IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8325

Train Delay Prediction System in India Using Machine Learning Techniques

Heena Gupta[1], Vidya Shree MG[2]

Assistant Professor, Department of Computer Science, Mount Carmel College, Autonomous, Bengaluru, India[1]

Student, Department of Computer Science, Mount Carmel College, Autonomous, Bengaluru, India[2]

Abstract: Indian railways are one of the largest railway networks which are managed by the Indian government. Indian trains are overcrowded most of the time due to huge demand and very less supply. Train delays are mainly caused by train speed restrictions imposed for safety reasons which can disrupt the train schedules for the connecting journeys. Real-world data was obtained to perform the research. The data was cleaned, pre-processed, and dimensionality reduction was done. Further, five machine learning models for classification were applied. The highest prediction accuracy was obtained by the Random Forest model.

Keywords: Machine Learning, Classification Models, Train Delay, Real-World Data, Delay Prediction, Transit Delays.

I. INTRODUCTION

Indian railways are one of the largest railway networks which are managed by the Indian government. Railways first came to India in the year 1853. Today, it is one of the busiest railway networks in the whole world carrying around 18million passengers daily and this number is seen to increase every year. As we are all aware, punctuality is the most important factor and a huge advantage of the trains when compared to other long-distance transports. It is also considered as one of the major performance measures for passenger trains. In India, there were about 25.3million people who traveled using trains in the year 2006. According to a survey in 2018, at least 80million people prefer to travel by trains. Under ideal situations, the trains follow the schedule in the timetable. However, the actual operation of trains is affected by various factors like weather, equipment failure, the number of travelers, the crossing of trains, etc. When one of the trains gets delayed, it causes problems to the connecting journeys. Another major factor to be considered here is, if one train gets late, the other trains who wait for the current train to pass at crossings will also get delayed causing a series of delays. The inability to handle such risks has become very important for the operators to improve the quality of the service provided and meet the passenger's expectations. The proposed system aims at predicting the train delay thereby helping the traveler to plan their journeys ahead of time.

II. BACKGROUND

Transport is an important part of India's economy; a major part of transport considers is covered by railways. Rail transport is one of the major transports used for traveling long distances within a country. In our country, most of the railway operations are taken care of by the state-owned organization, Indian Railways. Indian rail transport is the fourth-largest rail network in the world carrying billion of passengers, and this number of passengers increases annually. Indian trains are overcrowded most of the time due to huge demand and very less supply. The number of railway tracks is also very less in the Indian Railway system when we compare it to the ever-increasing demand. Over the years, trains have been introduced without any additional increase in the infrastructure. The trains come on time, but the delay begins when they enter Bengaluru Division. There is a huge migration of people from undeveloped states to comparatively developed states in search of work. These people travel to their native states in a year to celebrate festivals like Diwali and Holi. This leads to the crowding of the train.

A. *Problem Overview*

Most trains have 21 to 24 coaches and require 540 to 550 meters of space for stabling. There are only 350 meters of land available. The track record shows an average 20% increase in train delays in the past two years and the numbers are still increasing. The average speed of a normal train is ≤90 km/h (which is sadly not achieved in the majority of rail stretches). The reason for this may range from traffic congestion on the stretch to speed restrictions within small intervals. These delays can be classified as Track related delays and Train-dependent delays. Track-related delays are caused by trains that slow down because of some issues in the railway tracks and delays caused by the complete stoppage for a particular time. Congestion and bottlenecks at junctions due to single-line tracks cause track-related delays. At major junctions where tracks from multiple sides converge or diverge, trains of only one route are generally given the green signal, and others are held at the outer. This is because of the way the tracks have been laid. All trains cannot be let in or out

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8325

simultaneously as it will cause a collision at the crossing points where 2 tracks that are going in completely different directions cross each other. Thus, a delay in one train will lead to a delay of multiple trains. This is a chain reaction. The other type of delay that is, the train-dependent delays occur by a train that breaks down or is forced to slow down inline sections. overcrowding in trains due to which trains do not leave the stations on time also causes this delay. There are far too many people who commute and no space for them even after the addition of coaches in a train, so we need more trains, and to accommodate them more tracks are needed.

B. *Objectives*

- To study the historical data of the delays in train commute and to thus predict the train delay
- Using appropriate machine learning models and comparing their performances

III. LITERATURE SURVEY

Train delays are mainly caused by train speed restrictions imposed for safety reasons which can disrupt the train schedules for the connecting journeys. A gradient-boosted regression trees model was used to predict train delay time and the results showed that the trend of train delays can be accurately predicted [1]. In another study, after analyzing the delay propagation mechanisms and the data structure various factors that influenced the results were extracted and used as inputs for their models. It was found that the random forest model exhibits high prediction accuracy and fast call-back in terms of the training model, and ANN, XGBOOST, GBDT, and statistical algorithms are applied as benchmarks for comparison and the results demonstrated that the Random Forest (RF) technique had a good prediction effect [11].

In the railways, delay propagation means that if a delay occurs at any of the starting stations, it usually causes recurrent delays in the next stations too which may even lead to the interruption of the complete railway schedules. Delay prediction model (FCLL-Net) is a hybrid of two types of popular neural network models, namely fully connected neural network (FCNN) and long short-term memory (LSTM) train delays are affected by both operational and non-operational factors. FCLL-Net model considerably outperforms other widely used delay prediction models [10].

Trains in India get delayed frequently by which passengers must face a lot of inconveniences, and if we can predict in advance then it would help the passengers to plan their journeys accordingly. Sometimes people are not able to get the reservation of trains from a source to the destination directly, so the people generally prefer a break journey. The major drawback of break journey is - if they found their first train late then probably, they will miss the second one. 3 different machine learning methods (Multivariate regression, Neural Network, and Random Forest) with different settings to find the most accurate method. The Back Propagation (BPN) feed-forward network provided a good prediction with the least error [5].

The LSTM model is used to build a prediction model of train arrival delay, and the results show that it gives better results than RF and ANN models. The performance of the LSTM model is the highest, which was also proved by the data validation results. its prediction accuracy reaches 86.91% within 30s [8]. The SVR shows small prediction errors and best performance when compared with the random forest, KNN, and ANN [2]. The proposed bi-level random forest predictive model inherently avoids the inconsistencies by training a unique regression model for each of the class labels [4]. The results showed that the prediction accuracy depends on the current delay and type of service for trains that are operational where the ensembles performed better than constituent models as expected [3].

A real-time Bayesian Network (BN) model was proposed to predict both the spatial and temporal propagation of interruptions on train operations which outperformed the regression models as the regression model only captures the influence of the independent factors on the dependent factor [7]. CLF-Net, uses individual factors with unique attributes as input to achieve better performance, the LSTM layers that treat the temporal data as a sequence are developed to capture the temporal dependencies, but the FCNN, which is given non-time-series features, is designed to find out the influence of the static factors. The final simulation showed more accuracy in the CLF-Net than conventional models which are not equipped with the ability to address multi-attribute data which is generated by simultaneously running trains [13].

Bayesian network learning approach was used to identify the delay interactions among the stops as well as to reveal delay propagation patterns in a complex commuter rail network; two network delay metrics were developed which can eventually assist policymakers and practitioners in making investment plans and identifying target stops for improvement [9].

Visualization techniques using the python library are used to better understand the data. The first part was modeled as classes and used the two methods Random Forest and SVM. The authors separated the data into two classes not delayed (delay from 0 to 15 minutes) and delayed (> 15 minutes of delay), to know the factors corresponding to the delay. They were using a Gaussian kernel which proved good for predicting delays, but the training time was very high. In the end, to select the final model cross-validation was finalized. Both SVM and random forest methods led to very good results [12].

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8325

After identifying the major variables that affected the data was put through many machine learning models like Multiple Linear Regression, SVM, ANN, and finally ANN outperformed all others with 78.5% accuracy [14]. The authors have proposed a critical point search algorithm to integrate domain knowledge as an inference engine to categorize the data and find the primary delays. Later, deep learning models were applied to achieve accurate predictions. As a result, the system extracts valuable information, which is directly visualized to the system users for the planning and control rail services at the operational level. The algorithm was then applied in the time series forecasting models to improve the prediction of the primary train delays [6].

IV. METHODOLOGY

A. *Process Flow*

The below flowchart shows the complete flow of the whole project as per how it was carried out. A total of 14 research papers were studied and then the relevant machine learning models were studied. Then the data was cleaned, pre-processed, and used for plotting various types of graphs. Further, dimensionality reduction was done and the whole dataset was converted to numeric to build the models. The accuracy was checked and if it turned out bad, parameter tuning was done and then obtained the results for 5 classification models.

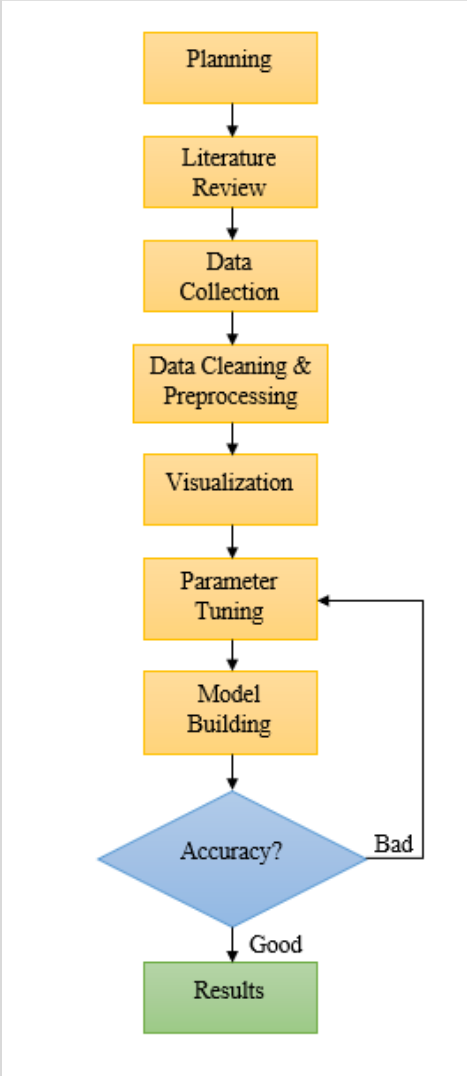

Fig. 1 Process Flow Diagram

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8325

B. *Data Collection and Pre-processing*

Real-time data was obtained from the Indian Railway Department, Southern Division after going through 3 rounds of approvals from the Manager, the Senior Manager, and the Head of Railways, Bengaluru Division. The obtained data consisted of the daily commute of only the delayed trains for the years 2018 and 2019.

Since the obtained data was raw and unprocessed, there were a lot of cleaning processes that had to be done to make the data usable. The actual timetable was obtained through web scraping [15] and [16]. The obtained information was further cleaned and processed and added along with the delay data. The arrival and departure schedule for each train was recorded for all the dates and was added to the data including the dates when the train had arrived on time.

Next, the scheduled time and the delayed time were added which helped to derive the actual time of train arrival. Further comparisons were done and through a few formulas, the target column was marked (0 or 1) by looking at the delay data.

Sr.	Train_No.	Start_Date.	Detention	Det_SubCode	Scheduled_Date	Div	Section_Code	Det_Time	Done By	Remarks
1	6579	27-Apr-18	PATH	INDOOP	27-Apr-18	SBC	BWT-BWT	7	SBC4815	10 mins at BWT home signal for cross movement of 12552 exp.(running late by 245 mins)
2	6579	20-Apr-18	PATH	INDOOP	20-Apr-18	SBC	YPR-JTJ	15	SBC4815	8 mins late out YPR for cross movement of 76525 pass (commuter train), 07 mins at MZV for c/s of 16520 exp.
3	11005	27-Apr-18	PATH	INDOOP	29-Apr-18	SBC	BAND-KJM	5	SBC4815	25 mins GHL precedence for 12080 Exp.
4	11005	29-Jul-18	INC		31-Jul-18	SBC	BSM-BSM	13	SBC4815	20 mins BSM c/s of 12658 exp.REP; T.No 12658 exp loco No 30391, stopped in mid-section b/w BSM-KPN at km 262/18 (SLR), Loco at km 262/1 DJ tripped & found one person on roof of loco got electrocuted & OHE tripped & unable to close DJ. Emergency power block imposed & detrained the person who was ALIVE & moving, he scaled down the loco himself & was entrained in front SLR & arranged 108 ambulance to shift the injured person Hospital. Train arrived spot at 00.07 hrs & left at 00.55 hrs.

Fig. 2. Sample Raw Data

C. *Machine Learning Models*

Machine Learning is an application of Artificial Intelligence that helps the system to learn and improve by using historical data. The main aim of ML is to allow the system to learn by itself automatically through algorithms without the intervention of humans. These ML algorithms are mainly of two kinds: supervised and unsupervised. Supervised learning algorithms can apply their learnings from historical data to predict future events. Unsupervised learning algorithms are used when the information at hand is not labeled. It does not predict or derive the right output. Instead, it explores the data and can draw analysis and inferences which can be used to explain hidden structures from unlabelled data. In this proposed system, we have made use of supervised algorithms like KNN, Logistic Regression, SVM, Decision Trees, and Random Forests.

The KNN algorithm is a very simple and easy-to-use supervised ML algorithm that is used for both classification and regression tasks [17]. It assumes that similar things tend to stay nearby.

Logistic Regression is used to predict categorical dependent variables using a set of independent variables. Since it predicts the output of a categorical dependent variable, the outcome will be a categorical value that is either 0 or 1, true or false, yes or no, etc. This is very similar to linear regression except for the way it is used. Linear regression is used to solve regression tasks, but logistic regression is used to solve classification tasks.

A Decision tree can also be used for both classification and regression but preferred mostly for classification tasks. It splits in the form of a tree to arrive at the classifications. A DT consists of the root node (parent node), internal nodes, branches, and leaf nodes.

On the other hand, the random forest model is an ensemble of multiple decision trees. It creates multiple decision trees and considers the majority votes of predictions among all DTs and predicts the results. More the number of trees means more accuracy and prevents overfitting.

The main goal of SVM is to create the best hyperplane such that the different classes are split exactly on either side of the plane. It chooses the points called support vectors, that help to create the hyperplane.

V. EXPERIMENTAL RESULTS

The experimental results obtained during the analysis are given below:

A.*Graphs*

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8325

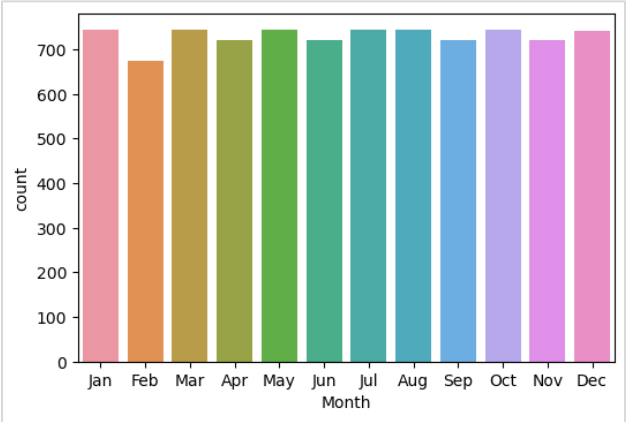

Fig. 3. Month-wise train delay

Fig. 3 shows a count plot between Month and count of delay. As we can see, the trains were delayed most of the time in all the months.

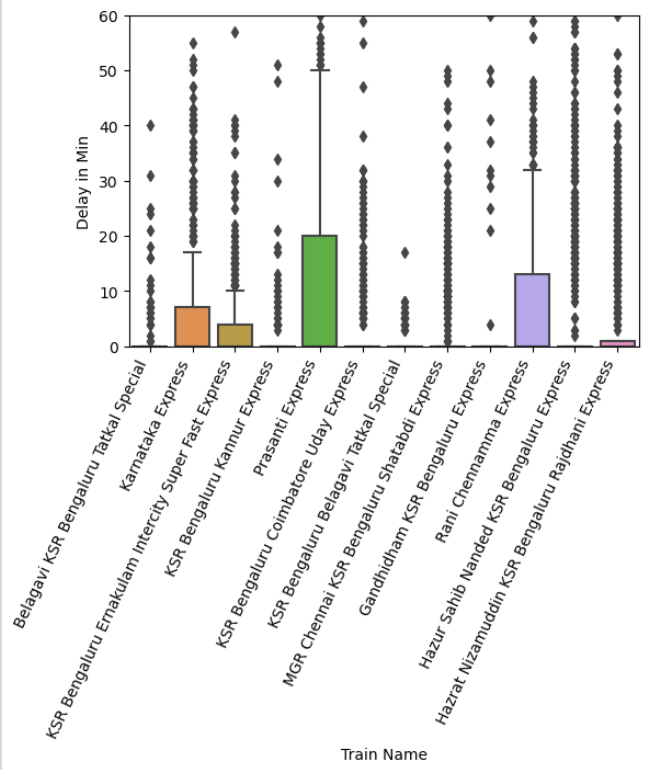

Fig. 4. Train-wise delay in minutes

Fig. 4. Shows a boxplot between Delay in Min and Train Name. It shows which train has delayed the most and which one has delayed for the least amount of time. Also, there are a lot of outliers in the data.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8325

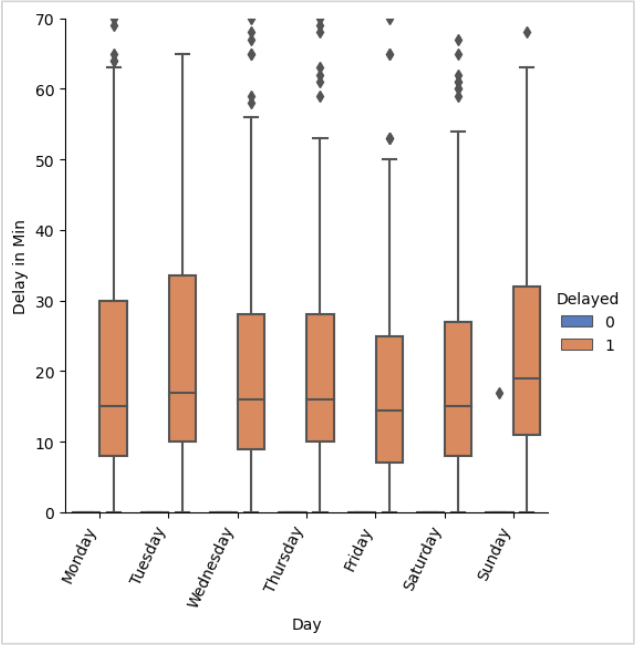

Fig. 5. Delay in minutes on each day of the week

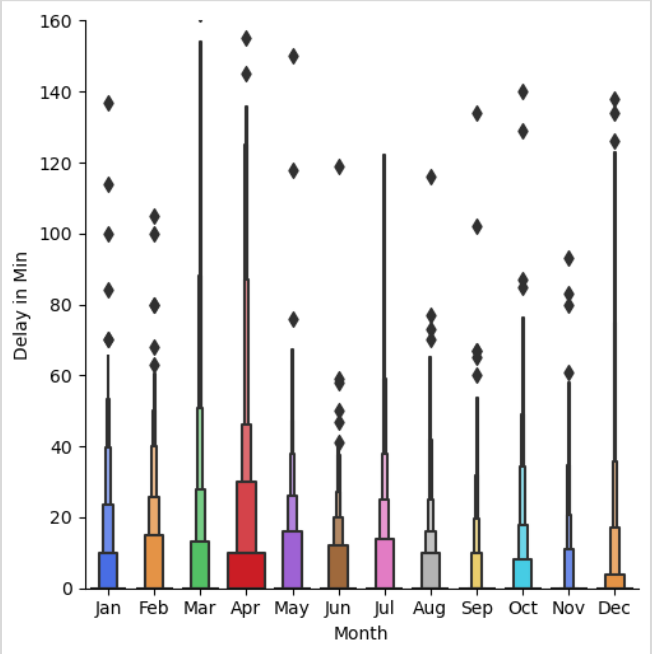

Fig. 6. Month-wise delay in minutes

In this study, after analyzing the delay propagation mechanisms and the data structure, various influencing factors were extracted and used as model inputs. Then, via optimized hyperparameter selection, four algorithms were applied to establish the delay prediction model. The KNN, DT, SVM, Logistic Regression, RF algorithms were used as benchmarks for comparison, and the results demonstrated that the RF technique had a good prediction effect. Finally, the importance of each input feature was calculated and analyzed, and the results revealed that the model was performing well, and the best accuracy was obtained by the RF model.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 3, March 2021

DOI: 10.17148/IARJSET.2021.8325

VI. CONCLUSION

This Project consists of various classification algorithms. The model measures the amount of delay (in minutes). It was shown that the proposed model was successful in providing accurate predictions for delay category and actual delay, and thus can be regarded as a viable approach in situations for which delay prediction involves a coupled classification - prediction task.

Since we have considered Real-world data from Indian Railways it shows that this project will be able to remarkably improve the accuracy of train delay prediction systems and would achieve less error. Due to the nature of the problem, which involved a coupled classification - prediction task, the RF model provided the most accurate predictions for the present dataset of the Indian railway system.

VII. SCOPE AND FUTURE ENHANCEMENTS

We used machine learning methods to predict train delays in India. Even though the study period was very short, we were able to test five models KNN, logistic regression, decision trees, SVM, and random forest methods which led to very good results. The work done in this article can be further continued by using more advanced visualization techniques from the beginning of the predictive analysis process that is, from data collection, preparation to deployment. In this document, we have used only a few machine learning methods, other methods can be applied to get better results.

REFERENCES

[1]. P. Wang and Q. Zhang, "Train delay analysis and prediction based on big data fusion," Transp. Saf. Environ., vol. 1, no. 1, pp. 79–88, 2019, doi: 10.1093/tse/tdy001.

[2]. P. Huang, C. Wen, L. Fu, Q. Peng, and Z. Li, "A hybrid model to improve the train running time prediction ability during high-speed railway disruptions," Saf. Sci., vol. 122, no. October 2019, p. 104510, 2020, doi: 10.1016/j.ssci.2019.104510.

[3]. R. Nair et al., "An ensemble prediction model for train delays," Transp. Res. Part C Emerg.Technol., vol. 104, no. April, pp. 196–209, 2019, doi: 10.1016/j.trc.2019.04.026.

[4]. M. A. Nabian, N. Alemazkoor, and H. Meidani, "Predicting Near-Term Train Schedule Performance and Delay Using Bi-Level Random Forests," Transp. Res. Rec., vol. 2673, no. 5, pp. 564–573, 2019, doi: 10.1177/0361198119840339.

[5]. M. Arshad and M. Ahmed, "Prediction of Train Delay in Indian Railways through Machine Learning Techniques," Int. J. Comput. Sci. Eng., vol. 7, no. 2, pp. 405–411, 2019, doi: 10.26438/ijcse/v7i2.405411.

[6]. J. Wu, L. Zhou, C. Cai, F. Dong, J. Shen, and G. Sun, "Towards a General Prediction System for the Primary Delay in Urban Railways," 2019 IEEE Intell. Transp. Syst. Conf. ITSC 2019, pp. 3482–3487, 2019, doi: 10.1109/ITSC.2019.8916868.

[7]. P. Huang et al., "A Bayesian network model to predict the effects of interruptions on train operations," Transp. Res. Part C Emerg. Technol., vol. 114, no. August 2019, pp. 338–358, 2020, doi: 10.1016/j.trc.2020.02.021.

[8]. C. Wen, W. Mou, P. Huang, and Z. Li, "A predictive model of train delays on a railway line," J. Forecast., no. October 2019, pp. 470–488, 2019, doi: 10.1002/for.2639.

[9]. M. B. Ulak, A. Yazici, and Y. Zhang, "Analyzing network-wide patterns of rail transit delays using Bayesian network learning," Transp. Res. Part C Emerg. Technol., vol. 119, no. June, p. 102749, 2020, doi: 10.1016/j.trc.2020.102749.

[10]. P. Huang et al., "Modeling train operation as sequences: A study of delay prediction with operation and weather data," Transp. Res. Part E Logist. Transp. Rev., vol. 141, no. June, p. 102022, 2020, doi: 10.1016/j.tre.2020.102022.

[11]. Z. C. Li, C. Wen, R. Hu, C. Xu, P. Huang, and X. Jiang, "Near-term train delay prediction in the Dutch railways' network," Int. J. Rail Transp., vol. 00, no. 00, pp. 1–20, 2020, doi: 10.1080/23248378.2020.1843194.

[12]. L. Sara, O. Soumaya, J. Houda, and A. Mohamed, "Predict France trains delays using visualization and machine learning techniques," Procedia Comput. Sci., vol. 175, pp. 700–705, 2020, doi: 10.1016/j.procs.2020.07.103.

[13]. P. Huang, C. Wen, L. Fu, Q. Peng, and Y. Tang, "A deep learning approach for multiattribute data: A study of train delay prediction in railway systems," Inf. Sci. (Ny)., vol. 516, pp. 234–253, 2020, doi: 10.1016/j.ins.2019.12.053.

[14]. C. Jiang, P. Huang, J. Lessan, L. Fu, and C. Wen, "Forecasting primary delay recovery of high-speed railway using multiple linear regression, supporting vector machine, artificial neural network, and random forest regression," Can. J. Civ. Eng., vol. 46, no. 5, pp. 353–363, 2019, doi: 10.1139/cjce-2017-0642.

[15]. "IRCTC Next Generation eTicketing System", Irctc.co.in, 2021. [Online]. Available: https://www.irctc.co.in/nget/train-search. [Accessed: 04- Apr- 2021].

[16]. Railyatri.in. 2021. IRCTC Train Ticket Booking, Indian Railways Train Status & Bus Tickets - RailYatri. [online] Available at: <https://www.railyatri.in/> [Accessed 4 April 2021].

[17]. "Machine Learning Basics with the K-Nearest Neighbors Algorithm", Medium, 2021. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761. [Accessed: 04- Apr- 2021].