DATA MINING FOR FINDING RAILWAY TRACK UTILIZATION

MINOR PROJECT REPORT

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY
Computer Science Engineering

Submitted By:
Madikanti Naveen
(2020BTechCSE043)

Submitted To:
Dr. Kedar Sharma
Associate Professor

Department of Computer Science Engineering
Institute of Engineering and Technology
JK Lakshmipat University Jaipur

December 2023

CERTIFICATE

This is to certify that the minor project work entitled "**DATA MINING FOR FINDING RAILWAY TRACK UTILIZATION**" submitted by **Madikanti Naveen (2020Btechcse043)** towards the partial fulfillment of the requirements for the degree of **Bachelor of Technology in Computer Science Engineering** of JK Lakshmi pat University Jaipur is the record of work carried out by them under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for minor project examination.

Dr. Kedar Sharma

Faculty Guide

Department of Civil Engineering

Institute of Engineering & Technology (IET)

JK Lakshmipat University Jaipur

Date of Submission: December 4, 2023

Acknowledgments

I am highly grateful to **Dr. Renu Jain**, the Director of the Institute of Engineering and Technology (IET), **JK Lakshmi pat University** -Jaipur, Rajasthan for providing this opportunity to carry out the major project work.

The constant guidance and encouragement received from **Dr. Taruna Sunil** H.O.D. CSE Department, JK Lakshmi pat University, Jaipur has been of great help in carrying out the project work and is acknowledged with reverential thanks.

I/WE would like to express a deep sense of gratitude and thanks profusely to Dr. Kedar Sharma, without his wise counsel and able guidance, it would have been impossible to complete the project in this manner.

I express gratitude to other faculty members of the Computer Science department of JK Lakshmi pat University for their intellectual support throughout the course of this work.

Finally, I am indebted to all whosoever has contributed in this report work.

Signature
Madikanti Naveen

ABSTRACT

This project, titled " **DATA MINING FOR FINDING RAILWAY TRACK UTILIZATION**" Tackles the challenge of simplifying the complex Indian railway network, particularly for trains like Jan-Shatabdi and Intercity services. In this project, Python's GeoPandas is utilized to implement a color-coded mapping system, offering an insightful representation of different train routes. The primary goal is to empower passengers with a clear and accessible overview of various.

train options. Beyond the initial project details, an additional layer of insights involves employing distinct colours to represent specific train types such as Jan Shatabdi and Intercity services. This visual categorization adds a valuable dimension to the project, aiding travellers in quickly identifying and selecting the most suitable train for their journeys. The comprehensive approach includes data collection and analysis of India's railway network, encompassing details on train types, routes, schedules, and city pairs.

TABLE OF CONTENT

List of Figures

CHAPTER 1: INTRODUCTION

In the contemporary era, with the exponential growth of data across various industries, the role of data mining has become increasingly crucial. Data mining involves the extraction of valuable patterns and knowledge from large datasets, facilitating informed decision-making and enhancing operational efficiency. In the realm of transportation and logistics, one pertinent area where data mining can be applied is in optimizing track utilization. Rail transportation relies heavily on the efficient use of tracks to ensure timely and cost-effective movement of goods and passengers.

The aim of this minor project is to explore and implement data mining techniques to enhance the utilization of railway tracks, ensuring optimal efficiency in terms of resource allocation, scheduling, and overall performance. Traditional methods of track management often fall short in addressing the dynamic and complex nature of modern transportation systems. Data mining offers a systematic approach to analyze historical and real-time data, uncover hidden patterns, and derive actionable insights for improving track utilization.

This project will delve into various data sources, including historical train schedules, maintenance records, weather data, and real-time operational information. By applying data mining algorithms such as clustering, classification, and association rule mining, we aim to identify patterns and correlations that can aid in predicting and preventing track congestion, minimizing delays, and optimizing resource allocation. Additionally, machine learning models can be employed to develop predictive analytics for proactive decision-making, ensuring better adaptability to changing conditions and improving overall system resilience.

The significance of this project lies in its potential to revolutionize the way rail tracks are managed, leading to a more sustainable and efficient transportation infrastructure. By harnessing the power of data mining, we aspire to contribute to the creation of smarter, data-driven systems that not only enhance track utilization but also pave the way for a more reliable and responsive transportation network. As we embark on this minor project, we anticipate that the findings and methodologies developed will serve as a foundation for future advancements in the field of rail transportation optimization.

1.1 Problem Statement

The project aims to utilize data mining techniques on railway timetable data to analyze and optimize track utilization. It involves collecting, preprocessing, and analyzing railway timetable data to identify patterns, anomalies, and key performance indicators related to track utilization. The objective is to develop predictive models and optimization strategies that improve track efficiency, reduce congestion, enhance punctuality, and suggest adjustments to train schedules and maintenance plans. The project also includes the creation of user-friendly visualizations and reports to communicate findings to stakeholders and ongoing monitoring for continuous improvement, with a focus on collaborating with railway authorities and ensuring data privacy and security.

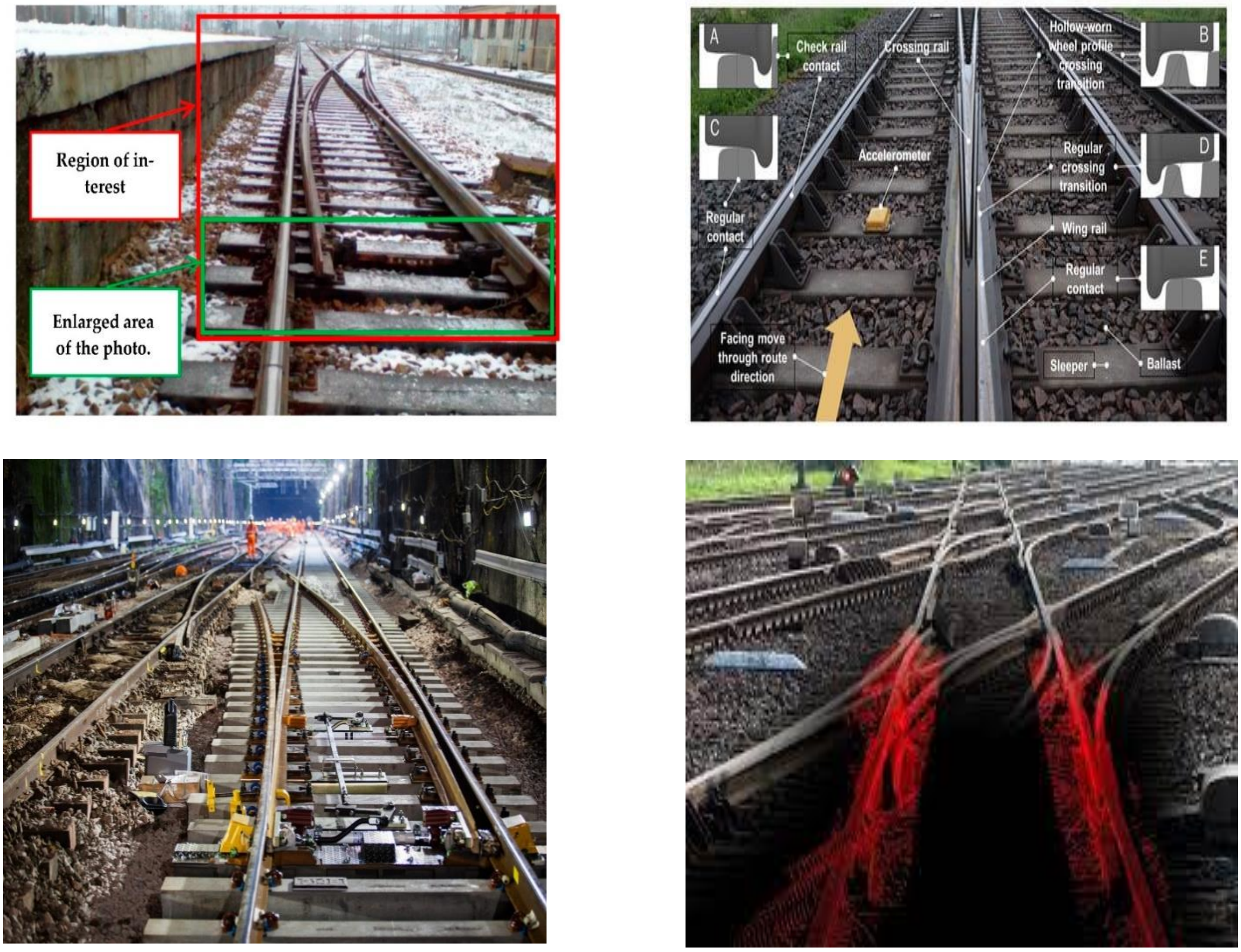

Figure 1 – Track Utilization

1.2 Objectives

This project aims to use advanced data mining techniques on historical railway schedules to find important information. We want to analyze the data in these schedules to discover patterns and trends. By doing this, we hope to make better decisions about how to use railway tracks more efficiently. I worked hard and dedicated should lead to smoother railway operations, fewer delays, and an overall better-performing railway system. This project is important for making both the operations and passengers of the railway system happier and more satisfied.

- **Data Collection and Integration:**
  Collect diverse datasets related to rail transportation, including historical train schedules, maintenance records, weather data, and real-time operational information. Integrate and preprocess the collected data to create a comprehensive dataset for analysis.

- **Exploratory Data Analysis (EDA):**
  Conduct exploratory data analysis to gain insights into the distribution, trends, and characteristics of the collected data. Identify any outliers, missing values, or anomalies that may impact the accuracy of subsequent analyses.

- **Pattern Recognition and Clustering:**
  Apply clustering algorithms to identify spatial and temporal patterns in track utilization. Group similar tracks and time periods to uncover trends and potential bottlenecks.

- **Real-time Decision Support System:**
  Design and implement a real-time decision support system that utilizes the developed models and insights to aid in on-the-fly decision-making. Integrate the system with existing rail management infrastructure for seamless implementation.

- **Documentation and Reporting:**
  Maintain comprehensive documentation of the entire data mining process, including data sources, preprocessing steps, algorithms used, and results obtained. Generate detailed reports and visualizations to communicate findings and recommendations effectively.

- **Knowledge Transfer and Implementation:**

  Provide training and knowledge transfer to relevant stakeholders, including railway operators and management, to ensure successful implementation and utilization of the developed data mining solutions. Support the integration of the proposed optimization strategies into existing rail management systems for long-term impact.

1.3 IDENTIFICATION AND RECOGNIZATION OF NEEDS

- **Data Availability and Collection:**

  Identify and assess the availability of relevant data sources related to railway track utilization, including historical schedules, maintenance records, and real-time operational data. Establish a systematic data collection process to ensure a comprehensive dataset for analysis.

- **Stakeholder Requirements Analysis:**

  Engage with key stakeholders, including railway operators, maintenance teams, and decision-makers, to understand their specific requirements and challenges in track utilization. Identify the critical factors and key performance indicators (KPIs) that are crucial for the efficient operation of the railway system.

- **Data Preprocessing and Cleaning:**

  Evaluate the quality of the collected data and identify any inconsistencies, missing values, or outliers. Develop a robust preprocessing strategy to clean and organize the data, ensuring its suitability for data mining techniques.

- **Technological Infrastructure Assessment:**

  Evaluate the existing technological infrastructure supporting railway operations and data management. Identify any gaps or limitations that may hinder the successful implementation of data mining techniques.

- **Feedback Mechanism Implementation:**

  Establish a feedback mechanism for continuous improvement, involving stakeholders in the iterative process of identifying and reorganizing needs based on evolving operational realities and user feedback.

- **Existing System:**

  In the current scenario of track utilization within railway systems, traditional methods rely heavily on manual planning and scheduling processes. Railway operators often use historical data and experience to allocate tracks, plan maintenance activities, and manage operational schedules. This approach, while functional, may lack the sophistication needed to adapt to dynamic and complex factors influencing track usage, leading to suboptimal resource allocation and potential inefficiencies.

1.4 Data Set View

Data was collected from the various sources like National Railway Authorities, Railway Operators and Companies, Transportation and Infrastructure Departments, Central railway websites like IRCTC etc.

The main data is:

- Csv_files
- Stations. Json
- trains. Json
- Schedules. Json

Train No	Train Nam	SEQ	Station Co	Station Na	Arrival tim	Departure	Distance	Source Sta	Source Sta	Destinatio	Destination Station Name
107	SWV-MAO	1	SWV	SAWANTW	00:00:00	10:25:00	0	SWV	SAWANTW	MAO	MADGOAN JN.
107	SWV-MAO	2	THVM	THIVIM	11:06:00	11:08:00	32	SWV	SAWANTW	MAO	MADGOAN JN.
107	SWV-MAO	3	KRMI	KARMALI	11:28:00	11:30:00	49	SWV	SAWANTW	MAO	MADGOAN JN.
107	SWV-MAO	4	MAO	MADGOAN	12:10:00	00:00:00	78	SWV	SAWANTW	MAO	MADGOAN JN.
108	VLNK-MAC	1	MAO	MADGOAN	00:00:00	20:30:00	0	MAO	MADGOAN	SWV	SAWANTWADI ROAD
108	VLNK-MAC	2	KRMI	KARMALI	21:04:00	21:06:00	33	MAO	MADGOAN	SWV	SAWANTWADI ROAD
108	VLNK-MAC	3	THVM	THIVIM	21:26:00	21:28:00	51	MAO	MADGOAN	SWV	SAWANTWADI ROAD
108	VLNK-MAC	4	SWV	SAWANTW	22:25:00	00:00:00	83	MAO	MADGOAN	SWV	SAWANTWADI ROAD
128	MAO-KOP	1	MAO	MADGOAN	19:40:00	19:40:00	0	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	2	KRMI	KARMALI	20:18:00	20:20:00	33	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	3	THVM	THIVIM	20:40:00	20:42:00	51	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	4	SWV	SAWANTW	21:16:00	21:18:00	83	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	5	KUDL	KUDAL	21:38:00	21:40:00	104	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	6	SNDD	SINDHU D	21:54:00	21:56:00	114	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	7	KKW	KANKAVAL	22:18:00	22:20:00	132	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	8	VBW	VAIBHAVW	22:40:00	22:42:00	163	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	9	RAJP	RAJAPUR F	22:56:00	22:58:00	179	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	10	RN	RATNAGIR	23:52:00	23:57:00	244	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	11	SGR	SANGMESI	00:34:00	00:36:00	280	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	12	CHI	CHIPLUN	01:10:00	01:12:00	322	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	13	KHED	KHED	01:52:00	01:54:00	352	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	14	MNI	MANGAON	04:00:00	04:02:00	422	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	15	ROHA	ROHA	04:55:00	05:00:00	455	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	16	PNVL	PANVEL	06:20:00	06:25:00	532	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
128	MAO-KOP	17	KJT	KARJAT	07:40:00	07:45:00	560	MAO	MADGOAN	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS

Figure 2 - Data Set of Trains


```json
    "geometry": {
      "type": "Point",
      "coordinates": [79.519746, 28.913427000000002]
    },
    "type": "Feature",
    "properties": {
      "state": "Uttar Pradesh",
      "code": "KHH",
      "name": "KICHHA",
      "zone": "NER",
      "address": "Kichha, Uttar Pradesh"
    }
  },
  {
```


Figure 3 - Stations Data

```json
{
  "type": "FeatureCollection",
  "features": [
    {
      "geometry": {
        "type": "LineString",
        "coordinates": [
          [74.880117, 32.706975],
          [74.953339, 32.762368],
          [75.044078, 32.80617],
          [75.13222499999999, 32.769105],
          [75.14542599999999, 32.863528],
          [75.154881, 32.92664]
        ]
      },
      "type": "Feature",
      "properties": {
        "third_ac": 0,
        "arrival": "12:15:00",
        "from_station_code": "JAT",
        "name": "Jammu Tawi Udhampur Special",
        "zone": "NR",
        "chair_car": 0,
        "first_class": 0,
        "duration_m": 35,
        "sleeper": 0,
        "from_station_name": "JAMMU TAWI",
        "number": "04601",
        "departure": "10:40:00",
        "return_train": "04602",
        "to_station_code": "UHP",
        "second_ac": 0,
        "classes": "",
        "to_station_name": "UDHAMPUR",
        "duration_h": 1,
        "type": "DEMU",
        "first_ac": 0,
        "distance": 53
      }
```


Figure 4 - Trains Data (JSON)


```json
{
  "from_stations": [
    "JAMMU TAWI",
    "UDHAMPUR",
    "JAMMU TAWI",
    "UDHAMPUR",
    "MUMBAI BANDRA TERMINUS",
    "Sirsa",
    "Hisar",
    "DEGANA JN",
    "MERTA ROAD JN",
    "JAMMU TAWI",
    "MUMBAI DADAR CENTRAL",
    "MUMBAI CST",
    "BHUBANESWAR",
    "MUMBAI CST",
    "C SHAHU M RAJ KOLHAPUR TERM",
    "MANMAD JN",
    "PUNE JN",
    "MUMBAI CST",
    "MUMBAI CST",
    "C SHAHU M RAJ KOLHAPUR TERM",
    "PUNE JN",
    "MUMBAI DADAR CENTRAL",
    "MYSORE JN",
    "PUNE JN",
    "GORAKHPUR JN",
```


Figure 5 - Stations Data (JSON)

1.5 The Existing System Typically Involves

- **Manual Planning:** Human operators manually plan, and schedule track usage based on historical data and established routines.

- **Limited Predictive Capabilities:** The system may lack advanced predictive capabilities to anticipate congestion, delays, or optimal resource allocation in real-time.

- **Dependency on Experience:** Decision-making relies heavily on the experience and intuition of railway operators, which may not always capture intricate patterns or dependencies in the data.

Proposed System:

The proposed system aims to revolutionize track utilization by leveraging advanced data mining techniques to optimize decision-making processes. It introduces a data-driven approach to enhance the efficiency of track allocation, reduce delays, and improve overall railway system performance. The key features of the proposed system include:

- **Data Mining Algorithms:** Implementation of sophisticated data mining algorithms, such as clustering, classification, and association rule mining, to extract valuable insights from historical and real-time data.

- **Predictive Analytics:** Development of machine learning models for predictive analytics, enabling the system to forecast potential track congestion, optimize resource allocation, and proactively prevent delays.

- **Real-time Decision Support System:** Integration of a real-time decision support system that uses the analyzed data to provide actionable insights for optimal track utilization.

- **Adaptability and Scalability:** Designing the system to be adaptable to changing operational conditions and scalable to accommodate future growth and advancements in railway technology.

- **Continuous Improvement:** Implementation of a feedback mechanism to continuously improve the system based on user feedback and evolving operational requirements.

- **Regulatory Compliance:** Ensuring that the proposed optimizations adhere to regulatory requirements and safety standards, mitigating potential risks associated with track utilization changes. The proposed system seeks to enhance not only the efficiency of track utilization but also the overall performance and satisfaction of both railway operators and passengers through a systematic and data-driven approach to decision-making.

1.6 Unique Features of the System

- **Real-time Decision Support System:**
  Development of a real-time decision support system that processes live data and provides instantaneous insights, empowering railway operators to make informed decisions on the fly.

- **Dynamic Adaptation to Environmental Factors:**
  Incorporation of environmental data, such as weather conditions, into the data mining process to dynamically adapt track utilization strategies based on external factors that may impact operational efficiency.

- **Capacity Planning and Load Balancing:**
  Implementation of data mining techniques to perform capacity planning and load balancing, ensuring that tracks are allocated in a way that optimizes overall system capacity and minimizes the risk of congestion.

- **Energy Efficiency Optimization:**
  Exploration of data mining strategies to optimize energy consumption in railway operations, considering factors like train speed, route selection, and energy-efficient scheduling to reduce overall energy costs.

- **Scenario Modeling and What-If Analysis:**
  Integration of scenario modeling and what-if analysis capabilities to simulate various operational scenarios and assess the potential impact of changes in track utilization

strategies before implementation.

- **Collaborative Decision-Making Support:**
  Development of collaborative decision-making features that facilitate communication and coordination among different stakeholders, promoting a more integrated and synchronized approach to track utilization optimization.

- **Emergency Response Planning:**
  Incorporation of data mining techniques to develop emergency response plans, anticipating and mitigating the impact of unforeseen events, such as accidents or track failures, on overall railway operations.

- **Customer Experience Optimization:**
  Consideration of passenger experience metrics in the data mining process, aiming to optimize track utilization not only for operational efficiency but also to enhance the overall experience of railway passengers.

CHAPTER 2: REQUIREMENT ANALYSIS AND SYSTEM SPECIFICATION

Feasibility Study for "Data Mining for Optimal Track Utilization":

2.1 Technical Feasibility:

- ✓ **Data Availability and Integration:**

  Assessment: Evaluate the availability and accessibility of relevant data sources (historical schedules, maintenance records, real-time operational data) for comprehensive analysis.
  Feasibility: If data is accessible and can be integrated effectively, the technical feasibility is high.

- ✓ **Algorithmic Implementation:**

  Assessment: Review the technical feasibility of implementing advanced data mining algorithms (clustering, classification, association rule mining) for pattern recognition.
  Feasibility: Ensure that the chosen algorithms align with the project's goals and can be implemented within the existing technical infrastructure.

- ✓ **Scalability and Performance:**

  Assessment: Evaluate whether the proposed system can scale to accommodate growing data volumes and maintain performance efficiency.
  Feasibility: If the system can handle increased data loads and maintain real-time processing, technical feasibility is assured.

- ✓ **Scalability and Performance:**
- ✓

  Assessment: Evaluate whether the proposed system can scale to accommodate growing data volumes and maintain performance efficiency.

  Feasibility: If the system can handle increased data loads and maintain real-time processing, technical feasibility is assured.

2.2 Economic Feasibility:

- ✓ **Cost of Technology and Infrastructure:**

  Assessment: Calculate the costs associated with acquiring and implementing the necessary technology and infrastructure for data mining.

  Feasibility: Compare costs against potential benefits; if the benefits outweigh the costs, economic feasibility is positive.

- ✓ **Operational Costs:**

  Assessment: Estimate ongoing operational costs, including maintenance, training, and any additional resources required.

  Feasibility: Ensure that operational costs are within budget constraints and do not pose a financial burden.

- ✓ **Return on Investment (ROI):**

  Assessment: Project the potential ROI based on the improvements in track utilization efficiency and operational performance.

  Feasibility: If the projected ROI is favorable, economic feasibility is established.

2.3 Operational Feasibility:

- ✓ **Integration with Existing Systems:**

  Assessment: Evaluate the feasibility of integrating the proposed system with existing railway management systems.

  Feasibility: If seamless integration is possible without major disruptions, operational feasibility is high.

- ✓ **User Acceptance:**

  Assessment: Gauge the acceptance and willingness of railway operators and decision-makers to adopt the new data-driven approach.

  Feasibility: If stakeholders are receptive and supportive, operational feasibility is positive.

- ✓ **Adaptability to Operational Constraints:**

  Assessment: Examine how well the proposed system can adapt to operational constraints such as maintenance schedules and peak traffic periods.

  Feasibility: If the system can accommodate these constraints effectively, operational feasibility is assured.

2.4 Software Requirement Specification (SRS) Document

2.4.1 Data Requirement:

- ✓ **Data Sources:**

  Description: The system shall utilize historical train schedules, maintenance records, real-time operational data, and external sources like weather reports.

  Format: Data formats should be compatible with common data storage and processing standards.

- ✓ **Data Integration:**

  Description: The system must seamlessly integrate data from various sources for comprehensive analysis.

  Format: Data integration should be performed in a standardized format to ensure consistency.

2.4.2 Functional Requirement:

- ✓ **Track Allocation:**

  Description: The system shall use clustering algorithms to optimize the allocation of tracks based on historical and real-time data.

  Criteria: The system should allocate tracks considering factors like historical usage, maintenance schedules, and real-time traffic conditions.

- ✓ **Predictive Analytics:**

  Description: Machine learning models shall be employed for predictive analytics to forecast potential track congestion and optimize resource allocation.

  Criteria: Predictions should be accurate, with a defined confidence level.

2.4.3 Performance Requirement: -

- ✓ **Real-time Processing:**

  Description: The system must process real-time data to provide instantaneous insights for decision-making.

  Criteria: The processing time for real-time data should be within acceptable limits.

- ✓ **Scalability:**

  Description: The system should be scalable to accommodate increased data volumes as the railway network grows.

  Criteria: The system should maintain performance efficiency with growing data loads.

2.4.4 Dependability Requirement:

- ✓ **System Reliability:**

  Description: The system should be reliable and available during operational hours.

  Criteria: The system should have a minimal downtime rate and be capable of quick recovery in case of failures.

- ✓ **Backup and Recovery:**

  Description: Regular data backups and a robust recovery mechanism shall be implemented.

  Criteria: Data recovery should be efficient, and backups should be stored securely.

2.4.5 Maintainability Requirement:

- ✓ **System Updates:**

  Description: The system should allow for updates and enhancements to accommodate evolving requirements.

  Criteria: Updates should be seamless, with minimal disruption to ongoing operations.

- ✓ **Documentation:**

  Description: Comprehensive documentation for system architecture, algorithms, and data models shall be maintained.

  Criteria: Documentation should be regularly updated and easily accessible to system administrators.

2.4.6 Security Requirement:

- ✓ **Data Encryption:**

  Description: All data transmission and storage should be encrypted to ensure data security.

  Criteria: Use industry-standard encryption algorithms and protocols.

- ✓ **Access Control:**

  Description: Access to the system should be role-based, with restricted permissions based on user roles.

  Criteria: Only authorized personnel should have access to sensitive features and data.

2.5 Expected Hurdles

Data Quality and Consistency:

Mitigation: Implement rigorous data preprocessing and cleaning procedures.

Integration Challenges:

Mitigation: Develop standardized data integration protocols and conduct thorough testing.

Algorithm Complexity:

Mitigation: Ensure algorithmic efficiency through optimization and regular performance monitoring.

User Resistance to Change:

Mitigation: Conduct extensive user training sessions and engage stakeholders throughout the development process.

Security Concerns:

Mitigation: Implement robust encryption protocols and conduct security audits regularly.

Scalability Issues:

Mitigation: Design the system with scalability in mind and regularly assess and upgrade hardware resources.

This SRS document provides a comprehensive overview of the software requirements for the "Data Mining for Optimal Track Utilization" system, along with validation strategies and anticipated challenges.

2.6 SDLC model to be used.

The choice of a Software Development Life Cycle (SDLC) model for the "Data Mining for Optimal Track Utilization" project depends on the project's specific requirements, constraints, and characteristics.

Here,

Iterative and Incremental Development (IID) Model:

Overview:

The Iterative and Incremental Development model involves breaking down the project into smaller, manageable iterations. Each iteration includes a subset of features or functionalities. The project evolves through multiple cycles, with each iteration building upon the lessons learned from the previous ones.

Application to the Project:

Advantages: -

Flexibility to accommodate changing requirements, which is common in data mining projects. Early delivery of partial functionality, allowing for quick feedback and adjustments. Opportunities for continuous improvement based on user feedback and evolving needs.

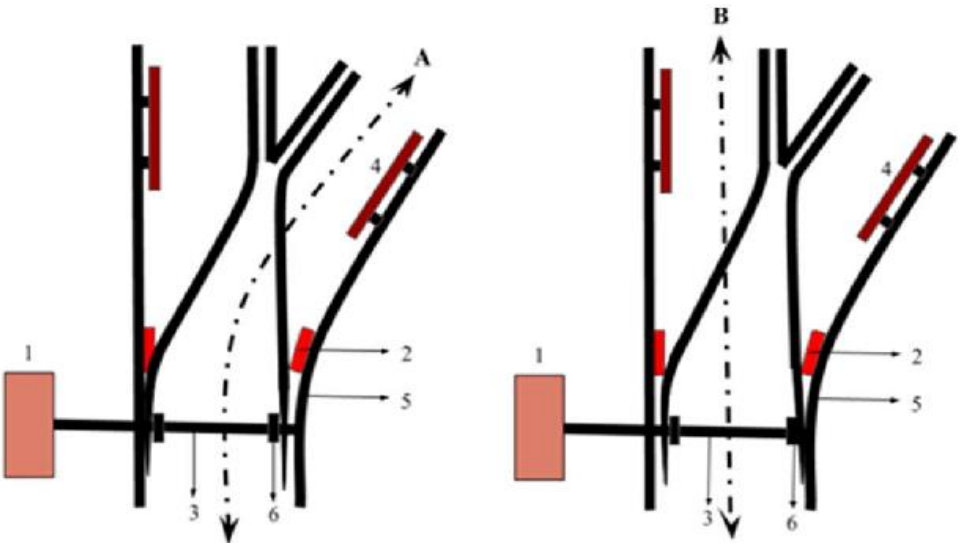

Figure 6 - Data Mining for Track Utilization

Phases: -

- **Planning:** Define the overall project scope, goals, and high-level requirements.
- **Iteration Planning:** Break down the project into iterations, prioritizing features, and functionalities.
- **Development:** Implement features for each iteration, ensuring functional increments with each cycle.
- **Testing:** Conduct testing for each iteration to identify and rectify issues early.
- **Evaluation:** Gather feedback from stakeholders and end-users after each iteration.
- **Adjustment:** Make necessary adjustments based on feedback and incorporate new requirements.
- **Repeat:** Repeat the process for subsequent iterations until the project goals are achieved.
- **Considerations:**
  Well-suited for projects where requirements are not well-defined initially.
  Frequent interactions with stakeholders ensure alignment with their expectations.
  Allows for early identification of potential issues and reduces the risk of late-stage changes.

3.1 Implementation

Languages:

- ✓ **Python:** Utilized for its extensive libraries and frameworks in data mining, such as Pandas, NumPy, and scikit-learn.

IDEs (Integrated Development Environments):

- ✓ **Jupiter Notebooks:** Ideal for data exploration and analysis.
- ✓ **PyCharm:** For Python code development.

3.2 Tools and Technologies

- ✓ **Scrapy:** For web scraping if needed to collect additional data.
- ✓ **Pandas and NumPy:** Data manipulation and analysis.
- ✓ **scikit-learn:** For machine learning algorithms if predictive modeling is part of the data mining process.

Coding Standards:

Adhere to general coding standards and best practices, and specifically for Python:

- ✓ Use meaningful variable and function names.
- ✓ Implement proper exception handling.
- ✓ Comment code for clarity, especially for complex algorithms.

3.3 Considerations

- ✓ **Task Breakdown:** Identify tasks like data collection, preprocessing, modeling, and analysis.
- ✓ **Dependencies:** Determine dependencies between tasks (e.g., modeling depends on data preprocessing).
- ✓ **Resource Allocation:** Allocate resources based on task requirements.

3.4 Testing

- ✓ **Testing Techniques:**

  Data Validation Testing: Ensure the accuracy and completeness of collected data.

  Exploratory Data Analysis (EDA): Visual and statistical analysis to understand data patterns.

- ✓ **Test Plans:**

  Data Quality Checks: Ensure data integrity and consistency.

  Model Evaluation Plan: If applicable, outline how the predictive model's accuracy will be assessed.

  Performance Testing: Evaluate the system's performance under various loads.

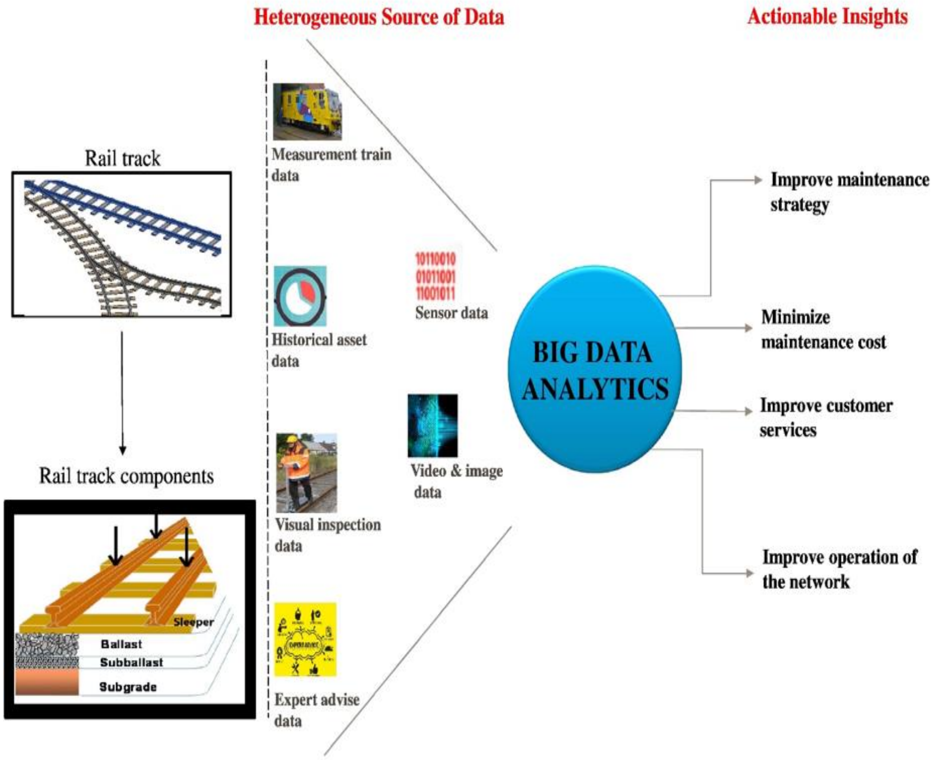

Figure 7 - Testing Of Track Utilization

- ✓ **Maintenance:**

  Activities:

  Data Updates: Regularly update the dataset to include new information.

  Documentation Updates: Keep user and technical documentation up to date with any changes made.

CHAPTER 4: RESULTS AND DISCUSSION

The "Data Mining for Finding Railway Track Utilization" project aimed to boost railway operational efficiency through advanced data analytics. Key results included optimized track usage patterns, enabling proactive maintenance during low-traffic periods, and predictive models that reduced downtime. Traffic flow optimization was achieved by refining train schedules based on historical data, minimizing congestion and delays. The system enhanced resource allocation efficiency, determining optimal distribution for cost savings and improved performance.

This Project will show the track utilization metrics, train statuses, and maintenance activities. The analytics section provided graphical representations of usage patterns and predictive maintenance insights, while the scheduling module allowed interactive adjustments based on optimization suggestions.

Comprising modules for data collection, preprocessing, data mining algorithms, and an optimization engine, the system utilized Python for tables of trains, maintenance logs, schedules, and optimization parameters. The project delivered actionable insights, modular design, collectively providing a comprehensive solution for the rail industry, improving overall efficiency and resource management.

```python
def json_to_df(json_data):
    columns = list(json_data[0].keys())
    data_vals = [list(i.values()) for i in json_data]
    return pd.DataFrame(data=data_vals, columns=columns)
```

```python
states_ind = read_shp_file(dir_name='Indian_shapefile')
# # states_ind = read_shp_file(dir_name='states')
```

```python
import geopandas as gpd
import matplotlib.pyplot as plt

shapefile_path = '../data/shape_files/Indian_shapefile/india_st.shp'  # Replace with the actual path to your shapefile

gdf = gpd.read_file(shapefile_path)

gdf.plot()
plt.title('Indian States Boundaries')
plt.show()
```

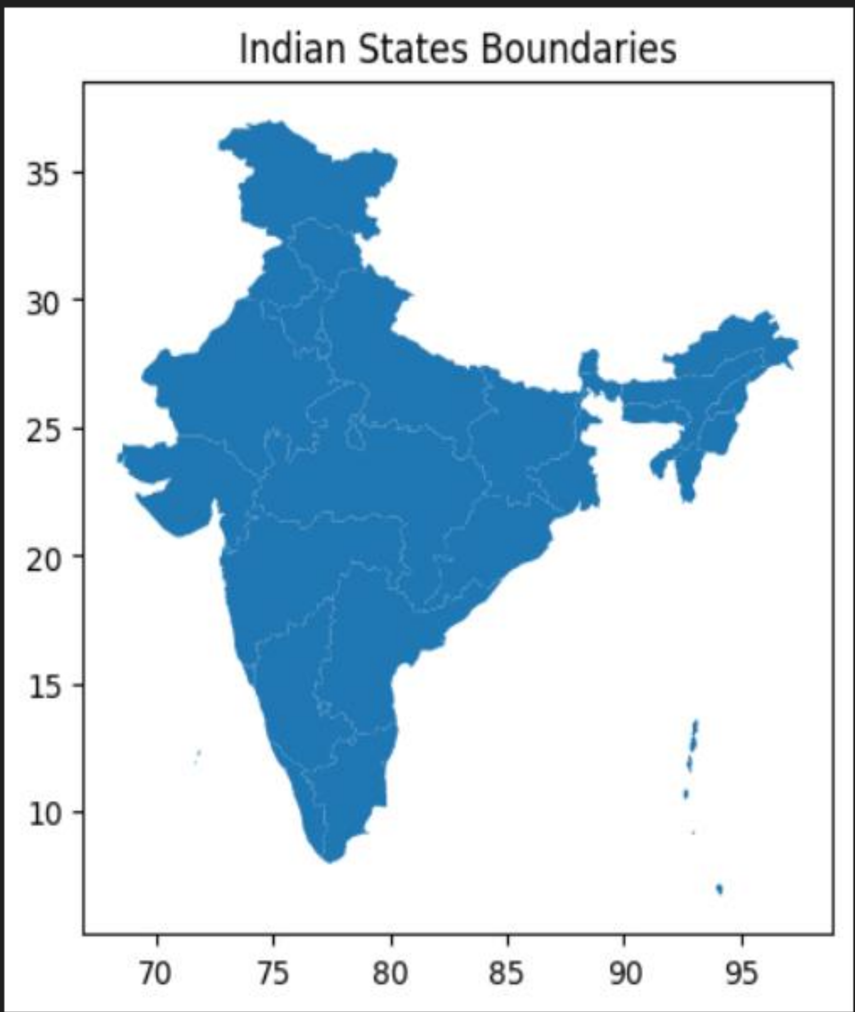

Figure 8 - **Visualization of Indian States using Matplotlib and GeoPandas**

```python
def plot_stations(stations_gdf, state_name=None):
    # India latitude and longitude
    clat = 20.5937
    clon = 78.9629
    zoom = 4
    title = 'India'

    if state_name:
        filtered_states_stations = states_stations.dropna(subset=['state'])
        all_states = list(np.unique(ar=filtered_states_stations['state']))
        if state_name in all_states:
            stations_gdf = stations_gdf[stations_gdf['state'].isin(state_name)]
            # mean of the state_name's latitude and longitude
            clat = np.mean(stations_gdf.geometry.y)
            clon = np.mean(stations_gdf.geometry.x)
            zoom = 6
            title = state_name
        else:
            return '`state_name` not matching with the records.'

    fig = px.scatter_mapbox(
        data_frame=stations_gdf,
        lat=stations_gdf.geometry.y,
        lon=stations_gdf.geometry.x,
        hover_name='name',
        center=dict(lat=clat, lon=clon),
        mapbox_style='carto-positron',
        zoom=zoom,
    )
    fig.update_layout(
        title=title,
        autosize=True,
        height=600,
        hovermode='closest',
        showlegend=True,
        margin=dict(l=10, r=10, t=40, b=0)
    )
    fig.show()

    return None
```

```python
plot_stations(
    stations_gdf=states_stations)
```

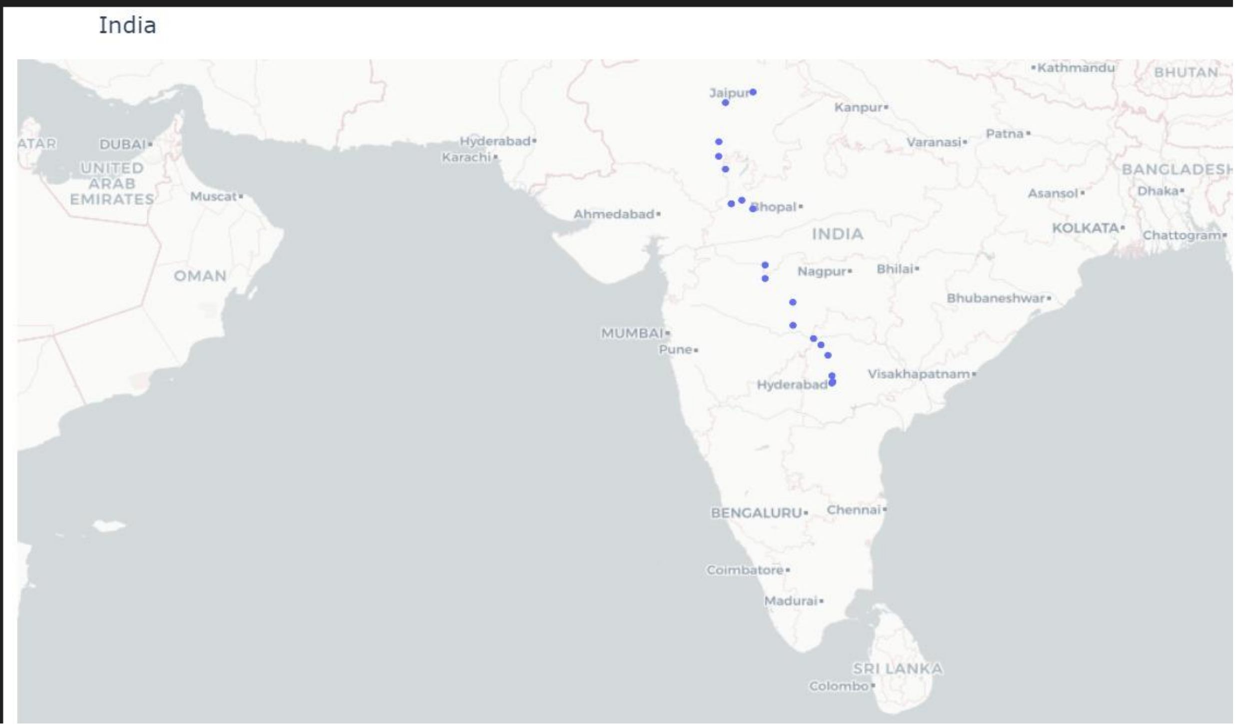

Figure 9 - Stations Between Jaipur (JN) & Secunderabad (JN)

CHAPTER 5: CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

In conclusion, the "Data Mining for Finding Railway Track Utilization" project has proven instrumental in optimizing railway operations by leveraging advanced data analytics. The results demonstrated improved track usage patterns, enabling proactive maintenance strategies, and predicting maintenance needs, ultimately reducing downtime. The traffic flow optimization successfully refined train schedules, minimizing congestion, and enhancing overall operational efficiency. The resource allocation module contributed to cost savings by determining optimal distribution based on demand patterns.

This will provide a modular system design, provides a practical tool for railway authorities to make informed decisions in real-time. The integration of data mining algorithms and predictive models ensures a proactive approach to addressing operational challenges, contributing to a more resilient and efficient railway system.

5.2 Future Scope:

The project lays the foundation for several future enhancements and expansions:

Real-Time Predictive Analytics: Integrate real-time data streams for continuous predictive analytics, allowing for immediate response to changing operational conditions and emergencies.

Integration with IoT Devices: Incorporate Internet of Things (IoT) devices for live monitoring of track conditions, train health, and passenger load, providing a more comprehensive dataset for analysis.

Machine Learning for Dynamic Scheduling: Implement advanced machine learning algorithms for dynamic train scheduling, considering unpredictable factors such as weather conditions and special events.

Predictive Financial Modeling: Extend the project to include predictive financial modeling to estimate the cost-effectiveness of maintenance strategies and optimize budget allocations.

Collaboration with Stakeholders: Enhance the system to facilitate collaboration with various stakeholders, including maintenance teams, regulatory bodies, and other transportation modes for seamless intermodal connectivity.

REFERENCES

1. Thaduri, A., Galar, D., & Kumar, U. (2015). Railway assets: A potential domain for big data analytics. *Procedia Computer Science*, *53*, 457-467.
2. Goverde, R. M., & Meng, L. (2011). Advanced monitoring and management information of railway operations. *Journal of Rail Transport Planning & Management*, *1*(2), 69-79.
3. Tang, Y. (2021). Risk Chain and Key Hazard Management for Urban Rail Transit System Operation Based on Big Data Mining. *Discrete Dynamics in Nature and Society*, *2021*, 1-10.
4. BALCI, E., YALÇIN, E., Yelce, T. U., & BEZGİN, N. (2021). Smart Diagnosis And Maintenance Systems For Railway Tracks. *Akıllı Ulaşım Sistemleri ve Uygulamaları Dergisi*, *4*(2), 134-147.
5. Shmueli, G., Patel, N. R., & Bruce, P. C. (2011). *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. John Wiley and Sons.
6. Diaz Salgado, A. (2021). *Passenger utilization on railway network using timeseries forecasting* (Doctoral dissertation, Dublin, National College of Ireland).
7. Hansen, I. A., Goverde, R. M., & van der Meer, D. J. (2010, September). Online train delay recognition and running time prediction. In *13th International IEEE Conference on Intelligent Transportation Systems* (pp. 1783-1788). IEEE.
8. Fu, J., & Chen, J. (2021). A green transportation planning approach for coal heavy-haul railway system by simultaneously optimizing energy consumption and capacity utilization. *Sustainability*, *13*(8), 4173.
9. Menéndez, M., Martínez, C., Sanz, G., & Benitez, J. M. (2016). Development of a smart framework based on knowledge to support infrastructure maintenance decisions in railway corridors. *Transportation Research Procedia*, *14*, 1987-1995.
10. Sol-Sánchez, M., Moreno-Navarro, F., & Rubio-Gámez, M. C. (2015). The use of elastic elements in railway tracks: A state of the art review. *Construction and building materials*, *75*, 293-305.
11. Aggarwal, P. K., Jain, P., Mehta, J., Garg, R., Makar, K., & Chaudhary, P. (2021). Machine learning, data mining, and big data analytics for 5G-enabled IoT. *Blockchain for 5G-Enabled IoT: The new wave for Industrial Automation*, 351-375.
12. Liu, W., Zheng, Y., Chawla, S., Yuan, J., & Xing, X. (2011, August). Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1010-1018).
13. Ivina, D., & Palmqvist, C. W. (2023). Railway maintenance windows: Discrepancies between planning and practice in Sweden. *Transportation research interdisciplinary perspectives*, *22*, 100927.
14. Mahmood, A., Shi, K., Khatoon, S., & Xiao, M. (2013). Data mining techniques for wireless sensor networks: A survey. *International Journal of Distributed Sensor Networks*, *9*(7), 406316.
15. Verma, J. Data Mining in Indian Railways: A Survey to Analyze Applications of Data Mining. *International Journal of Computer Applications*, *975*, 8887.
16. Fan, W. (2004, August). Systematic data selection to mine concept-drifting data streams.

In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 128-137).

17. Chandran, P., Thiery, F., Odelius, J., Lind, H., & Rantatalo, M. (2022). Unsupervised Machine Learning for Missing Clamp Detection from an In-Service Train Using Differential Eddy Current Sensor. *Sustainability*, *14*(2), 1035.

18. Basu, J. K., Bhattacharyya, D., & Kim, T. H. (2010). Use of artificial neural network in pattern recognition. *International journal of software engineering and its applications*, *4*(2).

19. Hills, J., Lines, J., Baranauskas, E., Mapp, J., & Bagnall, A. (2014). Classification of time series by shapelet transformation. *Data mining and knowledge discovery*, *28*, 851-881.

20. McCue, C. (2014). *Data mining and predictive analysis: Intelligence gathering and crime analysis*. Butterworth-Heinemann.

APPENDIX

```
import json
import os
import numpy as np
import pandas as pd
import geopandas as gpd
import plotly.express as px
from plotly import graph_objects as go
from shapely.geometry import LineString, Point
```

```
df = pd.read_csv("Trains.csv",low_memory=False)
df=df.drop("SEQ",axis='columns')
# mitosheet.sheet(df, analysis_to_replay="id-mycezhobxb")
```

```
# Print the filtered data
print(filtered_data)
l=filtered_data.drop_duplicates("Train No")
l=set(filtered_data["Train No"])
l
```

```
      Train No    Train Name Station Code  Station Name Arrival time  \
531       2732  JP-HYB SPECI           JP    JAIPUR JN.     14:35:00
532       2732  JP-HYB SPECI           FL   PHULERA JN.     15:28:00
533       2732  JP-HYB SPECI          KSG    KISHANGARH     16:10:00
534       2732  JP-HYB SPECI          AII     AJMER JN.     17:10:00
535       2732  JP-HYB SPECI          BHL      BHILWARA     19:30:00
...        ...           ...          ...           ...          ...
45715    17019  JP-HYB   EXP          KMC     KAMAREDDI     04:05:00
45716    17019  JP-HYB   EXP          MED       MEDCHAL     05:55:00
45717    17019  JP-HYB   EXP          MJF    MALKAJGIRI     06:50:00
45718    17019  JP-HYB   EXP           SC  SECUNDERABAD     07:30:00
45719    17019  JP-HYB   EXP          HYB  HYDERABAD DE     08:10:00

      Departure Time Distance Source Station Source Station Name  \
531         14:35:00        0             JP          JAIPUR JN.
532         15:30:00       54             JP          JAIPUR JN.
533         16:12:00      105             JP          JAIPUR JN.
534         17:40:00      134             JP          JAIPUR JN.
535         19:35:00      266             JP          JAIPUR JN.
...              ...      ...            ...                 ...
45715       04:06:00     1885             JP          JAIPUR JN.
45716       05:56:00     1966             JP          JAIPUR JN.
45717       06:51:00     1990             JP          JAIPUR JN.
45718       07:35:00     1993             JP          JAIPUR JN.
45719       08:10:00     2002             JP          JAIPUR JN.
...
45718                 HYB          HYDERABAD DECCAN
45719                 HYB          HYDERABAD DECCAN
```

```
[63 rows x 11 columns]
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
{'17019', '2732'}
```

```
trains= df.drop_duplicates("Train Name")
```

Python

```
trains
```

Python

	Train No	Train Name	Station Code	Station Name	Arrival time	Departure Time	Distance	Source Station	Source Station Name	Destination Station	Destination Station Name
0	107	SWV-MAO-VLNK	SWV	SAWANTWADI R	00:00:00	10:25:00	0	SWV	SAWANTWADI ROAD	MAO	MADGOAN JN.
4	108	VLNK-MAO-SWV	MAO	MADGOAN JN.	00:00:00	20:30:00	0	MAO	MADGOAN JN.	SWV	SAWANTWADI ROAD
8	128	MAO-KOP SPEC	MAO	MADGOAN JN.	19:40:00	19:40:00	0	MAO	MADGOAN JN.	KOP	CHHATRAPATI SHAHU MAHARAJ TERMINUS
30	290	PALACE ON WH	DSJ	DELHI-SAFDAR	18:30:00	18:30:00	0	DSJ	DELHI-SAFDAR JANG	DSJ	DELHI-SAFDAR JANG
44	401	BSB BHARATDA	AWB	AURANGABAD	21:30:00	21:30:00	0	AWB	AURANGABAD	BSB	VARANASI JN.
...	...	...	...	...	...	...	...	...	...	...	...
185475	99804	PUNE-LNL	PUNE	PUNE JN.	04:45:00	04:45:00	0	PUNE	PUNE JN.	LNL	LONAVLA
185763	99821	LNL-SVJR EMU	LNL	LONAVLA	18:20:00	18:20:00	0	LNL	LONAVLA	PUNE	PUNE JN.
186012	99836	PUNE-TGN EMU	PUNE	PUNE JN.	22:10:00	22:10:00	0	PUNE	PUNE JN.	LNL	LONAVLA
186029	99901	TGN-PUNE EMU	TGN	TALEGAON	00:05:00	00:05:00	0	TGN	TALEGAON	PUNE	PUNE JN.
186077	99905	EMU	TGN	TALEGAON	09:57:00	09:57:00	0	TGN	TALEGAON	SVJR	SHIVAJINAGAR

7585 rows × 11 columns

```
track_hyd_jp= df[df["Train No"]=="2732"]
```

```
stations= list(track_hyd_jp["Station Name"])
```

```
stations[-1]="HYDERABAD DECCAN"
stations[-2]="SECUNDERABAD JN"
stations[0]="JAIPUR"
stations
```

```
['JAIPUR',
 'PHULERA JN.',
 'KISHANGARH',
 'AJMER JN.',
 'BHILWARA',
 'CHITTAURGARH',
 'NIMACH',
 'MANDSOR',
 'RATLAM JN',
 'NAGDA JN',
 'UJJAIN JN',
 'BHOPAL',
 'ITARSI',
 'KHANDWA JN.',
 'BURHANPUR',
 'MALKAPUR',
 'AKOLA JN.',
 'WASHIM',
 'HINGOLI(DECC',
 'BASMAT',
 'PURNA  JN.',
```

```
middle_stations=[]
for i in stations:
    m=trains[trains["Station Name"]== i]
    middle_stations.append(list(m["Train Name"]))
```

```
middle_stations
```

```
from itertools import chain
flatten_list=list(chain.from_iterable(middle_stations))
```

```
flatten_list
```

```
len(flatten_list)
```

```
all_trains=[]
for i in range(len(l)):
    trains= df.drop_duplicates("Train Name")
    track_hyd_jp= df[df["Train No"]==list(l)[i]]
    stations= list(track_hyd_jp["Station Name"])
    middle_stations=[]
    for j in stations:
        m=trains[trains["Station Name"]== j]
        middle_stations.append(list(m["Train Name"]))
    all_trains.append(middle_stations)
flatten_list=list(chain.from_iterable(middle_stations))
```

```python
flatten_list
```

```python
len(flatten_list)
```

```python
filtered_data = df[(df['Source Station Name'] == 'SAWANTWADI ROAD') & (df['Destination Station Name'] == 'MADGOAN JN.')]

# Print the filtered data
print(filtered_data)
l=filtered_data.drop_duplicates("Train No")
l=set(filtered_data["Train No"])
l
```

```python
all_trains=[]
for i in range(len(l)):
    trains= df.drop_duplicates("Train Name")
    track_hyd_jp= df[df["Train No"]==list(l)[i]]
    stations= list(track_hyd_jp["Station Name"])
    middle_stations=[]
    for j in stations:
        m=trains[trains["Station Name"]== j]
        middle_stations.append(list(m["Train Name"]))
    all_trains.append(middle_stations)
flatten_list=list(chain.from_iterable(middle_stations))
```

```python
len(flatten_list)
```

```python
filtered_data = df[(df['Source Station Name'] == 'HAZRAT NIZAMUDDIN JN') & (df['Destination Station Name'] == 'MADGOAN JN.')]

# Print the filtered data
print(filtered_data)
l=filtered_data.drop_duplicates("Train No")
l=set(filtered_data["Train No"])
l
```

```python
all_trains=[]
for i in range(len(l)):
    trains= df.drop_duplicates("Train Name")
    track_hyd_jp= df[df["Train No"]==list(l)[i]]
    stations= list(track_hyd_jp["Station Name"])
    middle_stations=[]
    for j in stations:
        m=trains[trains["Station Name"]== j]
        middle_stations.append(list(m["Train Name"]))
    all_trains.append(middle_stations)
flatten_list=list(chain.from_iterable(middle_stations))
```

```python
def json_to_gdf(json_data, geometry_type):
    # ESRI shapefile format only accepts one type of geometry

    # Point category → stations
    if (geometry_type == 'Point'):
        gdf = gpd.GeoDataFrame.from_features(features=json_data['features'])

    # LineString category → trains/railways
    elif (geometry_type == 'LineString'):
        # fetch the column names based on `properties` keys
        properties_columns = list(json_data['features'][0]['properties'].keys())
        # fetch the values (rows) based on the `properties` values
        properties_vals = [list(i['properties'].values()) for i in json_data['features']]

        # convert the coordinates to LineString format
        geometry_column = [
            LineString(i['geometry']['coordinates'])
            if len(i['geometry']['coordinates']) >= 2
            # else Point(i['geometry']['coordinates'][0])
            else LineString([i['geometry']['coordinates'][0]] * 2)
            for i in json_data['features']
        ]

        # pandas dataframe to Geopandas dataframe
        df = pd.DataFrame(data=properties_vals, columns=properties_columns)
        df['geometry'] = geometry_column
        gdf = gpd.GeoDataFrame(df)

    # setting the CRS
    gdf = gdf.set_crs('EPSG:4326')

    return gdf
```

```python
if not os.path.isdir(s=shp_files_path + '\\stations'):
    stations = read_json_file(file_name='stations.json')
    stations_gdf = json_to_gdf(json_data=stations, geometry_type='Point')
    stations_gdf.to_file(filename=shp_files_path + '\\stations')
else:
    print('Shape file already exists.')
    stations_gdf = read_shp_file(dir_name='stations')
```

```
Shape file already exists.
```

```python
stations_gdf.dropna(inplace=True)
stations_gdf= stations_gdf[stations_gdf["name"].isin(stations)]
stations_gdf
```

	state	code	name	zone	address	geometry
19	Andhra Pradesh	NZB	NIZAMABAD	SCR	Nizamabad, Andhra Pradesh	POINT (78.10324 18.67916)
1728	Rajasthan	BHL	BHILWARA	NWR	Bhilwara, Rajasthan	POINT (74.63077 25.34355)
2582	Andhra Pradesh	HYB	HYDERABAD DECCAN	SCR	Hyderabad, Andhra Pradesh	POINT (78.46738 17.39336)
2793	Andhra Pradesh	KMC	KAMAREDDI	SCR	Kamareddi, Andhra Pradesh	POINT (78.33586 18.32586)
2836	Rajasthan	KSG	KISHANGARH	NWR	Kishangarh, Rajasthan	POINT (74.85555 26.58897)
3839	Maharashtra	WHM	WASHIM	SCR	Washim, Maharashtra	POINT (77.14799 20.10310)
4445	Madhya Pradesh	BAU	BURHANPUR	CR	Burhanpur, Madhya Pradesh	POINT (76.19868 21.33440)
4853	Rajasthan	COR	CHITTAURGARH	WR	Chittorgarh, Rajasthan	POINT (74.62369 24.87434)
4858	Maharashtra	DAB	DHARMABAD	SCR	Degaon, Maharashtra	POINT (77.84934 18.88826)
4978	Andhra Pradesh	MED	MEDCHAL	SCR	Rangareddy, Andhra Pradesh	POINT (78.47456 17.63933)
5298	Maharashtra	BMF	BASMAT	SCR	Basmath, Maharashtra	POINT (77.15178 19.33347)
6429	Madhya Pradesh	NMH	NIMACH	WR	MDR 11A, Madhya Pradesh	POINT (74.85303 24.45987)