ANNEXES

Annex-1: The original collected sample data

Track line	trips	Train type	station	Speed	Head way	passenger	power	rolling sta	signal	Daily oper	punctual r
up line	critical	single trai	EW	50	17	low	on	0	0	flat hours	Punctual
up line	critical	single trai	EW	50	18	low	on	0	1	flat hours	Delay
up line	critical	single trai	EW	50	17	low	on	0	1	flat hours	Delay
up line	critical	single trai	EW	40	20	low	on	0	1	flat hours	Delay
down line	minor	single trai	EW	50	17	high	on	0	1	flat hours	Delay
down line	critical	single trai	EW	40	18	high	on	0	1	flat hours	Delay
down line	critical	single trai	EW	50	17	high	on	0	1	flat hours	Delay
down line	critical	single trai	EW	40	18	high	on	0	1	flat hours	Delay
down line	critical	single trai	EW	50	16	low	on	0	0	flat hours	Punctual
down line	minor	single trai	EW	50	17	high	on	0	1	flat hours	Delay
up line	critical	couple tra	EW	50	17	low	on	0	1	flat hours	Delay
up line	critical	couple tra	EW	40	20	low	on	0	1	flat hours	Delay
up line	critical	couple tra	NS	50	18	low	on	0	1	flat hours	Delay
down line	critical	couple tra	NS	40	18	high	on	0	1	flat hours	Delay
down line	critical	couple tra	NS	40	18	high	off	1	1	flat hours	Delay
down line	critical	single trai	NS	50	18	low	on	0	1	flat hours	Delay
down line	critical	single trai	NS	50	17	low	on	0	1	flat hours	Delay
up line	minor	single trai	NS	50	18	low	on	0	1	flat hours	Delay
up line	critical	single trai	NS	50	17	low	on	0	1	flat hours	Delay
down line	critical	single trai	NS	50	18	low	on	0	1	flat hours	Delay
up line	critical	single trai	NS	50	18	low	off	1	1	flat hours	Delay
up line	critical	single trai	NS	60	17	low	on	0	0	flat hours	Punctual
down line	critical	single trai	NS	40	18	high	on	0	1	flat hours	Delay
down line	critical	single trai	NS	50	17	high	on	0	1	flat hours	Delay
down line	critical	single trai	NS	40	19	high	on	0	1	flat hours	Delay

Annex-2: The snapshot running information of J48 with percentage split technique


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) punctual rate
Start
Stop
Result list (right-click for options)
07:36:53 - trees.J48
Classifier output
Time taken to build model: 0.47 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 14351 95.4189 %
Incorrectly Classified Instances 689 4.5811 %
Kappa statistic 0.9026
Mean absolute error 0.0586
Root mean squared error 0.171
Relative absolute error 12.5819 %
Root relative squared error 35.4428 %
Total Number of Instances 15040
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC
0.961 0.050 0.918 0.961 0.939 0.903
0.950 0.039 0.977 0.950 0.963 0.903

Annex-4: J48 train arrival time management decision tree