ISSN 1648-4142 / eISSN 1648-3480

VILNIUS TECH
Vilnius Gediminas Technical University

TRANSPORT

2024
Volume 39
Issue 1
Pages 54–63
https://doi.org/10.3846/transport.2024.21385

Original Article

ON THE ANALYTICAL STUDY OF THE SERVICE QUALITY OF INDIAN RAILWAYS UNDER SOFT-COMPUTING PARADIGM

Saibal MAJUMDER[1], Aarti SINGH[2], Anupama SINGH[3], Mykola KARPENKO[4], Haresh Kumar SHARMA[5]✉, Somnath MUKHOPADHYAY[6]

[1] *Dept of Computer Science and Engineering (Data Science), Dr. B. C. Roy Engineering College, Durgapur, India*
[2] *FORE School of Management, New Delhi, India*
[3] *Dept of Strategic Environmental Management, Birla Institute of Management Technology, Greater Noida, India*
[4] *Dept of Mobile Machinery and Railway Transport, Vilnius Gediminas Technical University, Vilnius, Lithuania*
[5] *Dept of Operations Management and Decision Sciences, Birla Institute of Management Technology, Greater Noida, India*
[6] *Dept of Computer Science and Engineering, Assam University, Silchar, India*

Highlights:

- 7 vital attributes of the most popular trains of Indian Railways are considered;
- the overall performance of the trains is rated based on 7 important related attributes;
- a rough set decision support system based on several rules is put into place to analyse the importance of train attributes and assign a performance rating;
- a comparative analysis based on seven performance metrics is conducted, which eventually predicts the overall train rating by employing 3 ML estimators – the ETC, SVMC, and MNBC.

Article History:

- submitted 2 March 2023;
- resubmitted 6 April 2023;
- accepted 29 April 2023.

Abstract. Indian Railway Catering and Tourism Corporation (IRCTC) is among the busiest railways reservation systems since the Indian Railways (IR) is the vital and economical mode of transportation in India. Hence, rating of the trains seems to be critical aspect for selecting an appropriate train for travelling. In this study, we have considered 7 vital attributes of 500 popular trains and rate their performance based on 7 important related attributes. For this purpose, we have employed 2 different approaches to analyse of the train attributes, which eventually contribute to the overall performance of the trains. Here, we have developed a rule based rough set decision support system to analyse the criticality of the train attributes while rating the train performance. Furthermore, we have also used 3 Machine Learning (ML) model estimators: Extra Trees Classifier (ETC), Support Vector Machine Classifier (SVMC) and Multinomial Naive Bayes Classifier (MNBC) and perform their comparative analysis with respect to 7 performance metrics while predicting the overall train rating based.

Keywords: rough set theory, extra trees classifier, support vector machine classifier, multinomial naive Bayes classifier, performance metrics.

✉Corresponding author. E-mail: *hareshshrm@gmail.com*

Notations

ANP – analytic network process;
AUROCCS – area under the receiver operating characteristic curve score;
DEMATEL – decision making trial and evaluation laboratory;
DMS – data mining scaffolding;
ETC – extra trees classifier;
HL – hamming loss;
HL–RF – Hasofer Lind and Rackwitz Fiessler;
IR – Indian Railways;
IRCTC – Indian Railway Catering and Tourism Corporation;
IRRS – Indian Railway Reservation System;
MCC – Matthew's correlation coefficient;
ML – machine learning;
MNBC – multinomial naive Bayes classifier (MNBC);

RFECV – recursive feature elimination with cross validation;
ROSE – rough sets data explorer;
RST – rough set theory;
SVMC – support vector machine classifier;
UCI – University of California Irvine (Irvine, CA, US).

1. Introduction

The service quality is one of the crucial factors while opting any transportation system. In India, there are different transportation systems like road transportation system, air transportation and rail transportation system, however, Indian rail transportation has been considered as most cost effective transportation system for long distance. As rail transportation is widely opted, to travel it makes IR as most business transportation system and to smooth the process of such crowded system required computing. The soft-computing system has been adopted by the IR most easy and supporting process. In this study, the service quality system has been analysed under soft-computing of IR by using RST and ML.

Ever since RST (Pawlak 1982, 1991; Pawlak, Skowron 2007) has been proposed, it has become an extremely crucial mathematical technique for handling uncertainty and vagueness. Not only researchers but engineers also utilize the technique for as it is a valuable alternative for other theories like vague, fuzzy set and many more. The extent of utilization of RST includes medical and health sciences, banking sector, data mining, marketing and other systems. The primary goal of RST is to utilize the lower approximation set and upper approximation set or more commonly referred to as 2 crisps for identifying the approximation of a set. Both the sets have equal responsibilities; the upper approximation includes the categories that might (probably) belong to a set, whereas the lower approximation inculcates the categories that positively belong to the set. This collected data is represented in a tabular format, which is called the information table.

Zhang *et al.* (2019) applied the reliability theory, which was originated from the HL–RF algorithms for the railways transport chain data in China. Vojtek *et al.* (2021) studied the Radioblock Railway Safety System and how various deep analysis can improve and enhance the safety features further; the research also suggested the utilization in European Traffic Control System in comparison to Radioblock Railway Safety System. Guo & Dong (2021) statistically analysed the effect and function of railways in enhancing the trade in China. Rao (2021) utilized DEMATEL-ANP based multi-criteria decision-making model for understanding and evaluating sustainability indicators of intercity railway transport in Taiwan. Cascetta *et al.* (2020) cited a case study that depicted the effect of high speed railways on the transport accessibility and development of the economy in Italian Railways.

Donaldson & Hornbeck (2016) computed the data of 1890 by utilizing the market access approach and understood how railroads impacted the agriculture sector. He found that the value of land around the areas that were connected with railroads increased by 60% between the years from 1870 to 1890. As far as the decision-making process that revolves around the theory of rough set is involved, Sharma *et al.* (2018) projected the concept of RST. Furthermore they employed them for determining the weight values of any criteria under analysis through the best-worst method. Consecutively, Ye *et al.* (2021) proposed an approach that was based on fuzzy-rough set wherein there were multiple attributes that could aid in reaching to a decision. Further, they also demonstrated the practicability of the projected approach by solving the shape selection problem that was present in the UCI database. Of late, Hu *et al.* (2021) proposed a model for removing any unnecessary data, creating a valuable subsets of identified attributes and finally calculating the degree of dependency in a weighted neighbourhood rough set. This model can be further validated by applying it to biomedical and ML datasets. In addition to the above cited research, there is more research on railway safety and transportation (Rao 2021; Vojtek *et al.* 2021; Cascetta *et al.* 2020; Stević *et al.* 2017; Velaga *et al.* 2022; Sharma *et al.* 2018).

In recent times, there are some studies on transportation domain, where various ML approaches are adopted. Yan & Chen (2021) proposed a study related to the prediction of the traffic volume of various random section of the city using graph convolution neural network model. Rasaizadi *et al.* (2021) employed ML models to predict the traffic states of the segments of a road network of the rural region. Liu (2022) proposed a study on the refinement of the urban traffic conditions using ML and edge computing techniques. Raju *et al.* (2022) incorporated both ML and deep learning models for predicting the condition of the heterogeneous traffics. Subsequently, Huang (2022) used a SVMC for real time early safety warning of the traffic stream. Afterwards, He & Li (2022) proposed an economic forecasting model based on deep learning approach for predicting the traffic flow in a smart city.

Considering the above mentioned literature studies, to the best of our knowledge, any investigation on rating the trains based on their important inherent attributes is yet to be conducted by employing the rough set and machine learning approaches. The current article, revolves around the case study of the IR. The article utilizes the rough sets as the primary data mining tool. The major contributors of the research are:

- a DMS is applied based on the theory of rough sets, which would guide and aid the passengers in booking their train tickets for any journey effectively. The relevant information regarding the ticket reservation, which is important to IRCTC will then be processed effectively through a MCDM approach. This approach focuses on the decision-making, which is further controlled by various "if–then" decision rules. These decision rules are defined by the IRRS, which takes into account other

relevant and important factors regarding reservation of tickets;
- 3 ML estimators: ETC (Geurts *et al.* 2006), SVMC (Cortes, Vapnik 1995) and MNBC (McCallum, Nigam 1998; Manning *et al.* 2008) are used dataset under consideration and a comparative study is conducted to analyse the performance of the estimators.

In the present case study, the rules have been defined utilizing the RST and are based on 8 different and relevant criteria. The aim and output of the research is to direct the passengers efficiently while they are reserving tickets. This will be fruitful for the IRCTC, which will suggest the IR for enhancing and improving their services and will further upgrade their services for improving their customer's experience. The manuscript is organized as follows:
- current Section 1 an introduction;
- in Section 2 we discuss some rudimentary concepts of RST;
- the case study related to IRs is discussed in Section 3;
- subsequently, the analysis of the result for our considered dataset with respect to rough set and ML approaches are provided in Section 4;
- finally, the epilogue of our study (conclusions) is presented in Section 5.

2. Preliminaries

This section briefly introduces the concept of RST and its related properties in the following subsections.

2.1. Data table and indiscernibility relation of RST

RST is the derived mathematical solution for handling ambiguity and uncertainty (Figure 1). It is built on the hypothesis that, to a certain degree, the information is associated with every object in the discourse of the universe. The key recognising features of the RST are an upper and lower approximation (Pawlak 1982, 1991; Pawlak, Skowron 2007). The decision object and its related information in RST are usually constituted in the form of an information table, which is further represented through a 4-tuple information system $T = A, L, M, \alpha$, where A symbolise a finite set of entities and, L is the finite set of features and $\alpha : A \times L \to M$ is the information system.

For anyone set of features $N \subseteq L$ there exist an equivalence relation $IND(N)$ such that $IND(N) = \{\gamma_i, \gamma \in A \times A \mid \forall \beta \in N, \beta(\gamma_i) = \beta(\gamma_j)\}$, where $\beta(\gamma_i)$ represents the value of the attribute β for the element γ_i. $IND(N)$ is known as the indiscernibility relation.

Given the set of attribute $N \subseteq L$ and K in A, the lower and upper approximation of K are defined as follows:

$$\underline{K}_N = \cup\left\{K_i \mid [\gamma_i]_{IND(N)} \subseteq K\right\}; \quad (1)$$

$$\overline{K}_N = \cup\left\{K_i \mid [\gamma_i]_{IND(N)} \cap K \neq \varnothing\right\}. \quad (2)$$

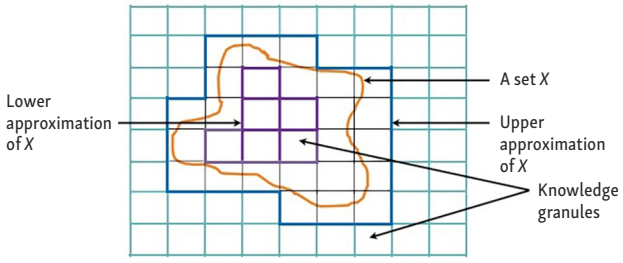

Figure 1. A schematic diagram of a rough set

The boundary region of set K is described as follow:

$$BND_N(K) = \overline{K}_N - \underline{K}_N. \quad (3)$$

In other words, the components, which are surely the members of the set K in the knowledge base N constitute the lower approximation set of K and is denoted by $\underline{K}_N$. Whereas, the probable components, which belong to K constitute the upper approximation set of K and is represented as $\overline{K}_N$. The objects, which cannot be resolutely classify as the objects of K is known as the boundary set of K and is expressed as $BND_N(K)$. If $BND_N(K) = \{\Phi\}$, then the lower and upper upper approximation will be equal. Otherwise, when $BND_N(K) \neq \{\Phi\}$, then the set K is referred to as a rough set with respect to N.

2.2. Accuracy of approximation and reduction of knowledge (Pawlak 1991)

Inaccuracy or vagueness in any set arises due to the presence of a boundary line region. This is inversely proportional to the accuracy of the set; that means that in order to increase the accuracy of a set; the boundary line region of the category should decrease.

In the context of numbers, we can define accuracy of approximation as follows:

$$\varphi_N(K) = \frac{|\underline{K}_N|}{|\overline{K}_N|}, \quad (4)$$

where: $|K|$ represents the cardinality of any set K.

Clearly, $0 \le \varphi_N(K) \le 1$. If $\varphi_N(K) = 1$, K is exact with respect to N, otherwise, K is rough (ambiguous) with respect to N, when $\varphi_N(K) < 1$.

2.3. Reduction of knowledge

Removing attributes that are unnecessary from the data set is important as it would shift the focus on the necessary attributes for any information system. Such type of attribute subset is termed as reduct and is a crucial part of any information system.

2.4. Reduct and core

The core contains the set of all the vital parameters that are extremely essential and immovable entities of the in-

formation table, they also possess a set of all common reducts of M, where M is the set of condition attributes.

Core can be denoted as follow:

$$CORE(M) = \cap RED(M). \quad (6)$$

2.5. Decision-making using RST

The reduct set is represented by the necessary attributes that are present in the initial data set without any possible reduction. The minimum rule of that is created from the information table possess the following steps from the RST. In order to simplify the structure of the decision table, rules are made and are listed through the following mandatory steps:

- generating an information table of the data set under study;
- calculating the upper and lower approximation value of the data sets;
- reducing the computation of certain condition attributes, which is equal to eliminating certain columns from the decision table;
- elimination of redundant attribute value;
- determining the core attributes and identifying minimal subset for all decision attributes.

A decision rule in knowledge base system can be obtained in the form of the "if–then", thus:

$$\text{if } \alpha(\gamma, \beta) = v_{\beta_1} \wedge \alpha(\gamma, \beta_2) = v_{\beta_2} \wedge \ldots \rightarrow$$
$$\text{then } \alpha(\gamma, d) = v_d$$
$$\text{for each } \beta_i \in M, d \in D$$
$$\text{and } v_{\beta_i}, v_d \in V, \text{ for every } \gamma \in A.$$

3. Case study

This section present a case study of the rating of Indian trains based on eight attributes. Accordingly, the data required for the case study is collected from the IR. A related illustration of the case study is presented in the following subsection.

Information of research problem

IR is the primary and most preferred mode of transportation opted by Indians. In terms of ranking, it is considered as the Asia's largest and World's 2nd largest rail network, which is fluently working under a single aegis. The parent organization's wing under Government of India is Ministry of Railways, which is responsible for its smooth functioning. The 1st railway track that was laid down in India ran between Mumbai and Thane in the year 1853, various units was further installed in India thereafter. The entire network thus generated were brought and nationalized under the single organization in 1951, the present day IR. In terms of furnishing and upgrading the services and amenities, IR introduced IRCTC and IRRS, the former is dedicated for catering and boosting tourism while the later eases the booking and reservation of tickets. The present study focuses on identifying the best 500 trains running in India. The study focuses on the attributes that are preferred and selected by any Indian before they plan a journey or trip to any destination. Identifying and selecting a suitable train and having a confirmed ticket for reaching the destination is the most integral part of any journey.

The data and information required for the success of the research has been obtained from various domain experts from the IR, IRCTC and the primary driving factor, feedback from the passengers (IRIS 2023). The decision of the passenger is controlled by a decision parameter, which is further dependent and governed by various conditional parameters. To obtain a valid outcome, certain variable of IRRS were incorporated in conditional parameters making them the driving force in the present study. These variables are:

- *punctuality* of the train (obtained from the interview taken of the Deputy Chief Controller / Dispatcher);
- *ticket availability* (source websites: IRCTC and IRRS);
- *cleanliness* and hygiene, quality of *food*, *rail-fanning* (discussions from various private contractors that are hired by IRCTC);
- *safety* (guidelines of the Railway Police Force);
- *unreserved coach* (source website: IR).

All the attributes, and variable that are considered and analysed for the study have been depicted and explained in Table 1.

4. Result and discussion

In this section, the analysis of the results and their discussion are done in 2 subsequent Subsections 4.1 and 4.2. Specifically, in Subsection 4.1, we have presented the

Table 1. Information related to IRRS variables

Attributes	IRRS variables	Explanation
Conditional attributes	*cleanliness*	cleaning condition for a specific train in accordance with IR policies
	punctuality	scheduled time of the train
	food	good quality edible
	ticket availability	selection of availability of train berth on the chosen route
	unreserved coach	the category, which is not available for reservation through IRCTC
	rail-fanning	the activities performed by rail-fans'
	safety	all safety and security precautions for passengers on a certain train
Decision attribute	*overall-rating*	conclusive rating of all attributes

analysis of the results using RST and in Subsection 4.2, we have employed the ML techniques and discussed its corresponding results. Both these techniques are used on the dataset, which we have prepared by gathering the relevant data from 2 popular railway websites of India namely:

- *https://www.indianrail.gov.in*;
- *https://www.irctc.co.in/nget/train-search*.

The dataset is prepared based on 500 trains for which we have considered 7 features:

- *cleanliness*;
- *punctuality*;
- *food*;
- *ticket availability*;
- *unreserved coach*;
- *rail-fanning*;
- *safety*.

The values of each feature as well as the target vector (*overall rating*) are categorical in nature and classified as:

- excellent (*e*);
- good (*g*);
- average (*a*).

4.1. Analysis using RST

The analysis of the dataset based on the RST are subsequently presented in the following subsections.

4.1.1. Approximation sets

Rough set analysis has been applied on the collected data set to calculate approximation sets and accuracy of approximation. The subsequent table depicts the approximation of sets and accuracy of approximation.

Table 2, reports and depicts the data set of train for the quality of classification and accuracy of approximation. The table also defines the classification of passengers behaviour being 0.8250, thus implies that the boundary region has very related to the information. In this case study, degree of dependency that is closer to 1 also shows that each member of the class has a better dependency among all conditional attributes. the Quality of lower approximation becomes 0.8250.

4.1.2. Decision rules using RST

The data was analysed using RST and the decision rules were obtained. These decisions were further applied to condition, decision attributes and their intra-relationship. A higher support to a particular attribute indicates a better decision, thus in this study a support of more than 40 has been considered. Predki *et al.* (1998) applied the algorithm for the creation of decision rules. Table 3 depicts that estimated results using the decision rules. Rule 1 is defined and identified by 2 attributes; *punctuality* and *safety*. Thus, if the train acquires and average rating in terms of *punctuality*, it will have an average rating in terms of *safety* and furthermore will having an overall average rating. The support of Rule 1 is by 59. Rule 2 is identified by the attributes of *cleanliness*, *food* and *rail-fanning*; if the train acquires an average rating in all 3 mentioned traits, it will have an overall average rating (support of 42). We have obtained decision rule and accuracy of data set using RST tools like ROSE2.

Table 2. Accuracy of approximation

	Average (*a*)	Good (*g*)	Excellent (*e*)
Lower approximation	98	141	176
Upper approximation	170	218	203
Boundary	72	77	27
Accuracy of approximation	0.5765	0.6468	0.8670

4.1.3. Significance of the condition attributes

The data obtained and the knowledge further generated from it can result only one reduct. Thus, after analysis every criterion is defined and determined by the conditional attributes of the data set. Their importance is clearly indicated by their presence in the decision rules. If a condition attribute has been frequently utilized in any decision rule and also has a high corresponding support value, it is thus an extremely important parameter in terms of the decisions taken by passengers about the train. Rule 5 and 6, depicted in Table 3 identify the attributes of *food*, *safety* and *cleanliness* are the most important and decisive criteria with a support of 65 and 64 respectively. The table also signifies and defines the decision rules where the most significant parameters are *punctuality*, *safety*, *food* and *cleanliness*. It can further be elucidated from the data of Table 3, that higher customer satisfaction for amenities and catering, higher will be the attention of passengers.

In this research, for better decisions, we consider the only decision rule that has support greater than 40, as higher support indicates a better and more significant decision rule.

The evaluation of the robustness of the decision rules is conducted by considering the accuracy based on the correctly classified objects. The confusion matrix, which reflects the total number of these correctly classified objects is reported in Table 4. The final accuracy percentage for the model calculated from Table 4 is 85.50%.

4.2. Analysis of results using ML techniques

In this section we have analysed the dataset using 3 ML algorithms:

- ETC (Geurts *et al.* 2006);
- SVMC (Cortes, Vapnik 1995);
- MNBC (McCallum, Nigam 1998; Manning *et al.* 2008)

and have performed a comparative study based on the importance of the features. For an in-depth analysis, we consider the dataset and apply 3 ML estimators on the dataset. Here, we have used 3 classifiers namely:

- ETC;
- SVMC;
- MNBC

Table 3. Certain decision rules of railway information system

Decision rule	Support
Rule 1: (*punctuality* = average) & (*safety* = average) => (*overall rating* = average)	59
Rule 2: (*cleanliness* = average) & (*food* = average) & (*rail-fanning* = average) => (*overall rating* = average)	42
Rule 3: (*cleanliness* = average) & (*unreserved coach* = average) & (*rail-fanning* = average) => (*overall rating* = average)	42
Rule 4: (*cleanliness* = excellent) & (*ticket availability* = good) => (*overall rating* = excellent)	41
Rule 5: (*cleanliness* = excellent) & (*safety* = excellent) => (*overall rating* = excellent)	65
Rule 6: (*food* = good) & (*safety* = excellent) => (*overall rating* = excellent)	64
Rule 7: (*cleanliness* = excellent) & (*punctuality* = excellent) & (*food* = good) => (*overall rating* = excellent)	47
Rule 8: (*cleanliness* = excellent) & (*punctuality* = excellent) & (*rail-fanning* = excellent) => (*overall rating* = excellent)	51
Rule 9: (*cleanliness* = excellent) & (*food* = good) & (*rail-fanning* = excellent) => (*overall rating* = excellent)	43
Rule 10: (*punctuality* = excellent) & (*unreserved coach* = good) & (*safety* = excellent) => (*overall rating* = excellent)	43
Rule 11: (*punctuality* = excellent) & (*rail-fanning* = excellent) & (*safety* = excellent) => (*overall rating* = excellent)	54
Rule 12: (*food* = good) & (*ticket availability* = good) & (*rail-fanning* = excellent) => (*overall rating* = excellent)	60

Table 4. Confusion matrix for 3 decision classes

	Average (*a*)	*Good* (*g*)	*Excellent* (*e*)
Average (*a*)	113	20	0
Good (*g*)	28	139	15
Excellent (*e*)	1	9	178

on the dataset to rate the trains with respect to its 7 important features:

- *cleanliness*;
- *punctuality*;
- *food*;
- *ticket availability*;
- *unreserved coach*;
- *rail-fanning*;
- *safety*.

The target vector for our dataset is *overall rating*, which rate the 500 trains of our dataset based on the above mentioned features. The analysis of the dataset based on the above mentioned classifiers are discussed in details in the subsequent subsections.

4.2.1. Data pre-processing and feature selection

For the dataset, in each feature, we count the occurrences of *excellent* (*e*), *good* (*g*) and *average* (*a*), which are tabulated in Table 5. Here, we observe that for each feature, total occurrences of *excellent* (*e*), *good* (*g*) and *average* (*a*) are more than 5%. This fact infer that there does not exist any outliers for all the features. For further analysis of the data, we have used *label encoder* of *scikit-learn* in *Python* library to convert the categorical data to numeric data. Further, to recalled the data, *min–max scaler* from the same library. Subsequently, to investigate the chance of the overfitting, we have also explored the Pearson correlation coefficient between a pair of features as depicted in Figure 2. In this figure, we observe that the maximum correlated value is 0.52 between *cleanliness* and *safety*, which we consider as significant. In Figure 3, we have presented the density plots of each features of our dataset. Here, the density plot is essentially the continuous and smooth version of the univariate set of observations of each feature visualized as histograms. In these density plot, the *x*-axis represents the value of the feature and the *y*-axis represents the probability density function for the kernel density estimation (Gaussian in our study). The values of a particular feature in the *x*-axis are 0, 1 and 2, which respectively represents the 3 discrete labels: excellent (*e*), good (*g*) and average (*a*). Furthermore, we observe that *cleanliness*, *food*, *ticket availability* and *unreserved coach* have similar distributions of data. Whereas, *punctuality*, *rail-fanning* and *safety* have identical distribution pattern of data.

Table 5. Occurrences of excellent (*e*), good (*g*) and average (*a*) for each feature in the dataset

Feature value	Cleanliness	Punctuality	Food	Ticket availability	Unreserved coach	Rail-fanning	Safety
Excellent (*e*)	282	254	265	258	229	286	291
Good (*g*)	131	145	210	152	204	166	109
Average (*a*)	90	104	28	93	70	51	103

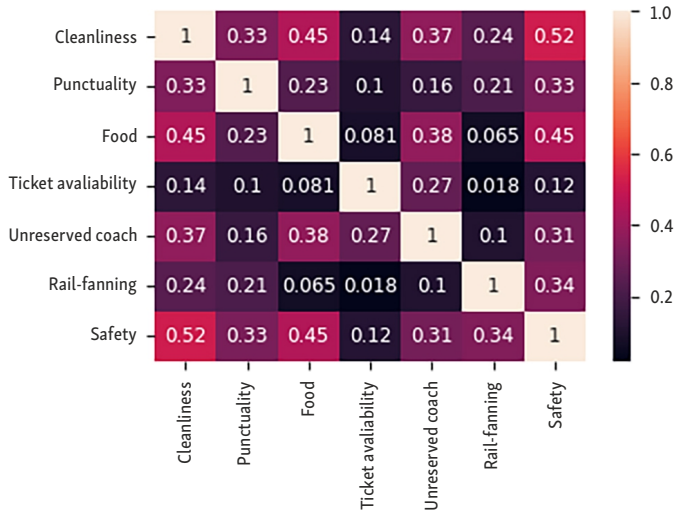

Figure 2. Correlation matrix between the pairwise features

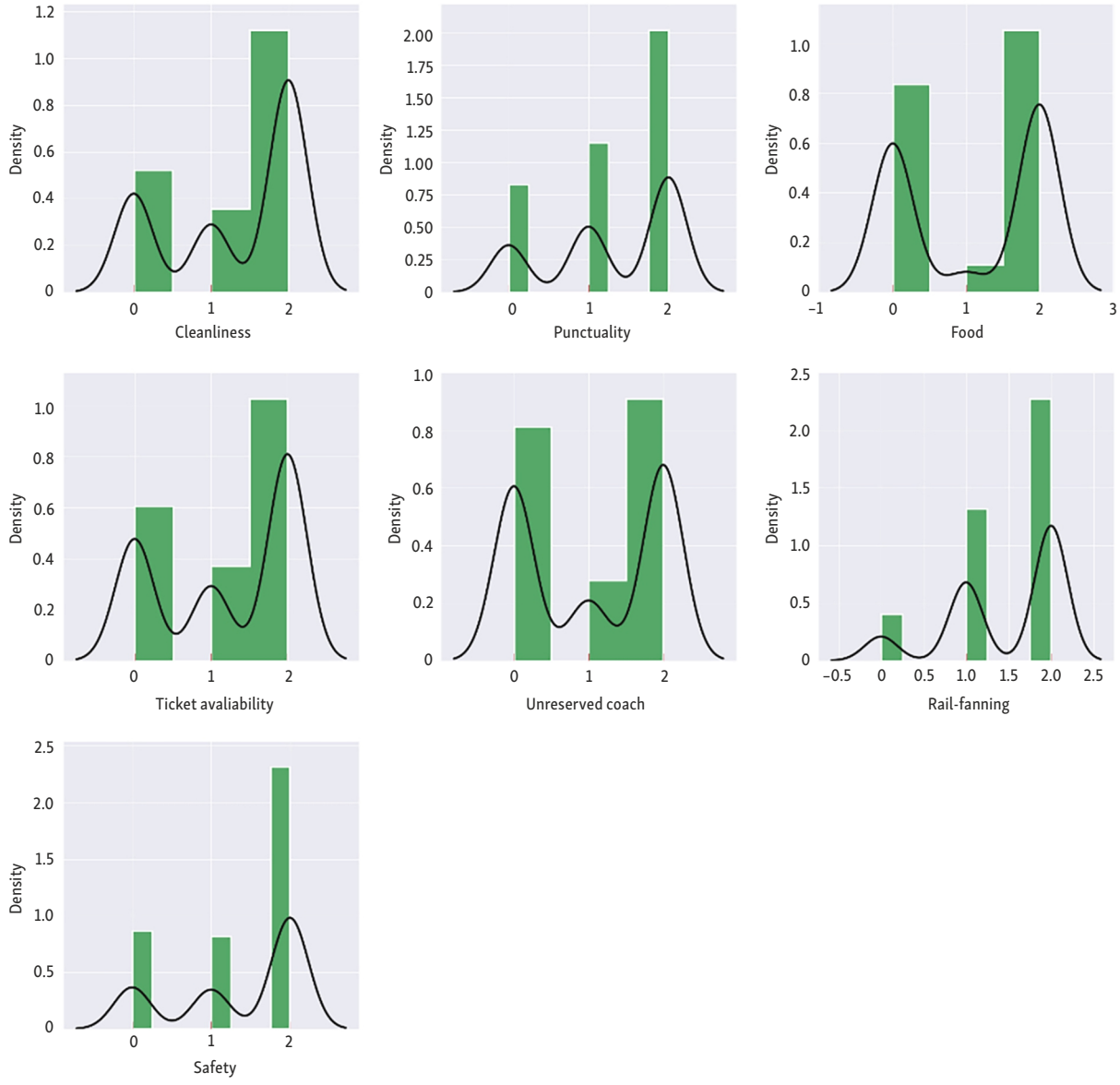

Figure 3. Density plot of all the features of the dataset

4.2.2. RFECV

In order to reduce the dimension of the feature matrices, in this study, we have considered *scikit-learn*'s in-built class RFECV, which is popularly used to select those features from a training dataset that are most pertinent as far as a prediction of the target variable is concerned. Usually, some ML algorithms can be deceived by irrelevant features, resulting in shoddier performance in their predictive capabilities. In such contexts, feature selection becomes useful, which select only a subset of features to enhance the effectiveness and efficiency of the ML algorithms. RFECV uses recursive feature selection with cross-validation loop to find the optimal features.

Using RFECV, we determine the optimal features of the dataset with respect to 3 the 3 classifiers:

- ETC;
- SVMC;
- MNBC.

Here, we observe 2 scenarios when each of the 3 classifiers are considered for RFECV. When, ETC is considered with RFECV, then *cleanliness*, *punctuality*, *food* and *safety* are selected as important features of the dataset. However, for the remaining 2 classifiers – SVMC and MNBC, all the 7 features are selected as important.

4.2.3. Performance analysis of classifiers

For the purpose of training the ML algorithms efficiently, some important hyperparameters of the classifiers are tuned. Accordingly, we have used the *Randomized Search CV* from the in-built *scikit-learn* library, to determine the optimal values of the hyperparameters of the classifiers. These hyperparameter values are provided below:

- ETC: *criterion* = *gini*, *min_samples_split* = 10, *min_samples_leaf* = 2, *max_features* = *auto*, *max_depth* = 10;
- SVMC: *C* = 0.5018745921487388, *kernel* = *linear*, *degree* = 1;
- MNBC: *alpha* = 0.909670656388511, *fit_prior* = *false*.

Subsequently, we compare the performance of the classifiers with respect to the 10 fold cross-validation scores calculated in terms of 7 performance measures:

- accuracy;
- precision;
- recall;
- f1-score;
- AUROCCS;
- HL;
- MCC.

These values are reported in Table 6. Accordingly, we observe that SVMC outperforms both ETC and MNBC in terms of all the performance metrics. Furthermore, we have also generated the confusion matrices depicted in Figure 4 corresponding for ETC, SVMC and MNBC. For this purpose, we train each of the classifiers with 80% of the data in the dataset and tested the prediction capability of the classifiers on the remaining 20% of the data. For this purpose, we have also set the *random_state* of ETC and SVMC as 47. From Figure 4, it can also be inferred that, the confusion matrix generated by SVMC in Figure 4b is superior to the remaining confusion matrices in Figure 4a and Figure 4c.

Table 6. Cross-validation score with respect to the performance metrics generated by ETC, SVMC and MNBC

Classifier	Accuracy	Precision	Recall	f1-score	AUROCCS	HL	MCC
ETC	0.841	0.864	0.844	0.849	0.881	0.158	0.766
SVMC	0.861	0.868	0.859	0.862	0.894	0.139	0.791
MNBC	0.713	0.714	0.712	0.712	0.789	0.287	0.575

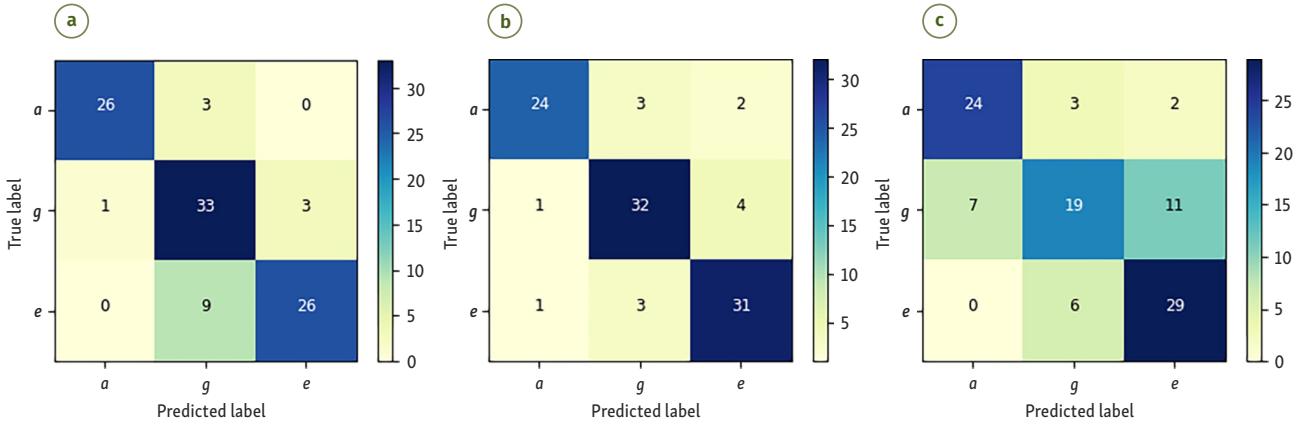

Figure 4. Confusion matrices of: (a) – ETC; (b) – SVMC; (c) – MNBC

5. Conclusions

In this article, we have proposed a case study on the IR by using and applying the key concepts of rough set analysis and ML. From the viewpoint of the RST, the uncertainty of the attribute-based data is processed and simplified using rough set. Although RST has a wide range of applications, its usage in understanding the behaviour of any passenger in selecting a train based on its *overall rating* on the basis of various vital parameters including cleanliness, punctuality, food quality, ticket availability, availability of unreserved coach, rail-fanning and safety is limited. Therefore, our approach in this article is to implement the RST technique to understand the behaviour of passengers for selecting a train for their upcoming journey based on our proposed rough set rule base, which constitute of twelve rules. The result obtained following these rules depicts how RST can be effectively implemented for mining the data whenever a passenger will select a suitable train as per their preference for their journey. RST methodology adopts the ideology that while analysing if no change has been observed in the quality of approximation, the condition attribute can be avoided. The present study holds each attribute equally important based on the analysis of both rough set and ML approaches. Therefore, all the conditional attributes become a reduct as well as core attributes. Accordingly, each of these attributes holds significant value especially when it comes to identifying train preference by the passengers.

From the perspective of ML analysis on the dataset, it is observed that the attributes or features are scarcely correlated with a maximum value of 0.52 between Cleanliness and safety. Moreover, all the eight attributes are very highly correlated (≈0.93) on the target attribute *overall rating*. Both these facts infer that all the features are important and therefore are not eliminated. Furthermore, we have conducted a comparative analysis of the 3 ML estimators for predicting a suitable train on behalf of a passenger during the course of the journey. These ML estimators include ETC, SVMC and MNBC. The performance of these ML estimators is measured based on 7 performance metrics including accuracy, precision, recall, f1-score, AUROCCS, HL and MCC. Here, we observe that SVMC emerges as the superior estimator among the 3 with its enhanced predictive capability as far as our dataset is concerned.

In future, we would like to use the 2 ways analysis approach constituting the rough set and ML approaches in different spheres of the IR (including prediction of passenger ticket confirmation, prediction of vital criteria for improvement of railways platforms and predicting the suitable geographical cohorts for the construction of the railway tracks of the high speed bullet trains), with an objective to serve the IR so that this prestigious Government organization can provide a hassle free service to the passengers.

Author contributions

Conceptualization – *Saibal Majumder* and *Haresh Kumar Sharma*.

Methodology – *Haresh Kumar Sharma* and *Anupama Singh*.

Software – *Saibal Majumder* and *Haresh Kumar Sharma*.

Validation – *Haresh Kumar Sharma*, *Aarti Singh* and *Anupama Singh*.

Formal analysis – *Mykola Karpenko* and *Aarti Singh*.

Investigation – *Haresh Kumar Sharma* and *Mykola Karpenko*.

Resources – *Somnath Mukhopadhyay* and *Aarti Singh*.

Data curation – *Haresh Kumar Sharma* and *Saibal Majumder*.

Disclosure statement

The authors declare that they have no conflicts of interest.

References

Cascetta, E.; Cartenì, A.; Henke, I.; Pagliara, F. 2020. Economic growth, transport accessibility and regional equity impacts of high-speed railways in Italy: ten years ex post evaluation and future perspectives, *Transportation Research Part A: Policy and Practice* 139: 412–428. https://doi.org/10.1016/j.tra.2020.07.008

Cortes, C.; Vapnik, V. 1995. Support-vector networks, *Machine Learning* 20(3): 273–297. https://doi.org/10.1007/BF00994018

Donaldson, D.; Hornbeck, R. 2016. Railroads and American economic growth: a "market access" approach, *The Quarterly Journal of Economics* 131(2): 799–858. https://doi.org/10.1093/qje/qjw002

Geurts, P.; Ernst, D.; Wehenkel, L. 2006. Extremely randomized trees, *Machine Learning* 63(1): 3–42. https://doi.org/10.1007/s10994-006-6226-1

Guo, Y.; Dong, B. 2021. Railway and trade in modern China: evidence from the 1930s, *China Economic Review* 69: 101661. https://doi.org/10.1016/j.chieco.2021.101661

He, Y.; Li, X. 2022. Feasibility of economic forecasting model based on intelligent algorithm of smart city, *Mobile Information Systems* 2022: 9723190. https://doi.org/10.1155/2022/9723190

Hu, M.; Tsang, E. C. C.; Guo, Y.; Chen, D.; Xu, W. 2021. A novel approach to attribute reduction based on weighted neighborhood rough sets, *Knowledge-Based Systems* 220: 106908. https://doi.org/10.1016/j.knosys.2021.106908

Huang, M. 2022. SVM-based real-time identification model of dangerous traffic stream state, *Wireless Communications and Mobile Computing* 2022: 6260395. https://doi.org/10.1155/2022/6260395

IRIS. 2023. *Indian Railways Information System* (*IRIS*). Available from Internet: https://indiarailinfo.com/train

Liu, L. 2022. Refined judgment of urban traffic state based on machine learning and edge computing, *Journal of Advanced Transportation* 2022: 7593772. https://doi.org/10.1155/2022/7593772

Manning, C. D.; Raghavan, P.; Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press. 506 p. https://doi.org/10.1017/CBO9780511809071

McCallum, A.; Nigam, K. 1998. A comparison of event models for naive Bayes text classification, in *AAAI-98 Workshop on Learning for Text Categorization*, 26–30 July 1998, Madison, WI, US, 41–48. Available from Internet: https://aaai.org/papers/041-ws98-05-007/

Pawlak, Z. 1982. Rough sets, *International Journal of Computer & Information Sciences* 11(5): 341–356. https://doi.org/10.1007/BF01001956

Pawlak, Z. 1991. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Springer. 231 p. https://doi.org/10.1007/978-94-011-3534-4

Pawlak, Z.; Skowron, A. 2007. Rudiments of rough sets, *Information Sciences* 177(1): 3–27. https://doi.org/10.1016/j.ins.2006.06.003

Predki, B.; Słowiński, R.; Stefanowski, J.; Susmaga, R.; Wilk, Sz. 1998. ROSE – software implementation of the rough set theory, *Lecture Notes in Computer Science* 1424: 605–608. https://doi.org/10.1007/3-540-69115-4_85

Raju, N.; Arkatkar, S. S.; Easa, S.; Joshi, G. 2022. Data-driven approach for modeling the nonlane-based mixed traffic conditions, *Journal of Advanced Transportation* 2022: 6482326. https://doi.org/10.1155/2022/6482326

Rao, S.-H. 2021. Transportation synthetic sustainability indices: a case of Taiwan intercity railway transport, *Ecological Indicators* 127: 107753. https://doi.org/10.1016/j.ecolind.2021.107753

Rasaizadi, A.; Seyedabrishami, S.; Abadeh, M. S. 2021. Short-term prediction of traffic state for a rural road applying ensemble learning process, *Journal of Advanced Transportation Volume* 2021: 3334810. https://doi.org/10.1155/2021/3334810

Sharma, H. K.; Kumari, K.; Kar, S. 2018. Air passengers forecasting for Australian airline based on hybrid rough set approach, *Journal of Applied Mathematics, Statistics and Informatics* 14(1): 5–18. https://doi.org/10.2478/jamsi-2018-0001

Sharma, H. K.; Roy, J.; Kar, S.; Prentkovskis, O. 2018. Multi criteria evaluation framework for prioritizing Indian Railway stations using modified rough AHP-MABAC method, *Transport and Telecommunication Journal* 19(2): 113–127. https://doi.org/10.2478/ttj-2018-0010

Stević, Ž.; Pamučar, D.; Kazimieras Zavadskas, E.; Ćirović, G.; Prentkovskis, O. 2017. The selection of wagons for the internal transport of a logistics company: a novel approach based on rough BWM and rough SAW methods, *Symmetry* 9(11): 264. https://doi.org/10.3390/sym9110264

Velaga, N. R.; Sharma, H. K.; Majumder, S.; Biswas, A.; Prentkovskis, O.; Kar, S.; Skačkauskas, P. 2022. A study on decision-making of the Indian Railways reservation system during COVID-19, *Journal of Advanced Transportation* 2022: 7685375. https://doi.org/10.1155/2022/7685375

Vojtek, M.; Matuska, J.; Siroky, J.; Kugler, J.; Kendra, K. 2021. Possibilities of railway safety improvement on regional lines, *Transportation Research Procedia* 53: 8–15. https://doi.org/10.1016/j.trpro.2021.02.001

Yan, G.; Chen, Y. 2021. The application of virtual reality technology on intelligent traffic construction and decision support in smart cities, *Wireless Communications and Mobile Computing* 2021: 3833562. https://doi.org/10.1155/2021/3833562

Ye, J.; Zhan, J.; Xu, Z. 2021. A novel multi-attribute decision-making method based on fuzzy rough sets, *Computers & Industrial Engineering* 155: 107136. https://doi.org/10.1016/j.cie.2021.107136

Zhang, R.; Li, L. Jian, W. 2019. Reliability analysis on railway transport chain, *International Journal of Transportation Science and Technology* 8(2): 192–201. https://doi.org/10.1016/j.ijtst.2018.11.004