# Data-Enabled Proactive Management of Delays in the French Railway Network: A Seasonal Approach

MSc Research Project
Data Analytics

Prachi Mahajan
Student ID: x22158511

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

# National College of Ireland
# Project Submission Sheet
# School of Computing

| | |
|---|---|
| **Student Name:** | Prachi Mahajan |
| **Student ID:** | 22158511 |
| **Programme:** | MSc in Data Analytics |
| **Year:** | 2023-2024 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Prof. Cristina Hava Muntean |
| **Submission Due Date:** | 14 December 2023 |
| **Project Title:** | Data-Enabled Proactive Management of Delays in the French Railway Network: A Seasonal Approach |
| **Word Count:** | 7481 |
| **Page Count:** | 20 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Prachi Mahajan |
| **Date:** | 14th December 2023 |

## PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | □ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Data-Enabled Proactive Management of Delays in the French Railway Network: A Seasonal Approach

Prachi Mahajan

x22158511

**Abstract**

Transportation planning is a critical component of effective urban growth, but traditional methods which are relying on manual procedures such as set schedules, fixed travel routes, and on paper ticketing infrastructure have difficulty keeping up with real-time data and changing passenger demands. In contrast, by examining train delays, their causes and utilising machine learning models such as SVR, ANN, Random forest, Decision tree, the research aims to improve the effectiveness of system. The study makes use of hyperparameter tuning, exploratory data analysis and model evaluation metrics like MSE,RMSE,R2. Using a dataset with transit records from the French transportation network, models predicted delays caused by various factors with excellent accuracy. The created Power BI dashboard allows meaningful data exploration and acted as a useful decision support tool for optimising delay. In findings, the ANN was the most effective model with R-squared value 0.95 which is greater performance in anticipating delays. This demonstrates ANN's strength and applicability for optimising proactive delay strategy in the challenging context of France railway system.

**Keywords:** Support Vector Regressor(SVR),Random Forest(RF), Decision Tree(DT), Artificial Neural Network(ANN).

## 1 Introduction

The transportation business has experienced significant revolutions in recent years, with technological developments and solutions based on data increasingly incorporated into public transit networks around the world. As cities throughout the world struggle with increasing population expansion and urbanisation, the need for efficient and ecologically sustainable transportation solutions is greater than ever. The transportation sector is expanding to satisfy the diversified and dynamic needs of a growing urban population, from extensive public transit networks to new mobility services. Machine Learning (ML) is one notable technology at the leading edge of these advances, with the potential to revolutionise the management and enhancement of urban transit networks.

In terms of public transport in France, the country's railway network is a pillar of both urban and intercity travel. In France, the rail transport landscape has witnessed a strong emphasis on passenger traffic, with high-speed train networks playing an important role in the transportation infrastructure of the nation. As per Wikipedia (Foundation; 2023) France has the second largest rail network in Europe, surpassed only by Russia, with 29,901 kilometres of track, and its passenger infrastructure ranks fifth globally, highlighting the country's reliance on efficient rail transit. However, as per (ConnexionNews; 2023) report there is still much opportunity for improvement in order to improve the effectiveness and efficacy of railway transportation.

Recognising the importance of rail transport in France and the crucial need to improve its efficiency, this **thesis project aims** to use machine learning to forecast delays and optimise railway transit. This study **intends to apply machine learning and data analytics** to address these issues by employing cutting-edge techniques such as ANN and other regression models like SVR, RF, DT, and hopes to modernise the way rail systems run. Research aims to eliminate inaccuracies and improve the whole passenger experience by forecasting delays and adopting strategic optimisations. Project will use data visualisation to improve our understanding of the railway transportation situation in addition to powerful machine learning algorithms. To that aim, an interactive Power BI dashboard is be developed to provide stakeholders with clear insights into the efficiency of the rail network and to facilitate data-driven decision-making.

Study is **motivated by the following research questions**: What are the patterns and seasonality of train delays in France's railway system? How could possibly predictive models such as ANN, SVR be constructed based on previous trends and factors that contribute to delays? How can a responsive Power BI dashboard offering clear insights be used to improve the decision-making process and strategic planning for the French rail system?

This work is organised as follows: Section 2 contains the Literature Review, in which we delve into previous studies to identify knowledge gaps. It establishes the research's foundation by emphasising the relevance of using machine learning to deal with difficulties in railway transportation. Section 3 describes the Research Methodology, which includes methods for integrating the use of machine learning algorithms, data analysis, creating models and the creation of interactive dashboards. This section also covers the assessment measures that will be employed.Section 4 describes the Final model evaluation and results.

The project conclusion, in Section 6, summarises the results and contributions, leading to in an extensive plan for transforming transit to drive beneficial improvements in the area of rail transportation, welcoming more effective, responsive and passenger-centric railway network in France and acting as an important guide for other transit systems worldwide. Through this comprehensive structure, the study attempts to give an in-depth plan for utilising ML and interactive dashboards in order to turn transit systems into functional, data-driven networks.

# 2 Related Work

Predicting High-Speed Train Dispatching Delays The Spatio-Temporal Graph Convolutional Network (TSGCN) created by Zhang et al. (2022) offers a novel delay prediction method that considers the overall effect of delays at each station over a specific time period.The robust TSTGCN model uses deep learning, spatiotemporal attention, and convolution to find complex correlations in high-speed train operational data. Train delay prediction from a real-world train dispatching perspective is a major contribution of the article. MAE, RMSE, and MAPE show that TSTGCN outperforms ANN, SVR, RF, and LSTM.Its methodology and findings can improve train dispatching systems, complementing this study's goals and methods in transportation planning.

(Marković et al.; 2015) used machine learning models to analyse passenger train arrival delays using support vector regression. The study examines the link between these delays and several railway system parameters. Comparison of SVR and ANN in train delay analysis is the main topic, and expert opinions are used to determine infrastructure impact. It analyses train delays at operational and tactical levels using historical data for strategic planning.SVR and ANN models are shown, along with a statistical comparison showing SVR's superior performance. Reviewed research compares SVR and ANN in train delay prediction, whereas this thesis includes additional methods. Both studies attempt to improve transportation network decision-making, but their methods and focus differ, yielding complementary train delay forecasting conclusions.

The study uses Bayesian networks to anticipate train delays in real-time, focusing on the dynamic character of probability distributions (Corman and Kecman; 2018). It uses regularly updated data to improve delay projections and address train activity time uncertainty. This study shows how dynamic stochastic modelling can be used, unlike a previous study that contrasted SVR and ANN. Informed passenger interactions, improved traffic planning, and control with real-time uncertainty removal are stressed. Building standalone applications and extending the concept to larger networks with a focus on computing effectiveness in real-time are suggested research objectives. This work improves train delay prediction and provides valuable insights for railway traffic control.

According to Lessan et al. (2019), a hybrid Bayesian network model for predicting train delays reflects the complex and interconnected structure of rail operations. The research compares heuristic hill-climbing, primitive linear, and hybrid BN methods using high-speed train operations data. The hybrid BN structure beats other models with over 80% accuracy in a 60-minute forecast window, utilising domain knowledge and experts judgements.The research emphasises distinguishing propagated delays from actual operation delays to improve generalizability and model comprehension. This reviewed study and the thesis address rail delays. However, their methods and focus vary. The BN-based model uses Bayesian networks to represent the superposition and interaction effects of train delays, but the current study uses data to develop a seasonal proactive management method.

The "Stochastic modelling of delay propagation in large networks" paper by Büker and Seybold (2012) offers a computationally efficient approach for determining delays in large railway networks. The paper introduces a distribution function, discusses the mathematical procedures, and addresses network procedural theory to appropriately adapt to real settings. Presenting the approach as software highlights its precision and effectiveness. Analytical approach insights, notably computational efficacy and propagation delay modelling, could augment thesis proactive management measures. In line with our thesis's emphasis on proactive delay management's iterative and adaptive nature, the paper's future research emphasises constantly refining and adjusting methods for accuracy and practicality.

The paper (Lapamonpinyo et al.; 2022) addresses passenger train delays by utilising machine learning techniques such as RF, gradient boosting machine, and multi-layer perceptron in prediction models. It examines two data input structures: RWH-DFS and RT-DFS. Ridership, weather, day of the week, location, and prior delay characteristics affect delay prediction models, according to the study. The study introduces and compares two data input architectures and analyses the weight of many external variables in real-time passenger train delay prediction. The RWH-DFS multi-layer perceptron approach performs better. The study found that certain sites require multiple station-to-station prediction algorithms to predict delays. Future studies should examine more adaptive models and factors like goods train data. When comparing this with our thesis, both works use machine learning to reduce train delays, but their parameters, datasets, and methods differ.

(Heglund et al.; 2020) addresses cascading delays in British railways in their study "Railway Delay Prediction with Spatial-Temporal Graph Convolutional Networks". Nonlinear spatiotemporal variable interactions cause cascading delays in the railway network. A unique graph-based formulation of a British railway network aspect is presented, using the Spatial-Temporal Graph Convolutional Network (STGCN) model to predict cascading delays. The model shows that Graph Neural Networks (GNNs) can anticipate delays better than statistical models that don't account for train network interactions. The British railway system, the oldest in the world, is experiencing cascading delays which affect commuters. The study emphasises the need of using the rail network structure to anticipate delays accurately. Future studies may examine the causes and spread of train delays and compare them to current models and different problem formulations.

A study by Wang and Zhang (2019) examines train delay complexities using a three-month dataset of weather, schedule records, and train delays. The study found that severe weather type affects delays during unfavourable conditions, while historical delay durations and frequency are more influential in normal weather. The study uses variables heavily associated with train delays to construct a machine learning model to predict train delays at each stop. This predictive technology helps train operators set better prices and schedules and lets passengers plan more reliable journeys. This research is notable for its long-term prognosis, which gives travellers early itinerary planning information and operators more time for proactive management. Big data fusion helps understand and resolve train delay concerns in the proposed model.

Study addresses unpredictable delays in railway operations by introducing a fuzzy Petri net (FPN) model for train delay estimation Milinković et al. (2013). The FPN model mimics railway traffic and train movements using hierarchy, colour, time, and fuzzy logic. Track segments are places, train motions are transitions, and trains are coloured tokens. Expert knowledge is used to integrate the fuzzy logic system in the absence of delay data, while an Adaptive Network Fuzzy Inference System (ANFIS) is used in systems containing history data. Animating train movement and presenting time-distance graphs validate the simulation. Fuzzy logic is essential for modelling traffic operations' subjectivity, ambiguity, imprecision, and uncertainty, according to the study. Future work will include a fuzzy logic module to handle rail route disputes to the FPN model. In future work, proactive delay management solution for the French railway network may incorporate fuzzy logic components to better handle complexities and uncertainties.

The study by Schlake et al. (2011) evaluates train delays, railcar quality, and the financial impact of railroad rolling stock in-service failures (ISFs). The study prevents ISFs and improves maintenance using automated roadside condition monitoring. ISF-caused train delays cost more than derailment damages, according to the report. Better railcar inspection and maintenance can save costs and increase network productivity, capacity, and reliability. This applies especially to automated technologies. Since unit coal traffic is dense and homogeneous in railcar design, the study evaluates the nonlinear impacts of traffic volume and ISF duration on train delays. The study also shows how peak traffic volume train delays affect rail service quality and are typically disregarded. Improved preventative maintenance saves money, and this study provides a rigorous analytical approach.

The study Chuwang and Chen (2022) uses a time series model to predict daily and weekly passenger demand for urban rail transit (URT) stations. Revenue management, operational strategy, and driving safety are stressed. The project employs Facebook Prophet algorithm and Box-Jenkins time series modelling on historical URT passenger data to improve prediction accuracy. Daily and weekly models reflect COVID-19's holiday and passenger demand effects. The goals are to create parametric models, analyse data, and evaluate forecasting using metrics. For weekly forecasts, ARMA (2, 1) is optimal and SARIMA (5, 1, 3) (1, 0, 0)24 for daily forecasts. Compare to Facebook Prophet, it forecasts daily better than Box-Jenkins weekly. URT station authorities benefit from insights that improve planning and operations. Future research should consider impact elements to increase forecast comprehensiveness, according to the study. It displays the Facebook Prophet algorithm's accurate predicting, improving URT passenger demand forecasts.

A comprehensive Traffic Management System (TMS) is proposed for real-time traffic optimisation in railroads, enhancing traffic fluency across broad networks with varied signalling systems Mazzarello and Ottaviani (2007). TMS analyses train position, speed, infrastructure status, and dynamic elements to predict and fix issues in real time. Key TMS components include Conflict Detection and Resolution (CDR) for real-time train scheduling and routing and Speed Profile Generator (SPG) for plan execution.TMS design, logic, and implementation are explained in the study.TMS capability and real-time pilot experiments show it can handle greater traffic and bottlenecks. Finally, TMS is essential for real-time traffic regulation in large railway networks. TMS architecture's versatility and ability to handle diverse signalling systems make it suitable for many railway network circumstances. Real-time railway operating issues may inspire this proactive delay control thesis.

# 3 CRISP-DM Methodology for Proactive Delay prediction in French Railway Network

This systematic study combines data collection, preprocessing, exploratory data analysis, an interactive Power BI dashboard, and machine learning predictive algorithms to anticipate French railway delays. This thesis follows the CRISP-DM methodological architecture outlined in Figure 1.

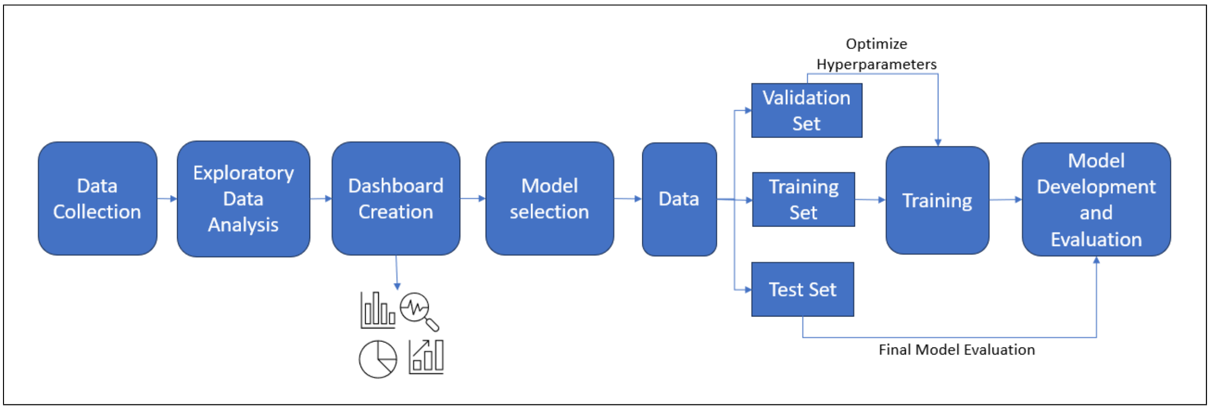


Figure 1: Architectural Overview of Research Approach

## 3.1 Data Collection

As the first step in any research project, data collection is an essential and crucial stage. The selection and collection of an excellent dataset serve as the foundation of the research, which aims to explore the proactive management of delays and resolve the complexity of the French railway network.

### 3.1.1 Data Source:

This study utilised a dataset from SNCF Voyage and Ile-de-France Mobilités, updated by DUBUC (2021), on kaggle, a platform that contains community-contributed datasets. Its reputation for high-quality datasets, transparency, and collaborative study projects made it the data platform of choice. The public accessibility of Kaggle is in line with the open science principles, guaranteeing that the dataset is easily accessible for verification and additional investigation by the researchers.

### 3.1.2 Data Variables:

The dataset contains critical variables related to the French rail transit system, such as train arrivals, departures, delays, and the fundamental causes of these delays. The choice of these factors is essential for the study's purpose. Arrivals and departures of trains provide insight into how the rail network operates on daily basis. The variable tracking delays is especially important since it serves as the focus area for the proactive management method studied in the present research.Furthermore, identifying the reasons of delays provides an in-depth awareness of the factors affecting the train system's efficiency.

### 3.1.3 Temporal Scope:

The dataset has a large temporal scope, spanning from January 2015 to June 2020 and monthly data collection is done for this investigation. This timeframe enables a thorough examination of historical patterns and trends in the French transportation system. Considerations for time-based analysis include identifying seasonal patterns that are influenced by elements such as weather, or other events. Also, the dataset allows for the investigation of long-term patterns, providing significant insights into the train network's past performance and efficiency.

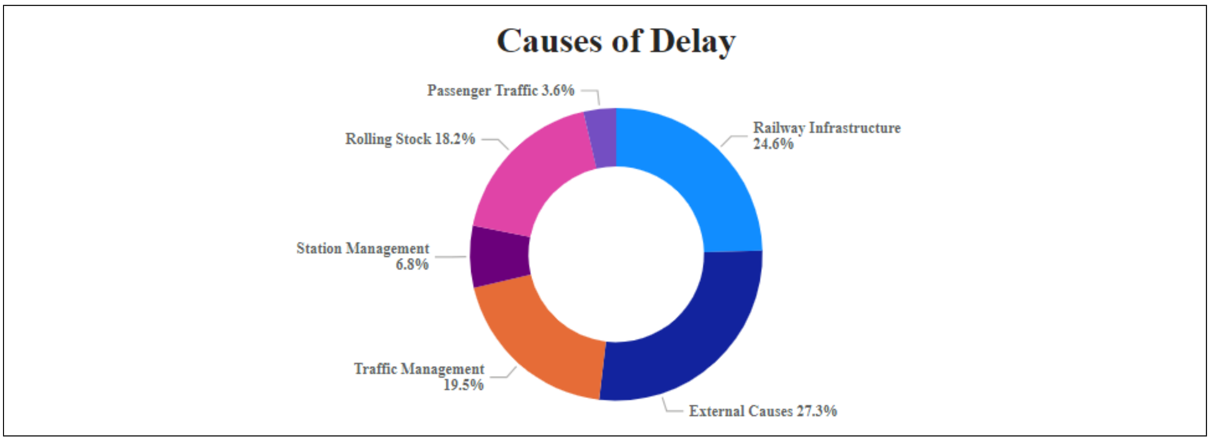


Figure 2: Insights into Causes of Train Delays

Refer to Figure 2 for a brief visual overview of delay causes. The donut chart shows delay reasons proportionally and quickly identifies their main contributors. It shows that 27% of delays are caused by external factors like weather, barriers, suspicious parcels, malevolence, and social movements. Railway infrastructure difficulties, including maintenance and works, are the second primary cause, emphasising the importance of proactive infrastructure management. The remaining categories include rolling stock, traffic management, station management, and passenger traffic, indicating a complete picture of French railway delay causes.

## 3.2 Data Preprocessing and EDA

After obtaining the dataset, the next phases of preprocessing and data cleaning are crucial in ensuring the quality of the data and its relevancy for thorough evaluation. The Python library pandas is a useful tool for this purpose.

### 3.2.1 Data Cleaning:

A rigorous strategy was used to address missing information in order to maintain the dataset's integrity. The dropna method was used to remove records with missing values, generating a revised dataset. Also, two columns for comment section were purposefully dropped during the data cleaning process because they were written in a language other than the primary language and were considered incomplete for the study objectives. This was necessary to simplify the dataset and remove unnecessary information, resulting consistent base for further studies.

Furthermore, numerical and categorical columns was meticulously separated to improve the dataset's quality. To handle any anomalies in the dataset, techniques for outlier detection and management have been implemented. To ensure the robustness outliers were identified and correctly handled. After completing a processing, the dataset which had initially 7806 rows and 32 column was reduced to 7520 rows and 30 columns.

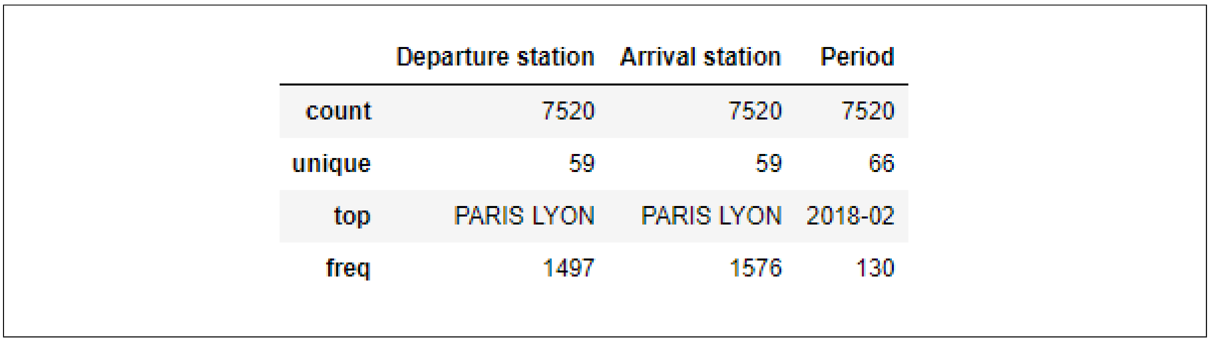
### 3.2.2 Data Standardization and Formatting:

To ensure uniformity and comparability across various aspects, a rigorous data standardization process was implemented. This involved standardising data formats and converting data types as appropriate. Particularly, categorical variables, like Departure station as well as Arrival station, were encoded applying one-hot encoding utilising the get-dummies method, allowing for a more efficient display of categorical data.

The dataset was standardized using the StandardScaler() function from the scikit-learn library to keep consistency among all variables and eliminate possible discrepancy due to diverse data formats. This phase assured that numerical features were standardised to a common scale, preventing specific variables from impacting the models excessively due to variances in size. The standardized dataset not only contributes to a cleaner dataset, but it also helped the modelling process by guaranteeing that the models were developed on data that has a consistent size, thus enhancing their overall effectiveness.

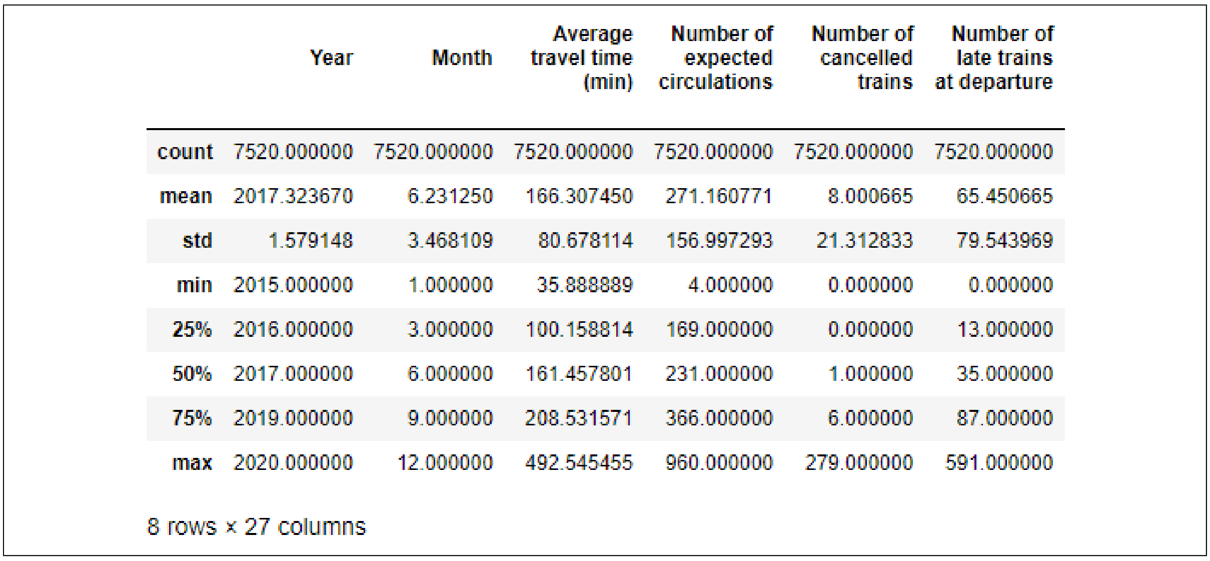
Figure 3 summarises category variables which provides a complete snapshot of the categorical data distribution by summarising key statistics such as count, unique values, top category and frequency.

Figure 4 summarises the statistical measures for numerical variables. This table comprises count, mean, standard deviation,minimum, quartiles, and maximum values, which provide significant insights into the distribution and primary patterns of numerical properties.



| | Departure station | Arrival station | Period |
|---|---|---|---|
| count | 7520 | 7520 | 7520 |
| unique | 59 | 59 | 66 |
| top | PARIS LYON | PARIS LYON | 2018-02 |
| freq | 1497 | 1576 | 130 |

Figure 3: Categorical Variable Overview



| | Year | Month | Average travel time (min) | Number of expected circulations | Number of cancelled trains | Number of late trains at departure |
|---|---|---|---|---|---|---|
| count | 7520.000000 | 7520.000000 | 7520.000000 | 7520.000000 | 7520.000000 | 7520.000000 |
| mean | 2017.323670 | 6.231250 | 166.307450 | 271.160771 | 8.000665 | 65.450665 |
| std | 1.579148 | 3.468109 | 80.678114 | 156.997293 | 21.312833 | 79.543969 |
| min | 2015.000000 | 1.000000 | 35.888889 | 4.000000 | 0.000000 | 0.000000 |
| 25% | 2016.000000 | 3.000000 | 100.158814 | 169.000000 | 0.000000 | 13.000000 |
| 50% | 2017.000000 | 6.000000 | 161.457801 | 231.000000 | 1.000000 | 35.000000 |
| 75% | 2019.000000 | 9.000000 | 208.531571 | 366.000000 | 6.000000 | 87.000000 |
| max | 2020.000000 | 12.000000 | 492.545455 | 960.000000 | 279.000000 | 591.000000 |

8 rows × 27 columns

Figure 4: Summary Statistics for Numerical Variables

### 3.2.3 EDA Techniques:

To uncover trends and insights within the dataset, a comprehensive Exploratory Data Analysis (EDA) was performed. Descriptive statistics, visualisations and category encoding were among the techniques used.EDA provides key insights on train delays, the pattern of delays across various stations, including the effect of categorical factors on delays.

In our EDA, Figure 5 shows count vs. arrival events in our exploratory data analysis (EDA), showing unique trends in arrival station event frequency. The station with the most arrivals is Paris Lyon recording 1576 arrivals. Paris Montparnasse is second-highest having 1050 arrivals. This shows Paris Lyon's importance as a railway hub that handles a lot of trains. Figure 6 shows count vs. departure occurrences, revealing departure station frequency trends. Paris Lyon remains important with 1497 departures and top rank. With 1051 departures, Paris Montparnasse ranks second. Our proactive delay

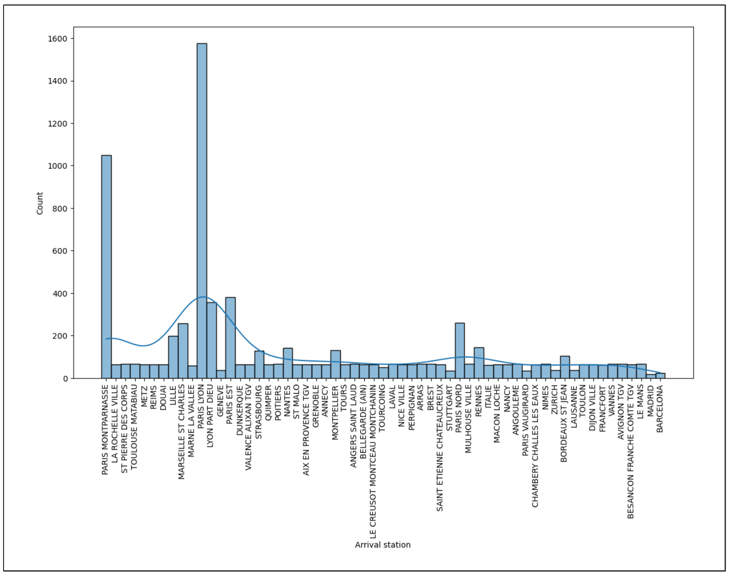


Figure 5: Total Trains by Arrival Station

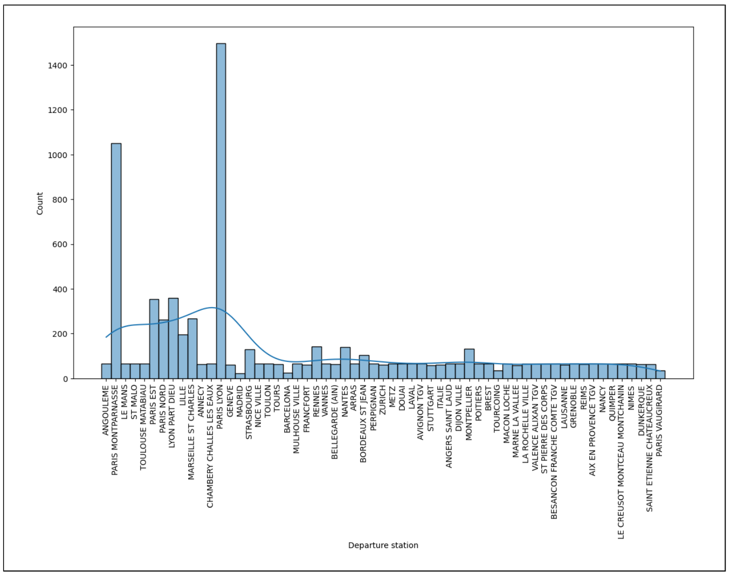


Figure 6: Total Trains by Departure Station

management solutions must focus on these critical stations' departures. This requires taking proactive steps to improve network performance and manage peak counts.

This methodical approach for preprocessing data and EDA guarantees the dataset is prepared for future analysis and evaluation, providing the basis for the use of machine learning models in the examination of proactive delay management across the French railway network.

## 3.3 Dashboard Creation and Analysis

Power BI was selected as the preferred visualisation tool for building interactive dashboards due to its extensive feature set and compatibility with the research goals. With its easy interface, wide range of data connectivity options, and powerful visualisation capabilities, Power BI is an excellent tool for turning complicated datasets into informative representations. Its simple integration with other data sources, such as transit system dataset, made data exploration and analysis more productive.

### 3.3.1 Dashboard Components:

Figure 7 shows interactive dashboard which is composed of multiple vital components that work together to present the French rail transit statistics in an extensive manner.

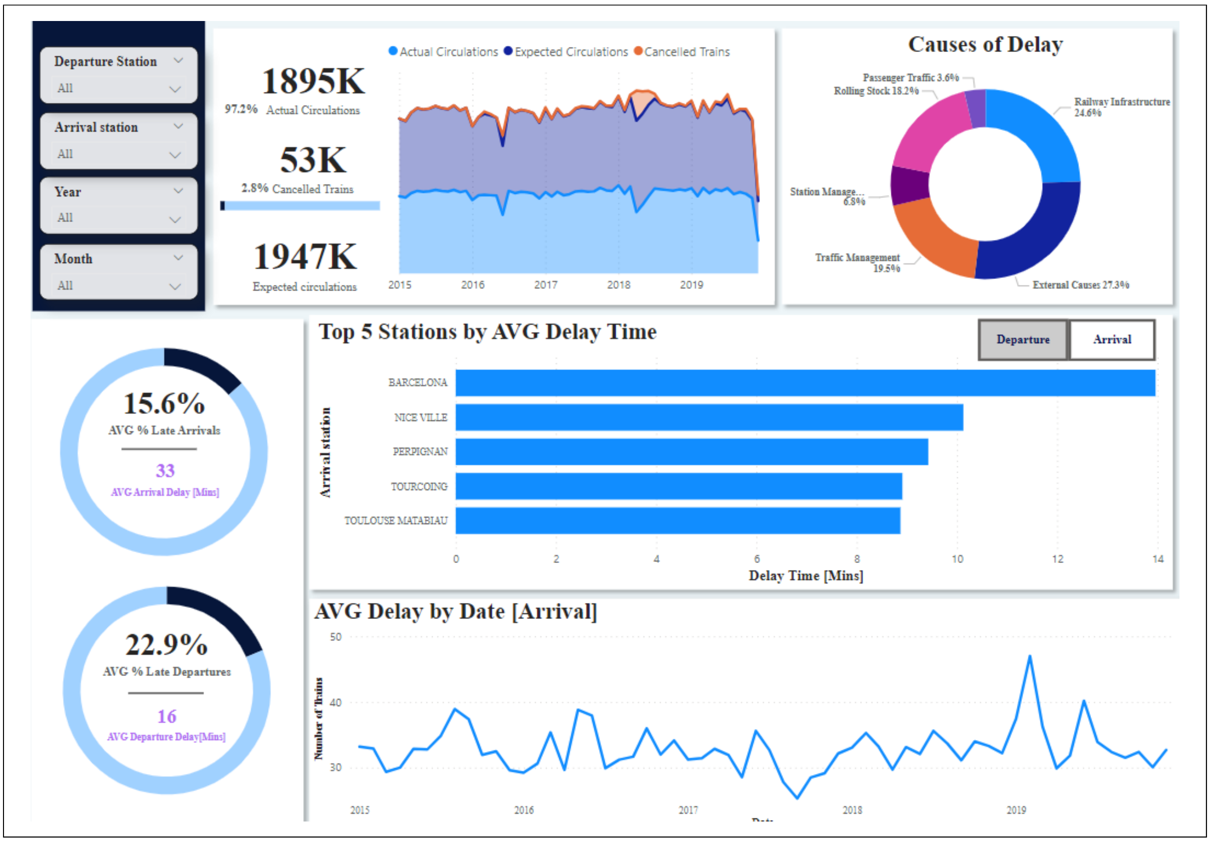


Figure 7: French Railway Network Dashboard

Dashboard components are listed below:

**Interactive slicer:** The interactive slicers allow users to interactively explore the dataset by year (2015-2019), months (1-12), and arrivals as well as departures stations, providing a detailed picture of railway data. In Figure 7, the dashboard covers all years, months, and stations.

**Cards and Area chart:** The cards display actual, expected, and cancelled train counts. The data shows that April 2018 had the most train cancellations, highlighting a period of major railway network problems. These indicators are visually represented over a period by an area chart.

**Donut chart:** The donut chart shows that external variables like the climate, obstacles, suspicious items, malevolence, and social movements account for 27.3% of delay causes across all years which is highest of all causes and and provides understanding of all causes by displaying breakdown in percentage.

**Bar Chart:** The clustered bar chart shows the top 5 arrival and departure stations by average delay time in minutes. Barcelona has the highest arrival and departure delays,

13.96 minutes and 4.29 minutes, respectively. This shows Barcelona's importance in train delay patterns.
**Line chart:** The line chart indicate arrival and departure station average delays by date. The figure shows a peak in arriving train delays in February 2019,47 minutes average. For departure, December 2016 had the longest average delay of 23.60 minutes. These findings highlight delay pattern temporal variability.
**Button:** An interactive buttons for switching among arrival and departure charts improves the dashboard's overall interactivity, enabling users to effortlessly switch between these crucial railway activities.

### 3.3.2 Purpose and Utility:

The requirement to fully comprehend data features and extract insightful knowledge is the driving force behind the choice to give dashboard construction priority. With the help of the interactive dashboard, users may examine correlations, patterns, and trends in the French rail transit dataset visually. Through the utilisation of interactive elements like slicers and visuals, stakeholders are able to comprehend in real time how different aspects affect train operations.

**Reason for Prioritising Dashboards:**
Creating a dashboard to obtain insights from data before constructing machine learning models is a wise and helpful method.It may better comprehend data's characteristics, spot patterns and choose appropriate attributes for machine learning models by visualising and examining it using a dashboard. This procedure is frequently included as part of the larger data discovery and preprocessing phase, and it can assist you in making better choices about feature selection, model selection, and other areas of your analysis.

## 3.4 Model Development

Machine learning is a transformative field within artificial intelligence that empowers computer systems to learn and improve from experience without explicit programming. It revolves around the development of algorithms that enable machines to discern patterns, make predictions and optimize decision-making based on data. In the context of this research, machine learning serves as a pivotal tool, allowing the creation of predictive models to anticipate and manage delays in the intricate dynamics of the French railway network.

### 3.4.1 Model Selection:

The selection of machine learning models was thoughtfully considered while creating a predictive framework for train delays within the complex environment of the French railway system. Every selected model contributes a distinct set of advantages that are consistent with the complexities and nonlinearity involved in forecasting train delays that are impacted by multiple dynamic variables. The Support Vector Regressor (SVR), Random Forest, Decision Tree Regressor, and Artificial Neural Network (ANN) are among the models in the ensemble.

**Support Vector Regressor** is a particularly important option because of its exceptional capacity to manage non-linear interactions in the dataset. SVR is a suitable option for modelling such non-linear dependencies since it makes use of support vectors in a high-dimensional feature space.

**Random Forest** is an ensemble learning technique that combines predictions from several decision trees, it is used. This approach is notable for its high accuracy and resilience to overfitting, which makes it a good choice for identifying complex correlations in the data. More consistent and dependable forecasts can be produced by RF by combining the outputs of several trees. Train delays are frequently influenced by a variety of factors and the randomization generated by integrating predictions from different decision trees in a RF reduces the danger of overfitting.

**Decision tree Regressor** was chosen for its adaptability in capturing complicated relationships within data via a hierarchical structure of decision nodes. This paradigm was chosen specifically for its interpretability and simplicity. Decision trees illustrate a series of decisions depending on input features, allowing for a more transparent decision-making process. When anticipating train delays, it is critical not only to acquire accurate predictions but also to understand the elements that contribute to those estimates. The hierarchical structure of Decision Tree Regressor allows stakeholders to clearly grasp the decision-making process, providing insights into the precise reasons causing train delays.

**Artificial Neural Network** As a deep learning model, this is chosen for its ability to learn complicated patterns in vast datasets. Because ANNs excel at capturing non-linear interactions, they are highly suited to complicated problem domains. ANN models combine elements of learning and adaptation as per Palit and Popovic (2006). ANNs can identify hidden patterns within data by exploiting the depth and interconnection of neural networks, helping to a more accurate and complete prediction of train delays in the French railway system.

### 3.4.2 Hyperparameter Tunning:

Optimising the effectiveness of each model required a vital step, which is hyperparameter tuning method.In order to improve the models' capacity to identify complex trends in the French rail transportation dataset, particular hyperparameters have been carefully chosen and adjusted.
**Random Forest:** Hyperparameters tunning for Rnadom forest is as follows:
N_estimator : Number of trees in forest.
Given parameters for n_estimator was [50,100,150,200,250] in which optimal parameter is 50. It means model performed best with 50 trees in forest.

**Impact on model:** Finding the ideal tree count in the forest requires fine-tuning n_estimators. This parameter has a direct impact on how well the model generalises and how well it can capture complex relationships in the data.
**Decision tree:** Hyperparameters tunned for decision tree are mentioned in following Table 1 which contains all considered parameters values and optimal values.

Description of parameters used in decision tree are as follows:
Splitter : The method used to select the split at every node.
max_depth: The max depth of tree.
min_samples_leaf: The minimum number of samples needed at leaf node.
max_features: The total number of features to take into account when choosing the ideal split.
max_leaf_nodes: Use the best-first method when growing trees with max_leaf_nodes.

**Impact on model:** The depth and layout of the decision tree are greatly affected by

| Grid Parameters | values | Optimal parameter |
|---|---|---|
| Splitter | [best,random] | best |
| max_depth | [2,4,6,8,10,12,14] | 14 |
| min_sample_leaf | [1,2,3,4,5,6,7,8,9] | 4 |
| max_feature | [sqrt,log2,None] | None |
| max_leaf_nodes | [None,10,20,30,40,50,60,70,80,90] | None |

Table 1: Parameters - Decision Tree

the tuning of these hyperparameters. It has a direct impact on the clarity, generalizability, and preventive overfitting of the model.
**Support Vector Regressor:** Refer Table 2 for Hyperparameters tunned for SVR which contains all considered and optimal parameters.

| Grid Parameters | values | Optimal parameter |
|---|---|---|
| estimator_kernel | [[linear, poly, rbf, sigmoid] | linear |
| estimator_C | [10, 100, 1000, 10000] | 10 |
| estimator_epsilon | [0.1, 0.01, 0.001] | 0.001 |

Table 2: Parameters - SVR

Description of parameters used in SVR are as follows:
kernel: Specifies the kernel type.
Estimator_C: Regularization parameter.
Estimator_epsilon: a tolerance gap in which errors are not penalised.

**Impact on model:** It is essential to adjust these hyperparameters in order to shape the SVR's capacity to detect non-linear patterns. It is essential for optimising the SVR's prediction capabilities while efficiently controlling the model.
**Artificial Neural Network:** ANN hyperparameter tuning method was a critical step in optimising its predicting performance. Various combinations of hyperparameters, and the best configuration was determined based on multiple scoring metrics as mentioned in Table 3 below.

| Grid Parameters | values | Optimal parameter |
|---|---|---|
| Layers | [20], [40, 20], [45, 25], [45, 30, 15], [40, 30, 20, 10] | [45, 30, 15] |
| Activation Function: | [sigmoid,ReLu] | ReLu |
| Batch Size | [128, 256] | 128 |

Table 3: Parameters - ANN

Description of parameters used in SVR are as follows:
Layer: The neural network's architecture, which specifies the total number of nodes in every layer.
Activation function: A function that is applied to each node in order to introduce non-linearity into the model.
Batch size: The amount of samples processed prior to adjusting the weights of the model.

Epochs: During training, the number of times the full dataset passes forward and backward across the neural network. To optimise parameters, the model is trained over 30 epochs.

**Impact on model:** The neural network's design was altered by adjusting the layers and activation function, allowing it to discover detailed patterns in the data.The use of ReLU as the activation function improves the network's ability to detect non-linear correlations. For training, a batch size of 128 and 30 epochs were determined to be best, balancing computational efficiency with model convergence.

The success of each model is tightly linked to the exact hyperparameters set, with the changes directly impacting their capacity to generalise, interpret outcomes, and capture complicated patterns in data.

### 3.4.3 Model Evaluation Metrics:

Choosing the right evaluation metrics is critical for analysing predictive model performance in the context of delay forecasting in the French rail network. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) are the metrics chosen because they provide complete insights into the models' accuracy and capacity to identify variations in delay patterns.

**Mean Squared Error(MSE):** The average squared difference between predicted and actual delays is calculated by MSE. MSE gives a quantitative measure of the algorithm's precision, demonstrating the size of errors, which makes it particularly important for evaluating the overall precision of delay forecasts.Stewart (2023)

**Root Mean Squared Error (RMSE):** The square root of MSE is RMSE, which represents the average magnitude of errors in the identical unit as the target measure.It provides a more transparent measure of errors in prediction and is sensitive to major errors, allowing for a more balanced evaluation of the model's effectiveness.(C3AI; 2021)

**R-Squared (R2):** R2 estimates the amount of the variance in a dependent variable (delay) that can be predicted by the independent variables (features). R2 measures the model's efficacy in terms of fit, or how well it represents the variability in the data. A greater R2 indicates a better fit, which improves the clarity of the model's forecasting abilities.Agrawal (2023)

**Cross-Validation and Refit:** During the evaluation phase, a 5-fold cross-validation (cv=5) technique was used to achieve robust model evaluations. This involves dividing the dataset into 5 subsets, training the algorithm on four of them, and assessing it on the fifth, and then repeating the procedure five times. Furthermore, the 'r2' measure was selected for refitting ('refit = 'r2'), emphasising model optimisation based on R2 scores. This method improves the model assessments' quality and generalizability.

When it comes to forecasting train delays in the French railway system, accuracy and interpretability are critical. The prediction error magnitude can be clearly understood using MSE and RMSE, and the model's capacity for defining the variability in the delay data can be recognised using R2. The results of this analysis support the goal of the research, which is to develop precise and understandable models to support proactive decision-making and management in the transportation industry.

### 3.5 Train-test Split

The dataset was systematically partitioned into testing and training sets in order to measure the generalization accuracy of the machine learning models. Splitting is an important step in ensuring that the models are tested on unknown data, offering an accurate representation of their prediction ability.

The dataset was divided using the commonly used train-test split methodology, using the scikit-learn library's train_test_split function. This approach involves using part of the data for training the algorithms and the remaining data for evaluating their ability to perform. The following are the specifics of the train-test split:

**Features and Target Variables:** Prior the split, features were standardised with a StandardScaler to ensure uniform scaling throughout the training and testing sets.

**Split Ratio:** The dataset was divided into training and testing sets (X_train, Y_train), with a test size of 20% (test_size=0.2).This proportion was chosen to find a balance among training data and independent set to evaluate model performance.

**Random State:** To ensure reproducibility, a random seed of 1 (random_state=1) was set, enabling an identical split to be produced every time the algorithm is executed.

This partitioning technique allows for robust model development on a specific portion of the data with extensive testing on another, resulting in an accurate evaluation of each model's predictive ability.

## 4 Model Evaluation and Selection

This section provides a complete summary of the final assessment of models, building on the evaluation metrics presented in the model selection section. The importance of key performance indicators such as Mean Squared Error (MSE), Root Mean Squared Error(RMSE) and R-squared (R2) score will be emphasised, and their implications for each model will be thoroughly examined. This comprehensive investigation seeks to provide conclusions about the predictive abilities of the chosen models and to find the most robust method for forecasting delays in the French transportation system.

### 4.1 Model Performance Overview

A summary of each model's primary performance indicators, comprising the R-squared, MSE and RMSE scores, are shown in Table 4. The table gives a thorough comparison, assisting in the evaluation of the model's performance across several evaluation criteria.

| Models | R squared | MSE | RMSE |
|---|---|---|---|
| **Random Forest** | 0.78 | 26.91 | 5.18 |
| **Decision Tree** | 0.84 | 13.38 | 3.65 |
| **Support Vector Regressor** | 0.98 | 2.23 | 0.00047 |
| **Artificial Neural Network** | 0.95 | 5.35 | 2.30 |

Table 4: Models Evaluation Metrics

## 4.2 Comparative Analysis of Model Performance

Each model displays distinct advantages and factors to be taken into account in the thorough evaluation of model performances, providing insightful information about the models' suitability for the particular study setting.The Figure 8 below depicts a visual representation of the comparison of model performance, with a specific emphasis on the R$^2$ values.

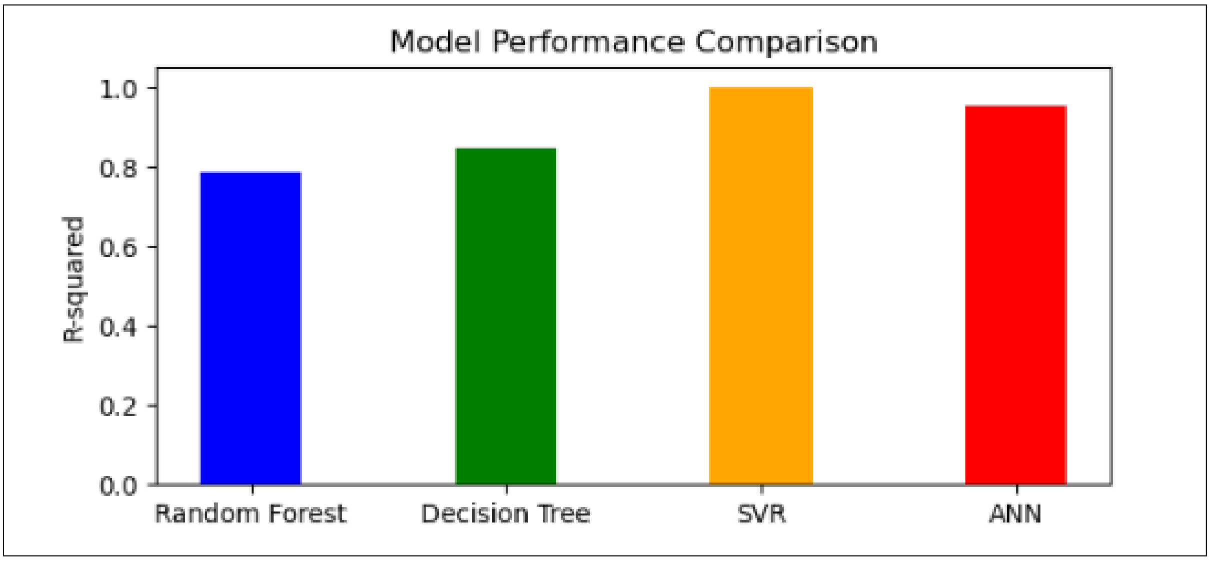


Figure 8: Model Performance Comparison

Both the Random Forest and Decision Tree models excel at capturing the subtle seasonal fluctuations in French railway delays. The RF provides a balanced trade-off between pattern recognition and precision, with a best mean cross-validated R-squared of 0.78 and an RMSE of 5.18. The Decision Tree, on the other hand, has an amazing R-squared of 0.84 and an RMSE of 3.65, demonstrating a robust capacity to capture variations.

With an almost perfect R-squared of 0.98 and negligible MSE and RMSE (2.23 and 0.00047, respectively), SVR emerges as an extraordinary performer. Its ability to model complex relationships within delay data is outstanding. However, its processing requirements may offer scaling issues in the vast French railway network. The use of SVR in real-time should be carefully examined, especially in circumstances where quick decision-making is required for delay control.

With an high R-squared of 0.95, a competitive MSE of 5.35, and an RMSE of 2.30, the ANN stands out. Because of its deep learning architecture, it can detect both linear and nonlinear correlations in delay data. Because of its ability to adapt to changing patterns and scalability, the ANN is well-suited to the dynamic character of the French Railway Network. While the ANN's training complexity is recognised, its capacity for precise and scalable forecasts makes it a viable tool for real-time decision assistance.

In the context of the French Railway Network, where delays vary seasonally, the ideal model is determined by the specific operational requirements. The trade-offs between pattern identification, accuracy, and computational effectiveness should be considered. Whether the goal is rapid decision-making or a deeper knowledge of delay patterns, the model chosen should be consistent with the overall objective of building an effective delay management system customised to the unique problems of the French railway landscapes.

## 4.3 Justification for Model Selection

The choice of an appropriate model is critical to the effectiveness of our proactive delay management solution for the French Railway Network. We thoroughly evaluated the advantages and drawbacks of each model when analysing performance indicators across several models like RF, DT, SVR, and ANN. Several major considerations support the decision to use the ANN as the final model.

With a Best MSE score of 5.35, RMSE score of 2.30, and R2 score of 0.953, the ANN model performed excellently. These indicators show a high level of accuracy and predictive ability, which is consistent with our primary goal of accurately projecting delays in railway operations. Because of the model's capacity to capture complicated patterns in data, as well as its stable performance in both training and validation sets, it is well-suited to the complex structure of railway systems.

The choice of the ANN over the SVR is driven by factors other than R-squared values. Although SVR had a higher R-squared value, we preferred ANN because of the nuances of train delays. Train delays are complicated; they involve not only simple relationships but also intricate, non-linear dependencies.This consideration is consistent with our objective of not only achieving numerical advantage but also making sure the chosen model is compatible with the complexities that accompany railway operations.

Furthermore, the ANN model's scalability distinguishes it as a forward-thinking option capable of accommodating the evolving nature of the French Railway Network. As railway operations become more complex, the adaptability of the ANN becomes a significant advantage, enabling effortless integration with future improvements and expansions.

While different models, like RF and DT, performed admirably, the ANN's better precision, adaptability, and complex pattern recognition make it the best fit for our proactive delay management strategy.

## 4.4 Results on Final Model

### 4.4.1 Quantitative Metrics:

The concluded Artificial Neural Network (ANN) model exhibited excellent results on the test dataset as well, indicated by the subsequent quantitative metrics:

R-squared: With an outstanding value of 0.9665 for the coefficient of determination (R2), the model is able to account for almost 96.65% of the variance in the test data.

MSE: The model demonstrates its accuracy in identifying the deeper trends within the test dataset by achieving a small mean squared variance between the predicted and actual values, as shown by its Mean Squared Error (MSE) of 3.8982.

RMSE: The Test RMSE, which is 1.9569, provides the average magnitude of the errors, which adds to the model's precision. A closer match between the expected and actual values is shown by a lower RMSE.

### 4.4.2 Visualization of Results:

Enhancing the numerical measurements, the scatterplot visualisation, namely ANN : Predicted Vs.Actual Causes of Train Delay in France shown in figure 9, is used to evaluate the accuracy of the final model that is developed to forecast delays caused by different factors. The scatterplot indicates actual delay on the x-axis and corresponding predicted delays on the y-axis, accompanied by a red dotted line indicating accurate predictions.

The scatterplot showed that most blue data points aligned diagonally along the red dotted line. This alignment indicates a strong correlation between projected and actual delay values, indicating that the model is successful in capturing data trends.

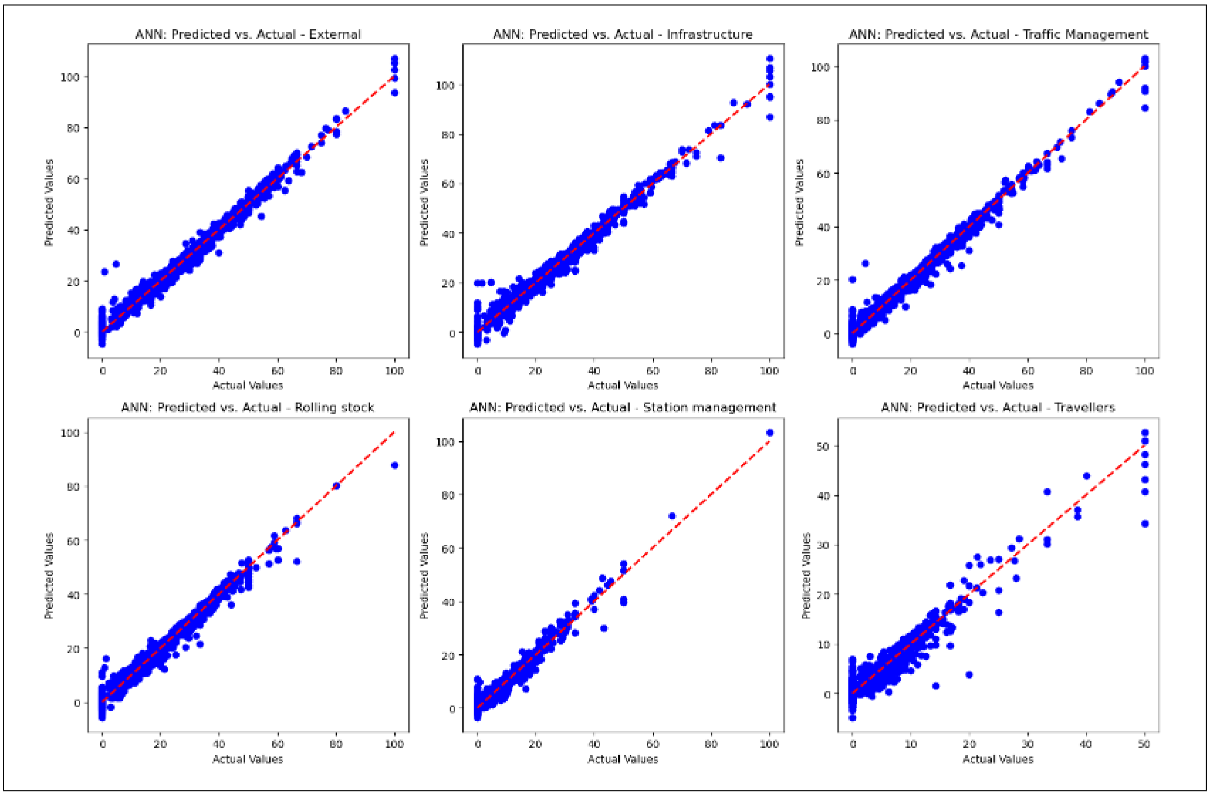


Figure 9: ANN : Predicted Vs. Actual Causes of Train Delay in France

To summarize, the scatterplot findings show a good correlation between projected and actual delay values, with a few deviations worth investigating. Considering real-world data's variability, the model's performance is adequate, and these findings provide significant insights for future modifications.

# 5 Conclusion

In conclusion, this research into delay management in the French Railway Network resulted in a robust framework that leverages the capabilities of modern machine learning models, most notably the ANN. The combination of these models provides a novel approach to proactive delay prediction and administration. The accompanying dashboard, which is designed to provide railway authorities with an intuitive visualisation of delay trends and model efficiency, serves as a critical decision support tool.

The model comparison, which includes Random Forest, Decision Tree, SVR, and ANN, emphasises the ANN's superiority in capturing the complex, nonlinear relationships inherent in railway delay data. While the SVR performed admirably computationally, the ANN's adaptability to nuanced patterns distinguishes it as an integrated approach aligned to the dynamic complexities of railway operations. This comprehensive structure not only tackles the immediate challenges of delay management, but also opens the way for a data-enabled, proactive method for optimising the operational efficiency of the French Railway Network.

# 6 Future work

The future holds exciting opportunities for improving and broadening our approach. integrating additional information from sources, such as weather and maintenance schedules, can improve our models' predictive capabilities. Furthermore, investigating ensemble models that combine the best qualities of multiple algorithms may improve accuracy even further.Constant collaboration with railway stakeholders, as well as the incorporation of real-time data streams, can help keep our models adaptable to changing operational scenarios. The envisioned future includes an effortless incorporation of our predictive algorithms into everyday operations of the French Railway Network, which will contribute to a more flexible and responsive rail transportation system.

# References

Agrawal, R. (2023). Know the best evaluation metrics for your regression model,analytics vidhya.
**URL:** *https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/h-r-squared-r2*

Büker, T. and Seybold, B. (2012). Stochastic modelling of delay propagation in large networks, *Journal of Rail Transport Planning Management* **2**(1): 34–50.
**URL:** *https://www.sciencedirect.com/science/article/pii/S2210970612000182*

C3AI (2021). Root mean square error (rmse).
**URL:** *https://c3.ai/glossary/data-science/root-mean-square-error-rmse/*

Chuwang, D. D. and Chen, W. (2022). Forecasting daily and weekly passenger demand for urban rail transit stations based on a time series model approach, *Forecasting* **4**(4): 904–924.
**URL:** *https://www.mdpi.com/2571-9394/4/4/49*

ConnexionNews (2023). French train delays last year among the worst seen in a decade.
**URL:** *https:https://www.connexionfrance.com/article/French-news/French-train-delays-last-year-among-the-worst-seen-in-a-decade*

Corman, F. and Kecman, P. (2018). Stochastic prediction of train delays in real-time using bayesian networks, *Transportation Research Part C: Emerging Technologies* **95**: 599–615.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0968090X18311021*

DUBUC, G. (2021). Public transport traffic data in france.
**URL:** *https://www.kaggle.com/datasets/gatandubuc/public-transport-traffic-data-in-france*

Foundation, W. (2023). wikipedia.
**URL:** *https://en.wikipedia.org/wiki/Rail*$_t$*ransport*$_i$*n*$_F$*rance*

Heglund, J. S., Taleongpong, P., Hu, S. and Tran, H. T. (2020). Railway delay prediction with spatial-temporal graph convolutional networks, *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6.

Lapamonpinyo, P., Derrible, S. and Corman, F. (2022). Real-time passenger train delay prediction using machine learning: A case study with amtrak passenger train routes, *IEEE Open Journal of Intelligent Transportation Systems* **3**: 539–550.

Lessan, J., Fu, L. and Wen, C. (2019). A hybrid bayesian network model for predicting delays in train operations, *Computers  Industrial Engineering* **127**: 1214–1222.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0360835218301025*

Marković, N., Milinković, S., Tikhonov, K. S. and Schonfeld, P. (2015). Analyzing passenger train arrival delays with support vector regression, *Transportation Research Part C: Emerging Technologies* **56**: 251–262.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0968090X1500145X*

Mazzarello, M. and Ottaviani, E. (2007). A traffic management system for real-time traffic optimisation in railways, *Transportation Research Part B: Methodological* **41**: 246–274.

Milinković, S., Marković, M., Vesković, S., Ivić, M. and Pavlović, N. (2013). A fuzzy petri net model to estimate train delays, *Simulation Modelling Practice and Theory* **33**: 144–157. EUROSIM 2010.
**URL:** *https://www.sciencedirect.com/science/article/pii/S1569190X12001700*

Palit, A. K. and Popovic, D. (2006). *Computational intelligence in time series forecasting: theory and engineering applications*, Springer Science & Business Media.

Schlake, B. W., Barkan, C. P. L. and Edwards, J. R. (2011). Train delay and economic impact of in-service failures of railroad rolling stock, *Transportation Research Record* **2261**(1): 124–133.
**URL:** *https://doi.org/10.3141/2261-14*

Stewart, K. (2023). mean squared error. encyclopedia britannica.
**URL:** *https://www.britannica.com/science/mean-squared-error*

Wang, P. and Zhang, Q.-p. (2019). Train delay analysis and prediction based on big data fusion, *Transportation Safety and Environment* **1**(1): 79–88.
**URL:** *https://doi.org/10.1093/tse/tdy001*

Zhang, D., Peng, Y., Zhang, Y., Wu, D., Wang, H. and Zhang, H. (2022). Train time delay prediction for high-speed train dispatching based on spatio-temporal graph convolutional network, *IEEE Transactions on Intelligent Transportation Systems* **23**(3): 2434–2444.

**Question 1: Briefly present the limitations of your research work.**

**Dependency on Historical Patterns:**

We can identify trends in train delays by using previous data. It can be challenging, though, to handle unforeseen problems that have never occurred before. It may be difficult for the models to anticipate these new issues, which makes it more difficult to handle unusual circumstances and the most recent difficulties with delay management.

**Data quality and Availability:**

The quality and availability of past railway delay data are critical factors that affect the predictive models' accuracy. The models' capacity to generate trustworthy predictions may be jeopardised by missing or erroneous data in the datasets, underscoring the importance of correct and thorough data for the best possible model performance.

**Continuous Monitoring Required:**

continuous monitoring is crucial for adapting our predictive models to the evolving dynamics of the railway system. As external influences and operational conditions change over time, the models may require ongoing adjustments to ensure their continued relevance and effectiveness in mitigating delays.

**External Influences:**

External factors that could affect our research, like changes in regulations or unforeseen events, could make our prediction models less accurate. These variables, which aren't stated clearly in past data, might affect how accurately the models forecast delays, highlighting the necessity of adaptability and flexibility in the face of unforeseen events.

**Question 2 : Justify why those models were investigated in your research.**

**Complex interactions**: A number of factors can impact complex and nonlinear interactions that are responsible for railway delays. To capture the complicated dependencies within the data, we selected ANN and Random Forest because of their reputation for managing complex patterns.

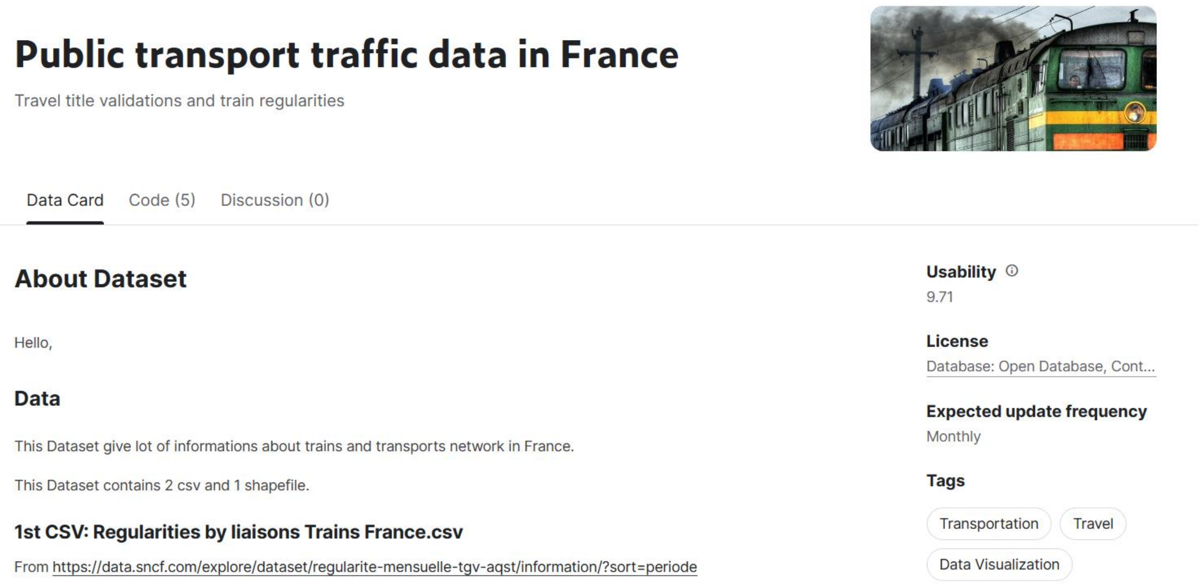
**Flexibility of random forest:** Random Forest's ability to handle both numerical and categorical variables makes it ideal for regression applications. Given the diversity of railway data, this model is adaptable to a variety of variables.

**Robustness of SVR:** SVR can capture both linear and nonlinear interactions, was used to solve the complexities of railway delays. The model's ability to handle varied patterns makes it appropriate for our research area.

**Ensemble Learning strengths:** Random Forest, as an ensemble learning technique, mixes many models to reduce the danger of overfitting while enhancing overall predictive performance. This feature is useful in improving the robustness of delay forecasts.

**Availability of data:** The availability and structure of the dataset were taken into account when selecting the models. The chosen models were able to manage the data's multilinear and temporal features in a way that was consistent with our study's objectives.

**Question 3: Sncf has open data platform for train data. was it considered for taking latest train data, why not? what was the reason to use data from 2015-2019 only?**



Public transport traffic data in France

Travel title validations and train regularities

About Dataset

Hello,

Data

This Dataset give lot of informations about trains and transports network in France.

This Dataset contains 2 csv and 1 shapefile.

1st CSV: Regularities by liaisons Trains France.csv

From https://data.sncf.com/explore/dataset/regularite-mensuelle-tgv-aqst/information/?sort=periode

Usability
9.71

License
Database: Open Database, Cont...

Expected update frequency
Monthly

Tags
Transportation
Travel
Data Visualization

When picking data for study, I prioritised the most recent and complete information accessible from 2015 to 2020. However, I noticed a difficulty with the 2020 data, which only spanned six months. Recognising the potential impact of skewed results and uneven distribution associated with such a short timeframe, I decided to exclude 2020 data from the analysis.

It is critical to have a balanced and representative dataset in order to gain significant insights. Furthermore, I would like to emphasise that the dataset used in this study, obtained from Kaggle, is a true reproduction of the information available on the SNCF portal, demonstrating the data's dependability and authenticity in our research.

**Question 4: "The use of SVR in real-time should be carefully examined, especially in circumstances where quick decision- making is required for delay control." explain the rationale for this statement.**

"The use of SVR in real-time should be carefully examined, especially in circumstances where quick decision-making is required for delay control," arises from the complex challenges created by Support Vector Regression (SVR) in the context of my study on proactive delay management in the French Railway Network. One key concern is the complex process of fine-tuning SVR parameters, which is required for optimal performance.

However, fine-tuning can be very difficult, and in real-time scenarios, the necessity for models with well-optimized parameters is critical to ensuring accurate and timely forecasts.

The complexity imposed by SVR's robustness and ability to capture complicated correlations in data may be a burden in instances where quick choices are required. In the dynamic environment of railway operations, where unexpected events and changes in conditions like accidents, protests necessitate quick response, the trade-off between model complexity and responsiveness is critical. As a result, a thorough evaluation of SVR in real-time applications is required, taking into account the complexities involved with parameter tuning and the need for models that can quickly adjust to changing conditions for effective delay management.