ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No 4, 287-295 (2020)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Optimization of the Procedures for Checking the Functionality of the Greek Railways: Data Mining and Machine Learning Approach to Predict Passenger Train Immobilization

Ilias Kalathas*, Michail Papoutsidakis, Chistos Drosos

Department of Industrial Design and Production Engineering, University of West Attica, Athens, 15354, Greece

A R T I C L E I N F O

Article history:
Received: 17 June, 2020
Accepted: 19 July, 2020
Online: 28 July, 2020

Keywords:
Machine learning
Railway
Train Immobilization
Data mining
Predictive Model
Malfunctions
Diagnosis

A B S T R A C T

Information is the driving force of businesses because it can ensure the ability of knowledge and prediction. The railway industry produces a huge ume of data, with the proper processing of them and the use of innovative technology, there is the possibility of beneficial information to be provided which constitute the deciding factor for the correct decision making. Safety is the railway comparative advantage that has to be reinforced by each business administration while making the optimum decisions. The main purpose of this paper is the investigation of the most important dysfunctions that arise in a train and can cause its immobilization at the main passenger rail, resulting in huge delays of conducting the routes setting the passengers at risk. Afterwards the total of malfunctions is assessed and the most important, potentially, malfunction is assessed, so as the executives of the Greek Railway company to plan and redefine the processes and the initial plan of the predictive maintenance. This paper demonstrates the effort of implementing innovative applications by making use of methods from the rapidly developed field of Data Mining to the Greek Railway Company that uses obsolete procedures for the control of the trains' functionality in order to investigate the data for the provision of specialized information which will be used as a tool for the faster, more accurate and precise decision making. This decision making approach is based on a specific algorithm's design in order to automatically detect faults and make periodic maintenance of trains easier. Holistic approach is performed in the management of real data from the Greek railway industry and a predictive model of Machine Learning is developed, for the optimization of the management's performance of the trains reinforcing the strategic target of the railway industry which is the transportation of citizens with safety and comfort.

1. Introduction

Businesses are more and more using sets of data in order to conduct their decisions. Developments in the field of Data Mining and the Machine Learning are expected to predominate in 2020 and to create, within the next decade, significant opportunities to all the companies [1]. The emerging technologies change the way that businesses collect and extract useful information from the data [2]. The Data Mining is an effective method for the analysis of a huge amount of collected data that extracts useful information. The Machine Learning is a field that was developed by the artificial intelligence and assists planning and development of algorithms and the eution of the performance related to empirical-operational data [3]. The railway industries are a field the performance of which progressively depends on their ability to extract information from complex sets of data and take the optimum action in real time [4]. The Data mining in conjunction with the Machine Learning has the capacity to improve the operational progress raising the level of efficiency in decision making and the overall procedures [5].

The railway is a system of passengers' and merchandise transportation with wheeled vehicles (trains) that roll on rails. The railways as a mean of transport is defined by three components, the functioning utilization, the infrastructure, and the rolling stock. With the term rolling stock we refer to all kinds of vehicles pulled or driven on rails that perform railway transports [6]. The Railway Rolling Stock is a complicated system the smooth operation of which plays a leading role on the exploitation of the rail system and is subject to progressive lesions (wears, erosions, malfunctions

*Corresponding Author: Ilias Kalathas, University of West Attica,
Tel: +306974731434 , i.kalathas@uniwa.gr

etc.) and therefor it is obliged to accept interventions of restoration – maintenance in order to ensure reliability, availability and security of its circulation [7]. The high quality of maintenance services of the rolling stock is an essential precondition for the smooth function of the whole railway system. The predictive models of machine learning are able to provide the management of railway industries with the appropriate information that will assist on making decisions for the restart of the process restoration - maintenance of rolling stock aiming to the increase of its reliability- effectiveness as well as the minimization of hazards and cost [8].

2. Problem Statement

For the smooth and orderly operation of the railway network necessary requirement is to avoid train immobilization on the main passenger track which can cause a traffic stop and consequently great delays in itineraries conduct and naturally question the most important goal, which is the passengers' safety [9] [10].

In this research work it is presented a method in order to optimize the management of the Greek Railway Company STA.SY. S.A. The prediction, diagnosis and the dealing with malfunctions that can immobilize a train in the main passenger track is the purpose of the application. It is trying to discover – detect which malfunction or malfunctions combined are capable of immobilizing a train. The effort of this research focuses on the way the application of Data Mining for the construction of a predictive model of Machine Learning which will be used as a tool for the output of useful results to the management in order to ensure faster and right decisions. The model of machine learning will meaningfully assist in the infrastructure and the improvement of the maintenance of trains, in the effective detection of malfunctions and the simplification of the programming relevant to the train traffic.

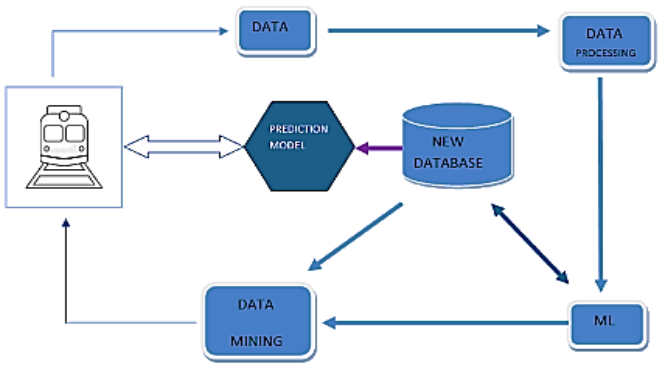

Figure 1: Predictive Malfunctions Solution

3. Theoretical background

Machine learning is a branch of artificial intelligence that covers the study or the construction of systems able to "learn" from a multitude of data. Machine learning comprises a field of information technology and belongs to the wider sector of artificial intelligence [11]. Algorithms of machine learning can learn from data and make predictions for new data that have not been dealt with.

3.2 Supervised learning

Supervised learning is a section of Machine learning and it is defined as the process of an algorithm constructing a function that depicts given inputs to known desired outputs, with ultimate aim to apply a model of prediction to new data that have not been seen before. It is used with problems of classification, prediction, and interpretation [12]. The creation of a model of prediction, which is necessary for the construction of a mechanism in order to make decisions or calculations, is achieved with supervised learning.

3.3 Data mining

The data mining is one of the stages of the process Knowledge Discovery in Databases – KDD. The assignments that are carried out at this stage incorporate algorithm applications for the construction of algorithmic models emphasizing on the discipline of Machine Learning [13]. Data mining is defined as the process of constructing a categorization model or prediction, with the use of algorithms aiming at the description and the prediction of data set for the production of exportable potentially useful knowledge.

3.4 Classification

Classification is one of the basic techniques of Mining Data, considered to be the most popular process because of the abundance of the effective applications. It is about a predictive technique aiming at the construction of a predictive model applying classification algorithms to present data and it is more easily translated than compared with the models which produce other techniques of data mining [14]. Classification is defined as the process of evaluating the most probable value a dependent variable will take, based on the already known data from the independent value of a known data set in order to predict a nominal discrete value. The classification algorithms are divided in five methods, Classification with Decision Trees, Classification with Neural Networks, Classification with Bayes Criteria, Classification with Support Vector Machines, and Classification with k – Nearest Neighbors [15].

3.5 Decision Trees.

The Decision Trees is one of the most widely known method of classification. It is one of the techniques of Machine Learning the construction of which is aiming at, the sequential split of a set of observations into subsets. It is easily apprehended by people because it applies simple representations for the classification of examples, and it is often used in the mining data sector [16]. The basic purpose of the Decision Trees is the prediction of the variable value that is trained by data created by other independent variables [17].

- Algorithm ID3 There are various algorithms for the data classification and the forecasting using the Decision Tree method, but the Iterative Dichotomies or ID3 has predominated along with its subsequent developments, the C4.5 and its commercial version C5.0 [18]. Its main characteristic is that it presupposes the existence of nominal fields only creating a tree with multiple courses (top down) that uses nodes with categorical variables which attract great sets of data at the beginning in order to be minimized during their courses.

- Algorithm C4.5 The algorithm C4.5 has been created by the ID3 which is its eution. Their main difference is the criterion of data separation that is called Gain Ratio and it is defined by the equation.

$$\text{GainRatio(S,A)}= \frac{\text{information(S,A)}}{\text{entropy(s,a)}} \quad (1)$$

The Gain Ration sets the rules, it regularizes the information gain as to the entropy. The Gain Ration improves the precision and minimizes the complexity of the trees [19]. An additional important difference between the two algorithms is that the ID3 produces a tree with multiple courses that uses nodes the categorical variables whereas the C4.5 can use a field of number signs setting border lines in order to split the data whether they are above the limit , under the limit or on the limit.

3.6 Weka

The WEKA is a software of Machine Learning of open source, its name comes from the initials Waikato Environment for Knowledge Analysis and it is used for Mining Data. It was developed at the University of Waikato in New Zealand and its development goes on from an international team of programmers [20]. It is publicly available according to the license terms GNU General Public License, which allows users its execution but also the making of free changes in the software and its wide acceptance [21]. WEKA provides graphical user interfaces which allow its application to users which don't provide knowledge of programming and it is capable of installing in all modern software platforms because it is written in Java programming language [22]. The software suit WEKA enables the preparation and visualization of the data and provides various techniques of knowledge extraction from data as the classification, clustering, association rules mining, and prediction [23]. The use of WEKA software is encountered in a great number of scientific papers and many books of mining Data, since the set of functions it offers in combination with the graphical reproduction of algorithms contributes to the analysis of data and the construction of predictive models.

4. Related work

In the present section of this work are briefly presented relevant researches and experiments regarding the techniques of machine learning, with the use or not of software WEKA aiming at data mining and prediction and they constitute the starting point for the development of a smart supporting tool of innovative processes. In the meanwhile, a comparison of machine learning models, which have been used in various transport sectors, takes place for the prevention and diagnosis of malfunctions in order to make the best decisions.

The article [24] refers to the detection and diagnosis of faults in parts of the rails of the railway network. It acknowledges that faults in railway tracks can cause catastrophic consequences to the passengers' safety. It presents a methodology for the in time detection of abnormalities in parts of the rails using sound data that have been collected by a rail vehicle model NS-AM Sehwa Company in Daejeon, South Korea, on the 1st of January 2016. Two different experiments take place (one for the total of data and one for the data characterized faulty) with the use of SVM algorithm at the WEKA software The results show that the system allows the effective detection and precise diagnosis of malfunctions that exceed 94.1 %. The proposed method provides reliable means of investigation of the railways for the comprehension of the railway conditions.

This research [25] refers to the use of Apriori algorithm and to clustering algorithm of WEKA tool that are applied to a set of data for road accidents at the area Alghat that belongs to the town Riyadh Province, in Saudi Arabia. The aim of this research is to find new approaches and new rules of connection between the set of movement data using the WEKA tool in order to discover new factors that cause road accidents. The study acknowledges that the software of machine learning WEKA is a very useful tool for mining data, which allows the user to choose from various algorithms and to compare them, in order to accurately reach the demanding results.

In addition, in the traffic accidents analysis based on the algorithm C4.5 using the WEKA software of machine learning refers to the article [26]. The study acknowledges that the data mining contributes to the prevention of accidents and the safety management of traffic. The research was held in China, in a part of Wenli motorway which is located in the central part of Zhejiang province connecting Lishui town with Wenzhou and it was based on the traffic accidents data from 2006 to 2013. Using a Decision Tree and applying the algorithm C4.5 to WEKA the effect of various factors which affect an accident was tested. The model that was created can effectively handle large sets of data and achieve total precision of prediction 80% having as a result to be considered a method capable of traffic control.

The article [27] refers to the investigation of the use of algorithms of machine learning in shipping. It assesses the smart predictive methods in two stroke low speed marine engine aiming to the effective tracking and classification of the faults/malfunctions displayed since the in time tracking of the malfunction secures the nonstop rhythm of the ship but also less consumption of fuels. The potential that the tool of mining data WEKA offers is used for the application of the optimum predictive tracking and prediction system of faults/failures. The experimental research was based on the collection and processing of engine data MAN BW 7S60MC using the engines' simulator of the Faculty of Engineers and Merchant Academy of Aspropirgos / Greece. The methods presented showed that they provide reliable diagnostic tools setting the method "AdaBoost", as the right choice with rate 96.5 %.

The reference [28] applies four techniques of machine learning for the automatic prediction of traffic maintenance using functions and historical records. Its purpose is to create a list of prioritizing tasks in order to avoid its future decline. The case study was held at the Portuguese motorway network of 539 km total length managed by the Infrastructures de Portugal Company and four techniques of machine learning DT, KNN, SVM and ANN were used for the optimum selection in making decisions. The results show that with enforced data base from historical facts a satisfying model of prediction is provided for the avoidance of adverse future conditions in the motorway network.

The article [29] refers to the prediction of airline routes delay with the use of machine learning. It recognized factors that are able to cause delay in flights (e.g. visibility, temperature, wind strength,

age and type of aircraft) the techniques of machine learning were applied (decision tree, k-means, J48, random forest and Bayesians classification) for the prediction of delay and the size of the incident. The tests wereheld on American flights data and on a big Iranian airlines network. With the methodology suggested the estimations for the flights' delay reaches more than 70% precision in the U.S.A and Iran.

The doctoral thesis [30] aims to the investigation of the possibilities of a ship to reduce fuel consumption of pollutant emissions based on the data from machine records. It presents results in order to save energy using tools of machine learning in real conditions. The data comes from machines' records of the cruise ship M/S Birka Stockholm in the Baltic Sea during four journeys between the years 2014 and 2015. Using algorithms and the method of supervised learning with the regression technique and the results showed that there is satisfying prediction ability of fuel consumption. The main conclusion of the doctoral thesis is that the data from a ship's tasks can be used for the observation of maintenance and performances but also to provide a solid base for the construction of a model that will improve energetically and functionally the performance of a ship.

5. Data set and Pre-Processing

The data of the present Case Study came from the informative system BAAN – RSD that the STA.SY S.A. uses and especially the Sector of the Rolling Stock Line 1 (former ISAP). The initial set of data with no processing contained 1000 instances and 24 attributes. The useful – exploitable data were derived from data bases of the company using SQL questions and were extracted in Excel spreadsheet format, which contained fundamental data – information (480 instances and 8 attributes) for the trainsets condition in a 6month period from 1/6/2019 to 31/12/19

Table 1: Attributes Considered

a/s	Attributes	Explanations of attributes
1	Number of trainsets	It includes coding of all the trainsets and aims at its coordinate observation and recording.
2	Date's coding	It includes the date that the trainsets were received and started their operation.
3	Operational system	It includes the operating levels of the trainset that are divided based on the equipment used.
4	Subsystem	It includes the subsystem of the equipment of the trainset with the same features.
5	Indications	It includes the probable problem of the trainset in an operative system.
6	Failure	It includes the problem that sets the trainset out of order.
7	Kilometres	It includes the kilometres travelled of a trainset from one maintenance to the other.
8	Restoration of the failure	It includes the way that the problem was dealt with using specialized alternatives.

The addition / improvement or delete / extraction of inappropriate information from the set of data, with appropriate criteria is the necessary precondition for the creation of a satisfying data base, appropriate to the research [31]. The appropriate criteria to be met are:

- The antiquity of the trainsets (dates arranged in year and months of the delivery of the trainsets.
- Malfunctions that set in dispute the most significant target of the railways which is the passenger's safety.
- The gravity of the malfunctions affects the traffic of the trainsets.
- The periodicity of the malfunctions that the trainsets face.

The process continued with the discretion of the data from Excel spreadsheet applying special filters that are based on the appropriate criteria. The failures that arise and are capable of causing malfunction to the traffic of the trainsets are:

- Main switch overcurrent cut-off (code: SWITCH)
- Door malfunction (code: DOOR)
- Generator failure (code: GENERATOR)
- Braking during course (code: BRAKING)

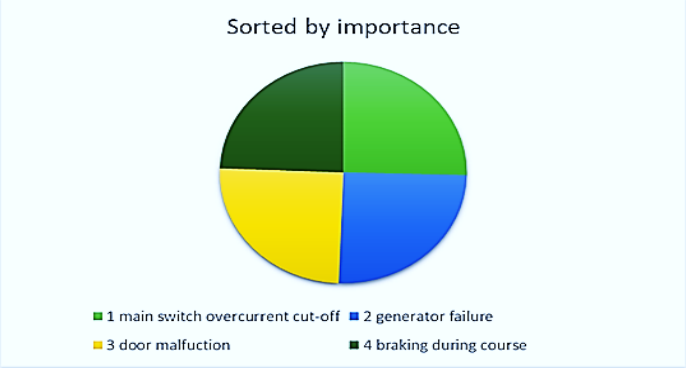

Figure 2: Results for malfunctions to the traffic of the trainsets

The braking during course (code: BRAKING) is divided in crucial (code: A) the cause of the failure is elicited from the malfunction of the engine of the vehicle and simple (code: B) the cause of the failure is elicited from the electronic display unit of the trainset.

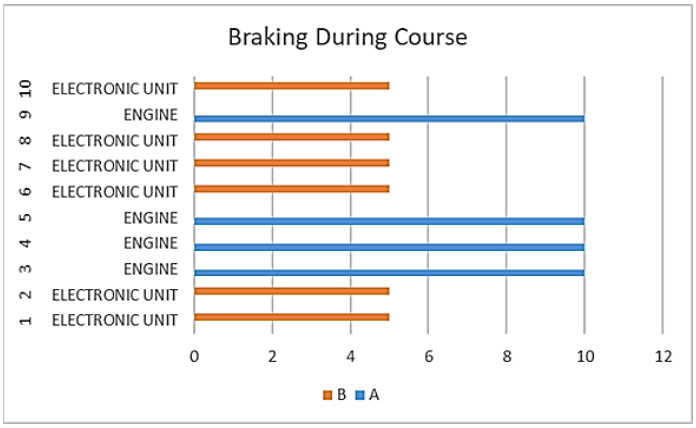

Figure 3: Results for malfunctions for braking during course

The number of repetitions, the importance of the failures that are able to cause malfunction to the traffic of the trainsets, during the 6month period of time, also defined the periodicity of the malfunctions.

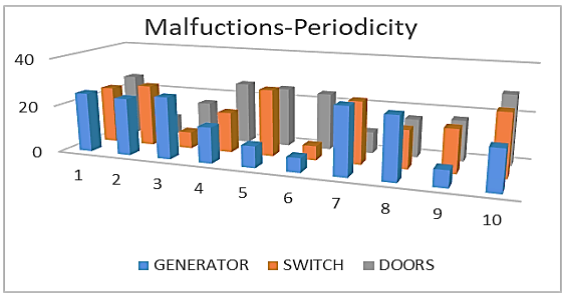

Figure 4: Results for Failure periodicity

- Frequent repetition:
  repetition from 20 to 30 times (code: FREQUENT)
- Moderate repetition:
  repetition from 10 to 19 times (code: MODERATE)
- Sparse repetition:
  repetition from 1 to 9 times (code: SPARSE)

Meanwhile, it is considered advisable to record that the Register of the trainsets referring to the date of delivery (antiquity) of the trainset.

Table 2: Receipt dates

Train register	Years	Months	Months	Months
8	1980	1 to 4	5 to 8	9 to 12
10	1992	1 to 4	5 to 8	9 to 12
11	2003	1 to 4	5 to 8	9 to 12

The result of discrimination in the division of the initial data in categories that can be easily apprehended and create a new base of data with malfunctions leading to immobilize a trainset on the main passengers' track.

TRAIN REGISTER	MALFUCTIONS- PERIODICITY				RESULT
TRAIN INVETERACY	GENERATOR	SWITCH	DOORS	BRAKING	PROPLEM-STANDSTILL TRAIN
8	FREQUENT	FREQUENT	FREQUENT	B	YES
8	FREQUENT	FREQUENT	SPARSE	B	YES
10	FREQUENT	SPARSE	MODERATE	A	NO
10	MODERATE	MODERATE	FREQUENT	A	YES
8	SPARSE	FREQUENT	FREQUENT	A	YES
10	SPARSE	SPARSE	FREQUENT	B	NO
11	FREQUENT	FREQUENT	SPARSE	B	NO
11	FREQUENT	MODERATE	MODERATE	B	YES
11	SPARSE	MODERATE	MODERATE	A	YES
11	MODERATE	FREQUENT	FREQUENT	B	NO

Figure 5: New database

The criteria for the selection of the appropriate method and the optimum techniques of machine learning – data mining is:

- The correctness / reliability of the data
- The amount of data
- The suitability of data
- The dynamic performance of data
- The constant update
- The readjustment (with the entry of new data)
- The satisfactory response to the research
- The definition of the objective purpose

In the scientific – investigative research in order to find a predictive model that will pinpoint which malfunction or malfunctions combined (generator failure, main switch overcurrent cut off, door malfunction and braking during course) are able to immobilize a trainset, the Categorizing method is advised (data mining technique). Therefore, the next step is the selection and application of the appropriate Categorizing method. In the present study case, the method applied for the creation of a predictive model and malfunction diagnosis that meets the appropriate criteria is the method of decision tree using the algorithm C4.5. The techniques above are provided by the software of machine learning WEKA.

6. Data Processing

The operation of WEKA software requires specific file formats. The WEKA software converts and saves as an ARFF file the data base that contains all the data that will be used [32].

```
@relation ENGLISH


@attribute TRAIN INVETERACY real
@attribute GENERATOR {SPARSE, FREQUENT, MODERATE}

@attribute SWITCH {SPARSE, FREQUENT, MODERATE}

@attribute DOORS {SPARSE, FREQUENT, MODERATE}

@attribute BRAKING {B,A}

@attribute PROBLEM-STANDSTILL {NO, YES}


@data

8,FREQUENT,FREQUENT,FREQUENT,B,YES
8,FREQUENT,FREQUENT,SPARSE,B,YES
10,FREQUENT,SPARSE,MODERATE,A,NO
10,MODERATE,MODERATE,FREQUENT,A,YES
8,SPARSE,FREQUENT,FREQUENT,A,YES
10,SPARSE,SPARSE,FREQUENT,B,NO
11,FREQUENT,FREQUENT,SPARSE,B,NO
11,FREQUENT,MODERATE,MODERATE,B,YES
11,SPARSE,MODERATE,MODERATE,A,YES
11,MODERATE,FREQUENT,FREQUENT,B,NO
```

Figure 6: ARFF file

Afterwards the immediate presentation of data takes place in the form of diagrams

- The first table GENERATOR presents in red the immobilization of a trainset after "generator failure" malfunction, whereas in blue when the "generator failure" malfunction does not result in immobilizing the trainset. They are additionally presented, in order, the periodicity malfunctions "SPARSE", "FREQUENT", "MODERATE".
- The second table SWITCH presents in red the immobilization of a trainset after "main switch overcurrent cut-off" malfunction does not result in immobilizing the trainset. They are additionally presented the choices of frequency malfunctions "SPARSE", "FREQUENT", and "MODERATE".
- The third table DOORS presents in red the immobilization of a trainset after a "door malfunction", whereas in blue when the "door malfunction" does not result in the immobilization of a train set. They are additionally presented, in order, the choices of frequency malfunctions "SPARSE", "FREQUENT", and "MODERATE".

- The fourth table BRAKING presents overall immobilizations of trainsets, in red the immobilization of a trainset after the "braking during course" malfunction, and in blue when the "braking during course" malfunction does not result in immobilizing the trainset. It is additionally presented in red the significance of the malfunction "A" or "B".
- The fifth table presents the results of the immobilization or not of a trainset, in blue the non-immobilization of a trainset, whereas in red the immobilization of a trainset.

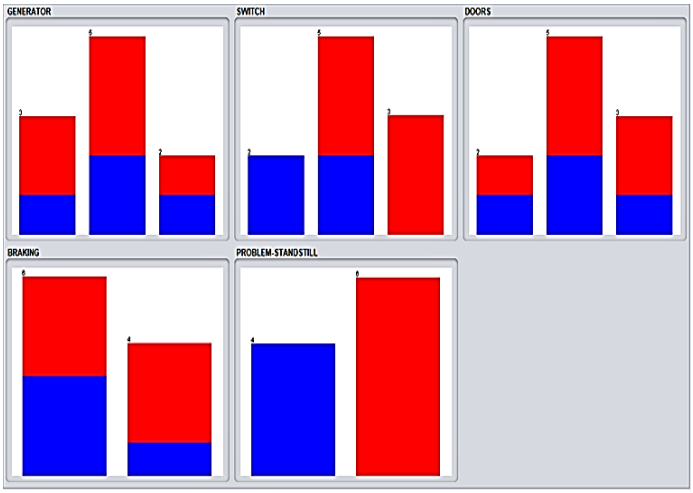

Figure 7: Data in the form of charts

The results after the completion of the algorithm are the following:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     ENGLISH-weka.filters.unsupervised.attribute.Remove-R1
Instances:    10
Attributes:   5
              GENERATOR
              SWITCH
              DOORS
              BRAKING
              PROBLEM-STANDSTILL
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

SWITCH = SPARSE: NO (2.0)
SWITCH = FREQUENT: YES (5.0/2.0)
SWITCH = MODERATE: YES (3.0)

Number of Leaves  :     3

Size of the tree :      4


Time taken to build model: 0 seconds
```

Figure 8: Run information of the algorithm

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances           8               80      %
Incorrectly Classified Instances         2               20      %
Kappa statistic                          0.5455
Mean absolute error                      0.24
Root mean squared error                  0.3464
Relative absolute error                 49.6552 %
Root relative squared error             70.6698 %
Total Number of Instances               10

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,500     0,000     1,000       0,500   0,667      0,612    0,875     0,786     NO
               1,000     0,500     0,750       1,000   0,857      0,612    0,875     0,875     YES
Weighted Avg.  0,800     0,300     0,850       0,800   0,781      0,612    0,875     0,839

=== Confusion Matrix ===

 a b   <-- classified as
 2 2 | a = NO
 0 6 | b = YES
```

Figure 9: The result of the algorithm

The WEKA software creates – visualizes the requested decision tree

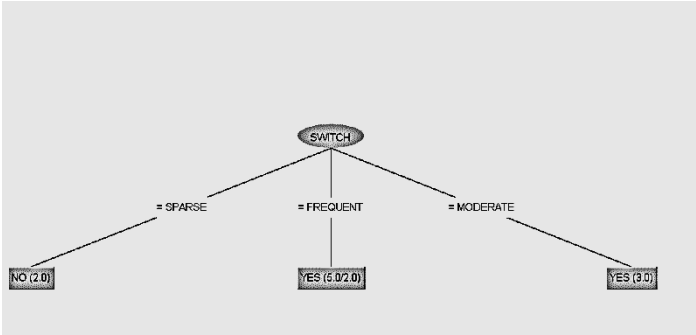

Figure 10: The output of C4.5 tree.

After the completion of the algorithm the conclusion that arises is that the "main switch" is the most important of malfunctions, which leads to the immobilization of a trainset. Setting the constant of antiquity to the algorithm an additional decision tree appears, whereas the root of the tree remains the same.

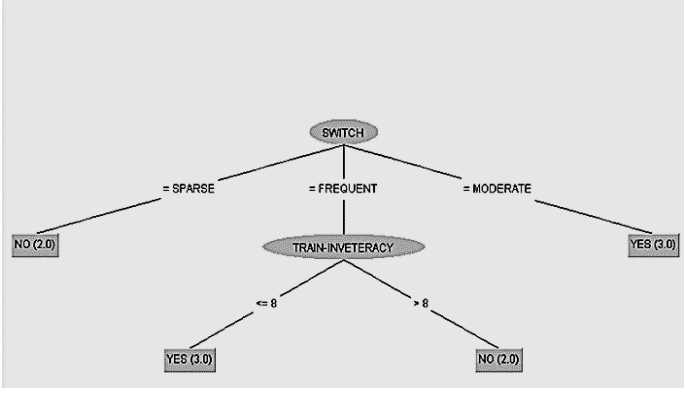

Figure 11: The output of C4.5 tree with variable 'train inveteracy'

From the final decision tree, the outcome is that the most important problem for the immobilization of a trainset is the

malfunction of the "main switch" and the antiquity of the trainset follows.

If another occasional option of the WEKA software is used, like the random tree it seems that the decision tree is different, which means that the software cannot decide which malfunction or combination of malfunctions will immobilize the trainset.

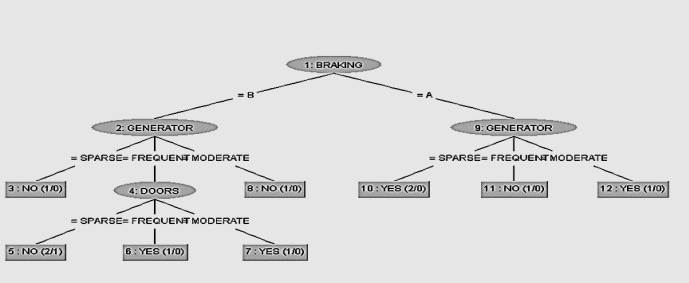

Figure 12: The output of random tree

7. Results and Discussion

For the precision of the characteristics (figure 12) used an assessment of data was held related to the periodic written reports of the maintenance technicians. The results presented totally 90 % success decreasing the danger of choice to minimum.

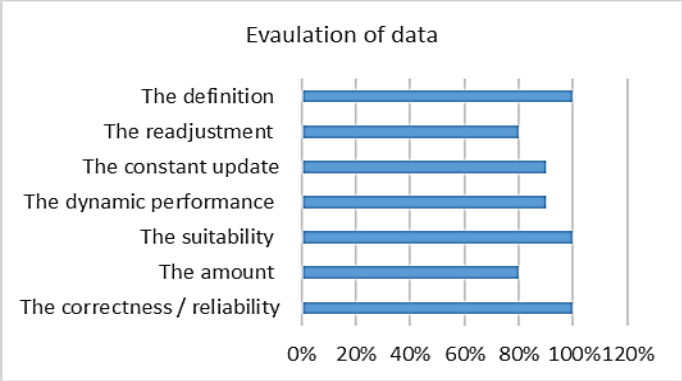

Figure 13: Success Rate Data

The algorithms assessment was held under specific criteria as the quantity, the adequacy, and the dynamic representation of data regarding the decision trees and the characteristic data of the clues.

The advantages of the C4.5 algorithm (figure 13) which are the speed and its independence from the classification model makes it the optimum choice. During the repeated procedure (5 times) the result appeared almost the same with an average of 80% success. [33]

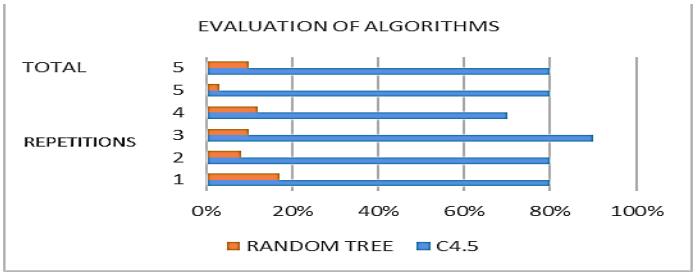

Figure 14: Success Rate Algorithms

During the pre-processing of data (figure 2) were presented the main malfunctions which affect the train movement and are capable of immobilizing the train at the main passenger rail. The malfunctions Main switch overcurrent cut-off (code: SWITCH), Door malfunction (code: DOOR), Generator failure (code: GENERATOR), Braking during course (code: BRAKING) appeared on every train in 25% each. The malfunction Braking during course (code: BRAKING) (figure 3) which is divided into crucial (code: A) and simple (code: B) appeared in a 6-month period as simple (code: B) 6 times and as crucial (code: A) 4 times. The periodicity of malfunctions (figure 4) in each train is not the same. Table 3 depicts the periodicity of each malfunction in every train.

Table 3: Periodicity-Malfunction

TRAIN	GENERATOR	SWITCH	DOOR
No 1	Repetition from 20 to 30 times	Repetition from 20 to 30 times	Repetition from 20 to 30 times
No 2	Repetition from 20 to 30 times	Repetition from 20 to 30 times	Repetition from 1 to 9 times
No 3	Repetition from 20 to 30 times	Repetition from 1 to 9 times	Repetition from 10 to 19 times
No 4	Repetition from 10 to 19 times	Repetition from 10 to 19 times	Repetition from 20 to 30 times
No 5	Repetition from 1 to 9 times	Repetition from 20 to 30 times	Repetition from 20 to 30 times
No 6	Repetition from 1 to 9 times	Repetition from 1 to 9 times	Repetition from 20 to 30 times
No 7	Repetition from 20 to 30 times	Repetition from 20 to 30 times	Repetition from 1 to 9 times
No 8	Repetition from 20 to 30 times	Repetition from 10 to 19 times	Repetition from 10 to 19 times
No 9	Repetition from 1 to 9 times	Repetition from 10 to 19 times	Repetition from 10 to 19 times
No 10	Repetition from 10 to 19 times	Repetition from 20 to 30 times	Repetition from 20 to 30 times

The results provide useful – necessary information, for the research conduction:

- The dysfunctions of the train can be pinpointed and immobilize the train are four: Main switch overcurrent cut-off (code: SWITCH), Door malfunction (code: DOOR), Generator failure (code: GENERATOR), Braking during course (code: BRAKING. The importance of each malfunction is the same with 25%.

- The malfunction Braking during course (code: BRAKING), is divided into crucial (code: A) and simple (code: B) and appears 10 times in total.
- The frequency of each malfunction Door malfunction, Generator failure, Main switch overcurrent cut-off in each train is different but equally important.

After the data analysis and processing with the use of WEKAsoftware (figure 10 and figure 11) arises:

- From the four main and equivalent malfunctions, the most important malfunction, in potential, on which the executives of a company must focus is the Main switch overcurrent cut-off (code: SWITCH) and after that the trains' antiquity which is not referred as a dysfunction but it exists as data.
- The prediction which takes place for the most important malfunction, potentially, which can immobilize a train in the main passenger rail, offers to the executives of the company the opportunity to focus on the functional systems of the train related to the main switch , rescheduling the plan of predictive maintenance.

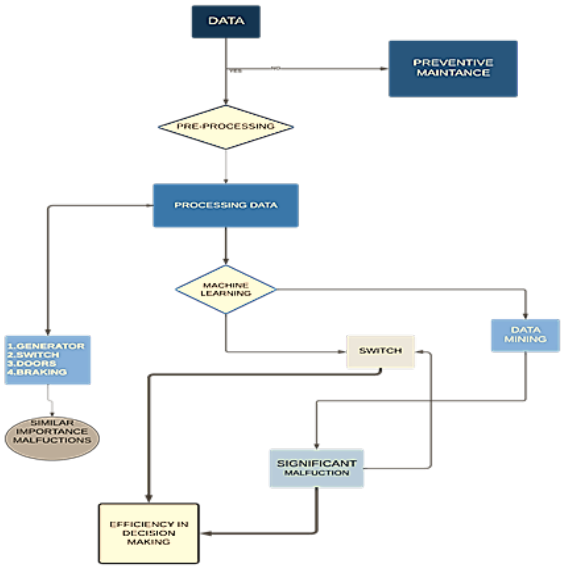

Figure 15:Flow chart diagram for Decision-making process

8. Conclusion

In the present paper the realization of the application of machine learning took place aiming to the prediction, diagnosis and dealing with malfunctions which immobilize a trainset on the main passengers' rail.

Detailed pre-processing of the data was carried out in the huge database provided by the Greek railway company STA.SY. S.A., resulting in the detection of malfunctions of equal importance (main device power failure, door failure, generator failure, braking during the course) capable of immobilizing the train on the main passenger rail.

Consecutively a new data base was created consisting of specialized - quality features, advantageous to the research like the trainsets' record, the significance, and the periodicity of malfunctions.

With the use of the Machine Learning software (WEKA) the most important factors which immobilize a trainset on the main passengers' rail were investigated and pinpointed. The results that arose from the method that was used are, that the malfunction "main switch overcurrent cut-off" and then the antiquity – year of the trainset, have the greatest significance on the immobilization of a trainset.

Additionally, the combination of simplicity and the clarity of the decision trees, with the use of strict criteria, for the accuracy, the quantity, the appropriacy, and the dynamic representation of the data, confirmed the accurate use of the C4.5 algorithm giving 80% result accuracy whereas the use of the random tree algorithm, with the same criteria, did not produce any quality results.

Finally, a new process was created which can successfully classify the malfunctions in order to make a precise and accurate prediction for the immobilization of the trains on the passengers' rail.

The suggested approach on the data analysis with the use of machine learning suggested to the STA.SY Company can develop a better method for the control of the trains' circulation.

The creation of a machine learning template for the prediction and mining data is capable of constituting the main tool for the improvement of the process of making decisions from the management, aim to the better programming and the effective management of the maintenance of trainsets, setting as dominant priority, the passengers' safety.

The new method of the malfunctions' classification with the use of innovative technologies, sets as main priority the passengers' safety eves the maintenance procedures and provides the company's executives with new knowledge to take the right decisions planning new maintenance processes.

The Data Mining is located in the center of every "smart", adaptive system that we can think of, and can offer huge benefits to the Railway companies. Whatever produces data, constitutes a potential target of Data mining, revealing therefor its importance and usefulness as a technology of the Future. In the meantime, the Machine Learning as a new approach to the business performances of railways, constitutes an interesting idea questioning the traditional ways procedures that were reliable in the past but now they have started showing their limitations.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

All authors would like to thank the University of West Attica for the financial support provided to them to undertake this research project.

References

[1] M. Weske "Business Process Modelling Foundation. In: Business Process Management", Published by Springer, 2019
[2] P. Aida-Maria, "Business Intelligence Methods for Sustainable Development of the Railways", Database Systems Journal **6**(2), 48-55, 2015. https://EconPapers.repec.org/RePEc:aes:dbjour:v:6:y:2015:i:23

[3] E. Kyrkos, 'Business intelligence and data mining' [eBook] Athens: Hellenic Academic Libraries Link. Chapter 4. 2015. http://hdl.handle.net/11419/1231
[4] L. Dai 'A machine learning approach for optimization in railway planning, PhD Delft University of Technology, 2018
[5] Petri, M., Pratelli, A., & Fusco, G. 'Data Mining and Big Freight Transport Database Analysis and Forecasting Capabilities'. Transactions on Maritimes Science, 2016. https://doi.org/10.7225/toms.v05.n02.001
[6] J.C. Wagenaar, J.C., Kroon, L.G., Schmidt, M. 'Maintenance Appointments in Railway Rolling Stock Rescheduling'. Transportation Science, 51(4) 1138-1160., 2017. https://doi.org/10.1287/trsc.2016.0701
[7] D. Ronanki, S. A. Singh and S. S. Williamson, 'Comprehensive Topological Overview of Rolling Stock Architectures and Recent Trends in Electric Railway Traction Systems,' in IEEE Transactions on Transportation Electrification, **3**(3) 724-738, 2017. https://doi.org/ 10.1109/TTE.2017.2765518
[8] E. Bosshca, 'Big Data in railway operation: using artificial neural networks to predict train delay propagation', University of Twente, PhD Thesis, 2016
[9] I. Öztürk, G. Güner, E. Tümer, 'The Root Causes of a Train Accident: Lac-Mégantic Rail Disaster' Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). Advances in Intelligent Systems and Computing, 823. Published by Springer, 2018. https://doi.org/10.1007/978-3-319-96074-6_21
[10] M. Heidarysafa, K. Koesari, L. E. Barnes, Brown 'Analysis of Railway Accidents-Narratives using Deep Learning' IEEE International Conference on Machine Learning and Application., 2018. https://doi.org/. 10.1109/ICMLA.2018.00235
[11] JudithHurwitz, Daniel Kirsch ´Machine learning´ IBM Limited Edition, Published by John Wiley & Sons, Inc., 2018.
[12] G.R. Devi, Karpagam, V. Kumar, 'A survey of machine learning techniques, Int. J. of Computational Systems Engineering, 3(4), 203 – 212, 2017. https://doi.org/10.1504/IJCSYSE.2017.089191
[13] R. Changala, D.R. Rao, T. Janardhana, P.K. Kumar, Kareemunnisa 'Knowledge Discovery Process: The Next Step for Knowledge Search' International Journal of Innovative Research in Computer and Communication Engineering **3**(5), 1-6, 2015. https://doi.org/10.15680/ijircce.2015.0305127
[14] S.S. Bhaskaran 'An Investigation into the Knowledge Discovery and Data Mining (KDDM) process to generate course taking pattern characterized by contextual factors of students in Higher Education Institution (HEI) , PhD Thesis, Brunel University, London, 2017
[15] Song, Yan-Yan, and Ying Lu. "Decision tree methods: applications for classification and prediction." Shanghai archives of psychiatry **27**(2), 1-6, 2015. https://doi.org/ 10.11919/j.issn.1002-0829.215044
[16] I.D. Mienyea, Y. Suna, Z. Wang 'Prediction performance of improved decision tree-based algorithms: a review' 2nd International Conference on Sustainable Materials Processing and Manufacturing, Published by Elsevier, 2019. https://doi.org/ 10.1016/j.promfg.2019.06.011
[17] M. Batra, R. Agrawal, 'Comparative Analysis of Decision Tree Algorithms. In: Panigrahi B., Hoda M., Sharma V., Goel S. (Eds) Nature Inspired Computing. Advances in Intelligent Systems and Computing, 652. Springer, Singapore, 2018. https://doi.org/10.1007/978-981-10-6747-1_4
[18] S.B. Begenova, T.V. Avdeenko, 'Building of fuzzy decision trees using ID3 algorithm': Journal of Physics: International Conference Information Technologies in Business and Industry, 2018. https://doi.org/ 10.1088/1742-6596/1015/2/022002
[19] A. Cherfi, K. Nouira & A. Ferchichi 'Very Fast C4.5 Decision Tree Algorithm', Applied Artificial Intelligence, **32**(2), 2018. https://doi.org/10.1080/08839514.2018.1447479
[20] E.G. Kulkarni and R.B. Kulkarni, "Weka Powerful Tool in Data Mining. IJCA Proceedings on National Seminar on Recent Trends" in Data RTDM, 2, 10-15, 2016.
[21] S. Akinola, O. Oyabugbe, 'Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study'. Journal of Software Engineering and Applications, 8, 470-477, 2015. https://doi.org/ 10.4236/jsea.2015.89045
[22] F.B. Márquez, 'Acquiring and Exploiting Lexical Knowledge for Twitter Sentiment Analysis', University of Waikato, PhD Thesis, 2017
[23] F. Škegro, J. Zoroja, and V. Šimičević, "Credit Scoring Analysis: Case Study of Using Weka" (September 7, 2017). 2017 ENTRENOVA Conference Proceedings, Available at SSRN: https://ssrn.com/abstract=3282504
[24] L. Jonguk, "Fault Detection and Diagnosis of Railway Point Machines by Sound Analysis." Sensors (Basel, Switzerland) **16**(4), 549-555, 2016. https://doi.org/10.3390/s16040549
[25] F.M.N. Ali & A.A.M. Hamed 'Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents', Journal of Information and Telecommunication, **2**(3), 231-245, 2018. https://doi.org/10.1080/24751839.2018.1448205
[26] J. Li, J. He, Z. Liu, H. Zhang, C. Zhang, "MATEC Web of Conferences 272, 01035 'Traffic accident analysis based on C4.5 algorithm in WEKA' School of Transportation", Southeast University, Nanjing 211189, Jiangsu, China, 2019. https://doi.org/10.1051/matecconf/201927201035
[27] G. Tsaganos, D. Papachristos, N. Nikitakos, D. Dalaklis, A.I. Ölçer, "Fault Detection and Diagnosis of Two-Stroke Low-Speed Marine Engine with Machine Learning Algorithms", Conference: 3rd International Naval Architecture and Maritime SymposiumAt: Istanbul, Turkey, 2018. https://www.researchgate.net/publication/324835430
[28] F. J. Morales, A. Reyes, N. Caceres, L. Romero, F. G. Benitez ' Automatic Prediction of Maintenance Intervention Types in Roads using Machine Learning and Historical Records' Transportation Research Record Journal of the Transportation Research Board, **2672**(44), 2018. https://doi.org/10.1177/0361198118790624
[29] H. Khaksar, A. Sheikholeslami, 'Airline delay prediction by machine learning algorithms' International Journal of Science and Technology, ume 26(5), 2019. https://doi.org/ 10.24200/SCI.2017.20020
[30] A. Fredrik, 'Reducing ships' fuel consumption and emissions by learning from data' Linnaeus University, PhD Dissertation, 2018.
[31] N. Barmpounakis 'Investigating the decision-making process of drivers during overtaking by Powered Two Wheelers' National Technical University of Athens, Department of Transportation Planning & Engineering PhD Dissertation, 2017.
[32] A. Kiranmai, J. Laxmi, "Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy", Prot Control Mod Power Syst 3, 29 2018. https://doi.org/10.1186/s41601-018-0103-3
[33] J. Shen, O. Lederman, J. Cao, F. Berg, S. Tang, A. Pentland, "Gina: Group gender identification using privacy-sensitive audio data", IEEE International Conference on Data Mining (ICDM), 457-466 2018. https://doi.org/ 10.1109/ICDM.2018.00061