Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 217 (2023) 72–81

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Industry 4.0 and Smart Manufacturing

Smart Trip Prediction Model for Metro Traffic Control Using Data Mining Techniques

Majid Mirbod[a], Hamidreza Dehghani[b]

[a]*Faculty of Management, Islamic Azad University, Tehran North Branch, Tehran, Iran*

[b]*Faculty of Management, Islamic Azad University, Karaj Branch, Karaj, Iran*

Abstract

In the urban rail transportation industry, collecting various data, including the number of trips, the number of fleet breakdowns, the number of breakdowns of railway lines, etc., contributes significantly to future planning and optimization of resources and costs. We are faced with big data analysis due to the daily production of large volumes of data in the urban rail transport industry. This study aims to find a trip prediction model for Tehran Metro to do trip decision-making. This paper provides a brief introduction to knowledge discovery from failure databases and presents the data mining methodology. The relevancy of data mining for a trip prediction model will be depicted. In this sense, considering the relevance of data collected on the metro control center and intelligent trip prediction, this paper presents a functional architecture of a predictive trip model using data mining techniques. Data Mining will identify behavior patterns, allowing a more accurate early detection of faults in the daily operation. This research has investigated three situations, including delayed trips, incomplete trips, and canceled trips, using actual data recorded over the past four years, including non-programmed failures and their impact on trips. A model with more than 90% (90.48%) accuracy suggests a trip status forecast on this line. Also, the research has used Rapid Miner software and the Decision tree technique in data mining and its rules. Also, the kappa index of 0.792 is calculated for the validation model. Therefore, the proposed model enables the prediction of three types of incomplete traffic, delayed trips, and cancellation of trips due to emergency failure at the time of the operation of the Tehran Metro.

Keywords: Data mining; Decision making; Smart Trip Prediction; Knowledge discovery; Tehran Metro Control Traffic.

*Corresponding author.
E-mail address: mjmirbod@yahoo.com

10.1016/j.procs.2022.12.203

1. Introduction

Virtually, all manufacturing enterprises use powerful data acquisition systems to collect analyses and transfer data from nearly all the organizational processes. These data may be related to machines, products, processes, maintenance, quality control, failure detection, etc., and are typically stored in databases at various stages.

The use of databases and statistical techniques is well established in engineering because intelligently analysed data is a valuable resource, providing new insights and significant competitive advantages. Traditionally, statistical techniques were used to find patterns in manufacturing data. However, recent proceedings in information technology, data acquisition systems, and storage technology, along with the emergence of machines with growing complexity and sensor equipment, lead to an overwhelming amount of data, permanently increasing and containing hundreds of attributes, which should be considered simultaneously to model the system's behaviour accurately. This complexity calls for new techniques and tools for the automated extraction of useful knowledge from huge amounts of raw data. Data Mining is about solving these problems by applying mathematical models to discover patterns in data already in databases automatically. Data Mining is, thereby, one step in Knowledge Discovery in Databases (KDD), which denotes the entire process of turning low-level data into useful knowledge. This process often includes the following important stages. The first step involves understanding the application domain, which is very important, especially when analysing manufacturing data. For the successful generation of new knowledge, a close collaboration of domain, data, and data mining experts is needed. The goals and tasks of the data mining process have to be determined, and all the factors, which might affect the manufacturing process under study, should be revealed and understood. The second step includes selecting the target data set. Since data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined in an acceptable timeframe. The data sets are often stored in various databases, which require additional integration. Next, the data sets have to be pre-processed, comprising inter alia transformation, handling missing data, and removing noise. The fourth step, finally, is data mining for the extraction of patterns from the data and involves the selection and application of appropriate (mathematical) data mining algorithms, a along with the development of a model, which describes the pattern. This step is often accompanied by 'traditional' statistical analysis and data visualization. In the last step, the extracted patterns must be interpreted and verified. Here, the evaluation from the domain experts is essential to transfer the patterns into new knowledge. Usually, some of the KDD steps have to be iterated several times to reach this goal. Once useful patterns are found and described, they allow for making (nontrivial) predictions on new data. Depending on the data and the intended outcome of the overall data mining process, two main goals can be distinguished. Regression is predicting a numeric quantity, in classification, and the outcome to be predicted is a discrete class. A wide range of data mining techniques is available to serve these goals, each with its advantages and disadvantages. Some of the most active applications of data mining have been in marketing and sales, for example, customer relationship management or market basket analysis. In recent years, this technique has been widely used in bioinformatics, genetics, and medicine. The use of data mining techniques in manufacturing began in the 1990s [3-5] and is currently a field of growing interest. In the following, some typical data mining applications for manufacturing are presented. The current study aims to show how a Corrective Trip Prediction helps to control the train decisions in the Tehran control centre. One of the important tasks of the Tehran Metro Traffic Control Centre is to decide on non-scheduled conditions. For example, in the event of an emergency, scheduled trips will be impaired, and in such cases, the control centre should avoid interruptions to the timetable with corrective actions. Therefore, the Traffic Control Centre needs to provide a prediction of emergency failures that can cause unplanned changes in scheduled trips. Thus, the problem we are looking for in this research is to obtain a forecasting model of the trip type in three categories of incomplete, missed, and delayed trips. For this purpose, recorded data of emergency failures are used in the delay record system. Briefly, this study focuses on forecasting the metro traffic in case of emergencies and models of collecting traffic data by distributed agents. We believe that Industry 4.0 and Smart Technologies are revolutionizing the transportation industry, and in particular, metro transport.

2. Literature review

There is a steady growing pressure on companies, urged by worldwide competition, to streamline operations involving product and product-related manufacturing system design, product manufacturing, and system maintenance [1]. As markets become more dynamic, there is more need to introduce concepts of flexibility and agility, forcing companies to deliver customized products and react promptly to fluctuating customer demands. E-collaboration and collaborative systems have allowed geographically dispersed teams to work together by supporting coordination and cooperation [2]. Companies are motivated to join collaborative networks to reach a competitive level of performance in terms of productivity, product quality, and system availability. Many companies have developed or are developing e-maintenance systems to provide better services and meet customers' needs, such as i) service suppliers, ii) service users, and iii) maintenance activities [3]. Maintenance activities are usually performed by integration of maintenance and process engineering functions at the phase of selection and application of machines and equipment, and also through proactive actions on those machines and equipment involving preventive and predictive maintenance [4]. In literature, it is possible to find three generic types of maintenance [5,6]

• The first type is corrective maintenance, consisting of repair actions when equipment or machine fails. The equipment is in action until its failure moment, when it will be repaired or replaced. The main disadvantages of this approach include fluctuant and unpredictable production, high levels of non-conforming products and scraps, and high levels of maintenance interventions motivated by catastrophic failures [7];

• Preventive maintenance, characterized by periodic maintenance operations to avoid equipment failures or machinery breakdowns, is determined through optimal preventive maintenance scheduling using a wide range of models that describe the degrading process of equipment, cost structure, and admissible maintenance actions [8];

• Predictive maintenance, which uses some parameters measured in the equipment to "feel" when a breakdown is imminent, intends to make interventions on machinery before harmful events may occur.

The requirement to respond to customer demand leads to a lot of pressure on the maintenance systems of factories. Therefore, industries should always keep their machines in working condition [9]. Nowadays, the amount of data generated and stored during industrial activities exceeds the capacity to analyse them without the use of automated analysis techniques. Due to such an increase in information, data processing has become more difficult and complex using traditional methods [10]. Conventional data analysis tools have limited capacity to detect patterns and discover the existing knowledge in data because they only use statistical methods [11]. Hence, there was a need to create a new generation of computational tools and techniques to assist humans in extracting useful information from data, or in other words, knowledge. Thus, the area of Knowledge Discovery in Databases (KDD) emerged in the late 80s, using models and data mining techniques to extract useful knowledge, patterns, and tendencies previously unknown in an autonomous and semiautomatic way [12]. The application domain of data mining and its related techniques, methodologies, and technologies have been greatly expanded in the last few years. The development of automated data collection tools and the tremendous data explosion, the urgent need for interpretation and exploitation of massive data volumes, along with the existence of supporting tools, have resulted in the development and flourishing of sophisticated Data Mining methodologies. Since Data Mining systems are comprised of several discrete but dependent tasks, they can be thought of as collaboration networks, yet autonomous units, which regulate, control, and organize all distributed activities involved in data cleaning, data transformation and reduction, algorithm application, and results in evaluation [13]. The research literature on intelligent agent system architectures has proven that such problems that require the synergy of several distributed elements for their solution can be efficiently implemented as a multi-agent system [14]. A multi-agent system consists of a group of intelligent agents that can take specific roles within an environment to cooperate with other agents [15].

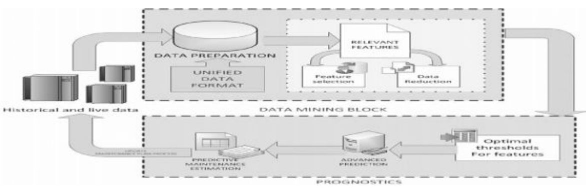

Figure 1: Data mining steps with maintenance data

Briefly, this process of Knowledge Discovery in Databases consists of a sequence of the following steps [16].

[1] Data cleaning – to remove noise and irrelevant data;

[2] Data integration – where multiple data sources are combined;

[3] Data selection – for retrieving only the relevant data from the database for the analysis;

[4] Data transformation – where data are transformed or consolidated into appropriate forms for mining;

[5] Data mining – the phase where the algorithms are applied to extract data patterns;

[6] Pattern evaluation – to find the interesting patterns which represent new knowledge; and

[7] Knowledge presentation – when visualization techniques are used to present the mined knowledge to the user.

Data mining [17] is another approach which is effective in the online prediction of train position and is a basic requirement for efficient route adjustment, traffic control, rescheduling, and passenger information. In practice, only the last measured train delays are known in the traffic control centres, and dispatchers must predict the arrival times of trains using only experience without adequate computer support. This often results in the simple extrapolation of the current delays for the expected arrival delays. Some railways use a linear shift of the timetable to extrapolate the current delays to the future. This method neglects that some trains may (partially) recover from a delay using running time supplements, while others may be (more) delayed due to route conflicts. Better predictions could be obtained by microscopic simulation models, but excessive computation times still prohibit the online application of microscopic simulation to densely operated large-scale networks. Blocking time theory provides the logic for building the process model from the log file. Signal passages are events that initiate processes such as blocking a part of the infrastructure and running over a block. Each complete train run can, thus, be represented as a graph built online by sweeping through the file. Moreover, route conflicts can be identified simultaneously by determining the time difference between relevant events and verifying if the train separation principles are respected. Given the large raw data log files, it is necessary to build an algorithm that sweeps through the file and visits every line only once, thus avoiding long computation times. An object-oriented approach is used to store the relevant data from the log files in infrastructure and train number objects, enabling the algorithm to revisit the objects and use and update the information therein. Figure 2 shows the attributes of each block and train object. All objects are created and updated on the fly while sweeping through a raw data log file. Static attributes in block objects ('Start signal', 'End signal', and 'Sections') are fixed when objects are created, using additional infrastructure files. They provide a unique description of a block or an interlocking route with the start and end signals and comprise track sections (Table 1).

Table 1: Objects and their attributes

Block	Train
Start Signal	Number
End Signal	Timetable
Section	Section (name, number, …)
Train (…)	Signal (name, number, …)

Dynamic attribute 'Trains' in a block object contains a chronologically sorted list of trains that traversed the block. Information about the block occupation time, release time, and train series is stored for each train. A Train object is defined by a train number and the list of traversed sections and signals, updated with every message from the log file related to the train. A list of scheduled departure and arrival times at each station is given in the attribute 'Timetable', which is an essential feature of the prediction model to capture the interactions of trains and the resulting conflicts and knock-on delays. As a result, the partitioning of the train movement data and the state of the control equipment of the movement path is of particular importance. The process mining algorithm, described by Kecman & Goverde [18], uses blocking time theory as the underlying process model and can, thus, identify all route conflicts. Released hindered running times are filtered out and excluded from further analysis.

2.1. An online railway traffic prediction tool

Figure 2 depicts the main components of the tool and the flow of data between them. The online prediction tool is based on a timed event graph with dynamic arc weights. The graph topology is built and updated considering the actual timetable, route and connection plan, and current positions of trains on the network. The route plan for a train is given as a planned sequence of block sections in the train route. Using the 'Sections' attribute of the block objects (Figure 2), a route plan can be translated to the level of track sections and used to determine the necessary headway arcs for routes with common track sections. The graph topology is updated by each change of the actual plans or new information from the real-time operations, i.e., changing the relative order of trains, adding or cancelling trains, modifying train routes, updating connections, and removing past events.

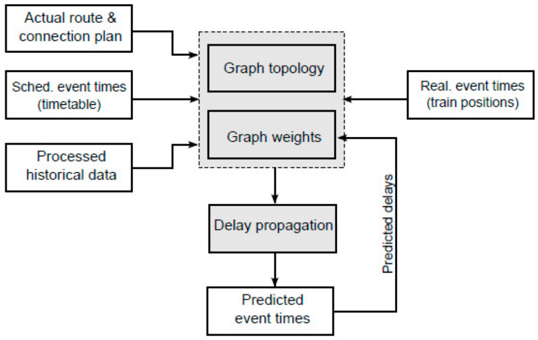

Figure 2: Components of the online prediction tool

Arc weights represent the minimum process times and are computed based on the current train positions and delays and processed historical data. The weight of an arc is time-dependent and assigned dynamically depending on the (estimated) starting time of the modelled process. Hence, the dependence of running and dwell times on current (predicted) delays are incorporated into the model. A prediction of event times of all reachable events is performed after every graph update (Figure 3). In the following subsections, the three main components of the tool (shaded boxes in Figure 4) and the input data will be explained in detail.

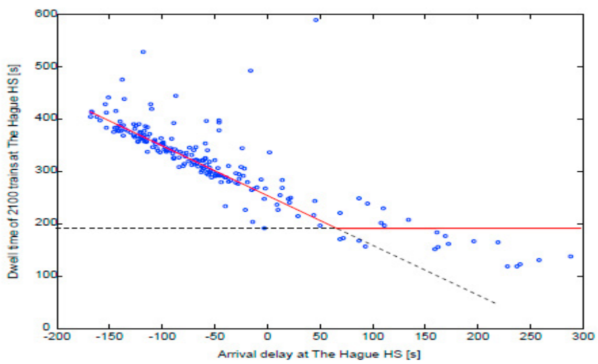

Figure 3: Dependence of dwell times in Den Hague HS on arrival delay

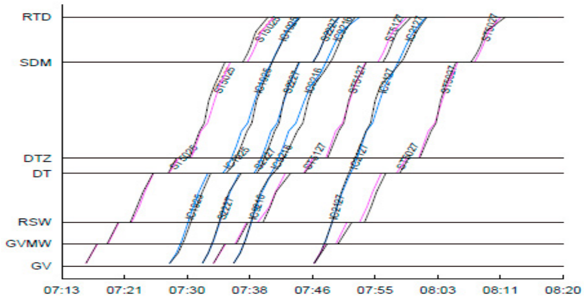

Figure 4: Time-distance diagram of predicted (at 7:13) and released train paths

The predictive model provides adequate decision support to signallers and traffic control and contributes to better utilization of railway infrastructure, improved reliability of train services, and more reliable and dynamic passenger information. The developed model will be embedded in a closed-loop model-predictive railway traffic control framework where online optimization algorithms will automatically resolve detected conflicts and propose control decisions to traffic controllers together with the predicted conflicts [19]. This way, an intelligent railway traffic management system will be obtained that proactively monitors the railway traffic and supports traffic controllers with decisions that optimize the traffic on a network level, beyond the traditional local control areas.

3. Data structure and data clean

For text analysis, a single data file with all the relevant text in a single field is required. The data construction involved deriving a combined text column from the relevant text fields in the source data. Input data was split into training and validation sets using simple random partitions since the distribution of the target variable was not considered skewed enough to require a stratified sample. Data volumes in each group were selected by balancing the training sets needed to have sufficient data to create a reliable model against the validation sets. Also, enough samples are required to provide a practical estimation of the model's performance. The final split was a training set containing 75% of the data and a validation data set with the test 25% records. The raw data entry in the metro database is shown in Table 2.

Table 2: Sample Database Structure

ID	Type of disorder	Direction	Dispatch time	Failure code	Delay in dispatch	Delay in receipt	Nonstop at station	Backward	ATP off
IT	Incomplete trip	A-B	07:50	210	0	0	0	0	0
DT	Delay trip	B-A	07:56	410	0	8	0	0	0
IT	Incomplete trip	B-A	08:06	120	0	0	0	0	0
DT	Delay trip	A-B	08:12	410	1	12	1	0	0
CT	Cancel trip	B-A	08:16	210	0	0	0	0	0
CT	Cancel trip	A-B	08:20	750	0	0	0	0	0

* IT: Incomplete trip - DT: Delayed trip - CT: Cancel trip

For the implementation of the data mining project, the data recorded during 2016-2021 were used. It is necessary to explain that emergency failures are reported to the control centre and metro command at each occurrence, and there is a record for each in the data bank, which provides the data table in Figure 1 for the 4 years. The characteristics examined in the records are as follows:

1. Type of trip contains: Incomplete trip - Delayed trip - Cancel the trip
2. Direction contains: A-B & B-A
3. Train number
4. Dispatch time: Scheduled time
5. Failure code: Emergency breakdown code
6. Delay in dispatch

7. Delay in receipt

8. Nonstop at stations

9. Backward

10. ATP-OFF: Automatic train protection

11. Connect two trains

Operator misspellings were often seen in logged error reports. A common error in the input data was words where spaces should have been used to separate the words, for example, "fixed" where the words "fixed" should have been typed. Therefore, data mining techniques, such as predicting missing data and correcting outliers, etc., were used to clear the database. While the task required user guidance, fixing the spelling errors was completed quickly and accurately without requiring other external tools. Generally, the raw data should be refined before they enter into the Rapid Miner software, and the following modelling should be checked:

1. Identify and Define Missing Data;

2. Identify and reduce noise data; and

3. Identify and reduce unrealistic data.

4. Model structure

The table shown in Fig.5 consists of six main modules in rapid miner

i. Data: Refined data input model.

ii. Set attributes: Attributes to be selected for entry.

iii. Set role: The attribute role that is expected to be predicted is label determined.

iv. Decision Tree: The first output model is the decision tree.

v. Naïve Bayes: The second output model is the probability prediction model based on the byes rules.

vi. Rule Induction: The third output model is rules induction.

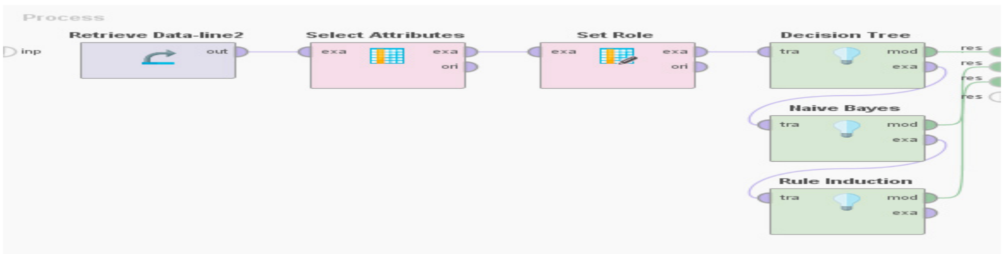

Fig. 5: Model Structure

4.1. Describe the model components:

4.1.1. Attributes:

Eleven attributes are selected in columns for the recorded data, as shown in Fig.5

4.1.2. Set Role:

Among the attributes defined for data, the type of the trip column is labelled for prediction.

4.1.3. Decision Tree:

The first output of the model is the type of trip to predict in the event of emergency failure of the decision tree, as shown in Fig. 6.

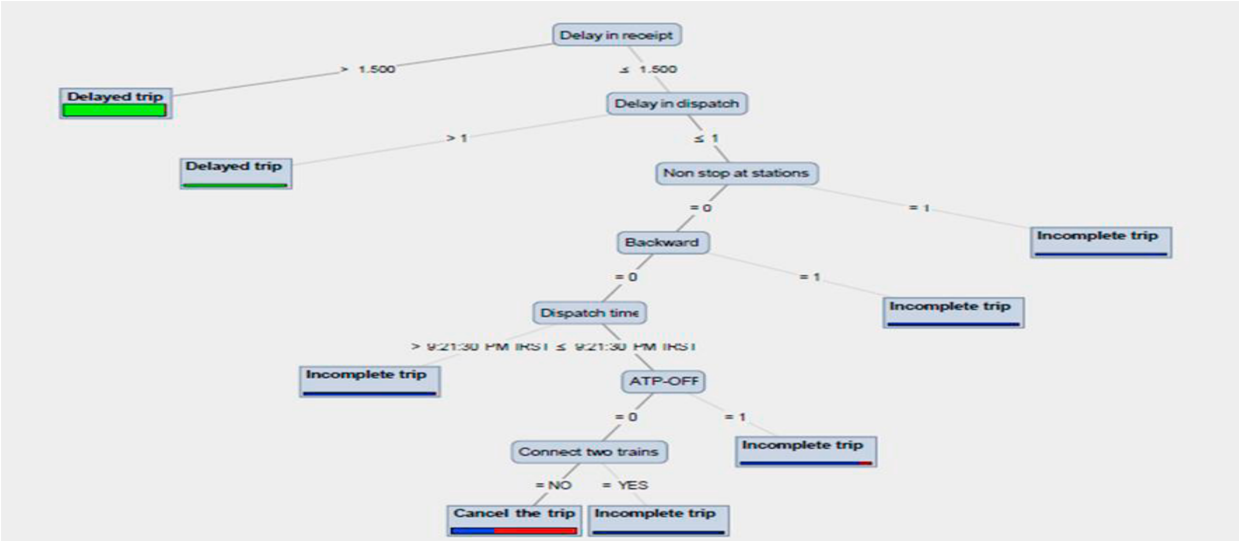

Fig. 6: Decision Tree

Table3: Decision Tree parameters & Model settings

criterion	Maximal depth	Confidence	Minimal leaf size	Minimal size for split	Number of pre-printing	Minimal gain
Gain-ratio	20	0.5	2	4	3	0.1

4.1.4. Interpretation of decision tree

The decision tree has been proposed based on the delays in arrival times and critical items of emergency failure and is designed for the metro line, which is headway for 4 minutes. The first level of the tree is divided into a delay of 1.5 minutes, and the final level is predicted as the type of trip. Needless to say, the numbers shown on the decision tree are of the type of time and in minutes.

5. Model Validation

We use the performance evaluation function operators in the Rapid Miner software to validate and evaluate the efficiency of the proposed model, as shown in Fig. 4. Also, in this section, we divide 70% of the test data and 30% of the input training data and teach the model using the rule of law. We can determine and compare the validity of decision tree models and associative laws separately in each step. In the form of four examples, the validation tree of the decision tree is shown in Fig 7.

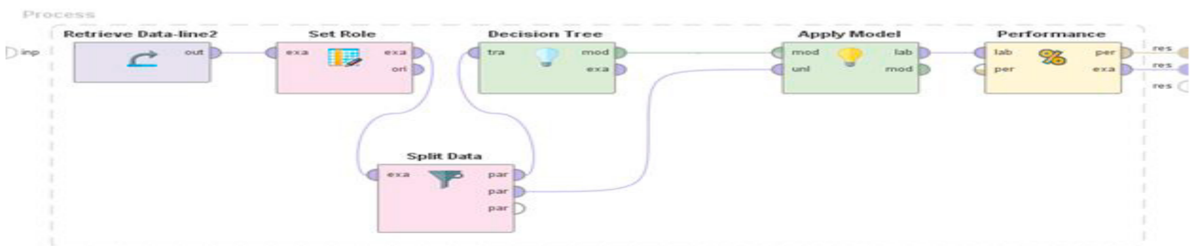

Fig 7: Schematic view of model validation

5.1. Description of the decision tree validation model

Rapid Miner software was used to predict the type of trip and validate the model through the following steps, as shown in Fig. 4:

1. Enter refined data into the validation model;
2. Specify the type of trip label between attributes;
3. Split Data based on the rule of 75% test data and 25% test data;
4. Apply the proposed decision tree to the Rapid Miner;
5. Use the pickup operator; and
6. Use the performance operator at the end.

5.2. Output validation model:

The outputs of the model after the training and testing are as follows*:

Table4: The decision tree model validation table with the efficiency index

accuracy: 90.48%

	true Incomplete trip	true Delayed trip	true Cancel the trip	class precision
pred. Incomplete trip	10	0	1	90.91%
pred. Delayed trip	0	221	0	100.00%
pred. Cancel the trip	29	0	54	65.06%
class recall	25.64%	100.00%	98.18%	

In this table, according to the knowledge extracted from the data, the future state of the trip is predicted with every failure.

Table5: The decision tree model validation table with the kappa index

kappa: 0.792

	true Incomplete trip	true Delayed trip	true Cancel the trip	class precision
pred. Incomplete trip	10	0	1	90.91%
pred. Delayed trip	0	221	0	100.00%
pred. Cancel the trip	29	0	54	65.06%
class recall	25.64%	100.00%	98.18%	

Table6: Split data table

ExampleSet (315 examples, 5 special attributes, 10 regular attributes) Filter (315 / 315 examples): all

Row No.	Type of disorder	prediction(Type of disorder)	confidence(Incomplete trip)	confidence(Delayed trip)	confidence(Cancel the trip)
2	Incomplete trip	Incomplete trip	1	0	0
3	Delayed trip	Delayed trip	0	1	0
4	Cancel the trip	Cancel the trip	0.346	0	0.654
5	Cancel the trip	Cancel the trip	0.346	0	0.654
6	Delayed trip	Delayed trip	0	1	0
7	Delayed trip	Delayed trip	0	1	0
8	Incomplete trip	Incomplete trip	0.900	0	0.100
9	Incomplete trip	Cancel the trip	0.346	0	0.654
10	Delayed trip	Delayed trip	0	1	0
11	Delayed trip	Delayed trip	0	1	0
12	Incomplete trip	Cancel the trip	0.346	0	0.654
13	Cancel the trip	Cancel the trip	0.346	0	0.654

* Due to the fact that the tables were taken directly from the software without changes, it is displayed as a screenshot.

6. Results & Conclusion

According to the investigations, the decision-makers in the metro traffic control centre can predict the travel situation and decide how to deal with it according to the knowledge obtained from the past records if they encounter any record of failure of the fleet, rail, signalling, etc. Therefore, the presented model will be a decision-making software for managers of the metro traffic control center. In summary, the results of the research are as follows:

(1) As an expert system, the proposed model can assist the control centre of the rail transport system supervisor in dealing with emergency failures in decision-making and forecasting of the trip type. (2) The proposed model can predict the type of the trip in the urban transport system in case of system errors and emergency failures with 90.48% accuracy. (3) The Kappa index for the proposed model is 0.792.

Therefore, the proposed model makes it possible to predict three types of incomplete traffic, delayed trips, and cancellation of trips due to the occurrence of emergency failure at the time of the operation of the Tehran Metro and then proceed to decide on dealing with unwanted situations. This research has been carried out for the Tehran Metro.

REFERENCES

[1] Sluga A., Butala P., and Peklenik J., A Conceptual Framework for Collaborative Design and Operations of Manufacturing Work Systems. CIRP Annals Manufacturing Technology, 2005. 54(1): p. 437-440.

[2] Wiendahl H.P. and Lutz S., Production in networks. in CIRP. 2002. [3] Persona A., Regattieri A., Pham H., and Battini D., Remote control and maintenance outsourcing networks and its applications in supply chain management. Journal of Operations Management, 2007. 25: p.

[4] Palmer R., Maintenance planning and scheduling handbook. 1999, McGraw Hill: New Jersey.

[5] Chu C., Proth J.M., and Wolff P., Predictive maintenance: The one-unit replacement model. Int. J. Production Economics, 1998: p. 285-295.

[6] Pinjala S.K., Pintelona L., and Vereecke A., An empirical investigation on the relationship between business and maintenance strategies. Int. J. Production Economics, 2006: p. 214-229.

[7] Swanson L., Linking maintenance strategies to performance. Int. J. Production Economics, 2001: p. 237-244.

[8] Yao X., Fernandez-Gaucherand E., Fu M.C., and Marcus S.I., Optimal preventive maintenance scheduling in semiconductor manufacturing IEEE Transactions on Semiconductor Manufacturing, 2004. 17(3): p. 345 - 356

[9] Bansal D., Evans D.J., and Jones B., BJEST: A reverse algorithm for the real time predictive maintenance system. International Journal of Machine tools and manufacture, 2006. 46: p. 1068-1078.

[10] Goebel M. and Gruenwald L., A survey of Data Mining and Knowledge Discovery Software Tools, in SIGKDD. 1999. p. 20-33.

[11] Michalsky R.S., Bratko I., and Miroslav K., Machine Learning and Data Mining Methods and Aplications. 1998, England: John Wiley

[12] Apte C., Liu B., Pednault E.P.D., and Smyth P., Business Applications of Data Mining. ACM, 2002. 45(8): p. 49-53.

[13] Symeonidis A.L., Chatzidimitriou K.C., Athanasiadis I.N., and Mitkas P.A., Data mining for agent reasoning: A synergy for training intelligent agents. Engineering Applications of Artificial Intelligence, 2007. 20: p. 1097–1111.

[14] Ferber J., Multi-agent systems: An introduction to distributed artificial intelligence. 1999, London: Addison-Wesley.

[15] Koulinitch A.S. and Sheremetov L.B., Coordination and communication issues in multi-agent expert system: concurrent configuration design advisor. Expert Systems with Applications, 1998. 15: p. 295–307

[16] Wirth and Hipp, R. "Towards a Standard Process Model for Data Mining". In Proceeding of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Pages 29-39, Nanchester, UK (2000)

[17] Pavle Kecman, Rob M.P. Goverde, An online railway traffic prediction model, Pages 5-9,2014

[18] Kecman, P. and Goverde, R.M.P. Process mining of train describer event data and automatic conflict identification. In Brebbia, C.A., Tomii, N., and Mera, J.M. (eds.), Computers in Railways XIII, WIT Transactions on The Built Environment, vol. 127, pp. 227–238, Southampton, 2012.

[19] Kecman, P., Goverde, R.M.P., and van den Boom, T.J.J. A model-predictive control
Framework for railway traffic management. In Proceedings of the 4th International Seminar on Railway Operations Modelling and Analysis (RailRome2011), pp. 1–15, Rome.