Indian Railway Trains Delay Dataset

Ankita Anand

April 2024

Indian Railway Trains Delay Dataset "To document [the dataset] motivation, composition, collection process, recommended uses, and so on." The dataset "Indian Railway Express Trains Delay Dataset" is created to provide comprehensive information about train routes, delays, and cancellations for express trains connecting Guwahati (Assam) to major metro cities in India in the year 2023, namely Delhi, Mumbai, Chennai, and Kolkata.

The motivation behind creating the "Indian Railway Express Trains Delay Datasets" was to address the persistent issue of delays and cancellations experienced by passengers traveling on express trains connecting Guwahati to major metro cities in India.

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The specific purpose of creating this dataset was:

1. **Analyzing Train Performance:** By collecting data on delays and cancellations, the dataset can be used to analyze the performance of various express trains operating between Guwahati and the metro cities of India. This analysis can help identify patterns, trends, and potential areas for improvement in the Indian railway system.
2. **Service Improvement:** Understanding the causes and frequency of delays can help railway authorities in devising strategies to improve service reliability and punctuality. This could involve optimizing schedules, addressing infrastructure issues, or enhancing operational efficiency.
3. **Research and Development:** Researchers and academia can leverage the dataset for studying various aspects of transportation management, logistics, and infrastructure planning. It can serve as a valuable resource for conducting empirical studies and modeling the dynamics of railway operations.

The dataset fills a significant gap by providing detailed information on train routes, delays at each station, and cancellation status for express trains connecting Guwahati to major

metro cities. This granularity allows for a thorough analysis of train performance, which can ultimately contribute to enhancing the efficiency and reliability of India's railway transportation system.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Ankita Anand an undergraduate student at Indian Institute of Technology, Guwahati has created this dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This dataset creation was self-funded

Any other comments?

This dataset is created by scrapping government of India Indian railways website.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in the dataset represent different aspects of train routes, delays, and cancellations for express trains connecting Guwahati (Assam) to major metro cities in India. Each instance corresponds to a specific train route, with detailed information about individual stations along the route, average delay times, categorization of delays in percentages (e.g., right time, slight delay, significant delay), and cancellation status.

The dataset comprises multiple types of instances, primarily focused on:

1. **Train Routes**: Each instance represents a unique train route connecting Guwahati to one of the metro cities, including information such as train number, train name, departure and arrival stations, and train type (e.g., Mail/Express, Superfast, Rajdhani).

2. **Station Delays**: Within each train route, there are instances corresponding to different stations along the route. These instances provide details about the average delay times at each station, categorized into different delay types in percentages (e.g., right time, slight delay, significant delay), as well as information about cancellations or unknown statuses.

How many instances are there in total (of each type, if appropriate)?

Train Routes: There are 42 trains instances, each representing a unique train route connecting Guwahati to one of the metro cities (Delhi, Mumbai, Chennai, Kolkata). Each instance contains details such as train number, train name, departure and arrival stations, and train type.

Station Delays: The number of instances corresponding to station delays will vary depending on the number of stations along each train route and the direction of travel (Guwahati to metro cities or vice versa). Since the dataset includes delays at each station,

the total number of instances for station delays will be determined by summing up the delays recorded at each station for all 42 train routes.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not.

The dataset is a sample since it only contain possible instances of express trains connecting Guwahati to metro cities. The larger set would include all express train routes and their delay information within the Indian railway system. Since the dataset was created by scraping the Indian government's railway website, it represents a subset of available data.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The dataset consists of two parts: a CSV file named "Train List" containing information about express trains connecting Guwahati to metro cities, and a folder containing CSV files with train route details and delay information. The features in the dataset include train number, name, departure and arrival stations, type of service, and delay statistics for each station along the route, expressed as percentages. Here's a description of the data included in each instance:

Train Route Information stored in "Train_List.csv":

- **Train Number**: Unique identifier for the train.
- **Train Name**: Name or designation of the train.
- **From Station**: Departure station of the train route (e.g., Guwahati).
- **To Station**: Arrival station of the train route (e.g., Delhi, Mumbai, Chennai, Kolkata).
- **Type**: Type of train service (e.g., Mail/Express, Superfast, Rajdhani, Tejas, Humsafar).

Station Delay Information stored in folder "Train_Route".:

- **Station**: Identifier for the station along the train route.
- **Station Name**: Name of the station.
- **Average Delay (%)**: Average delay time experienced at the station, expressed as a percentage.
- **Right Time (0-15 min's) (%)**: Percentage of instances where the train arrived or departed within 0-15 minutes of the scheduled time.
- **Slight Delay (15-60 min's) (%)**: Percentage of instances where the train experienced a slight delay (15-60 minutes) at the station.
- **Significant Delay (¿1 Hour) (%)**: Percentage of instances where the train experienced a significant delay (¿1 hour) at the station.

- **Cancelled/Unknown (%)**: Percentage of instances where the train service was canceled at the station or the status is unknown.

Is there a label or target associated with each instance? If so, please provide a description.
No, there isn't an explicit label or target associated with each instance in the dataset. Each instance contains features describing train routes, station delays, and cancellation statuses.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No , there is no information missing from idividual instances.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
Yes, relationships between individual instances are made explicit in the dataset. Specifically, the relationships are established between train routes and the corresponding delay information at each station along the route. This allows for the visualization and analysis of stations with maximum delays over the course of the journey for all trains. By linking the delay information to specific train routes and stations, the dataset facilitates the identification of patterns, trends, and variations in delay occurrences, enabling stakeholders to make informed decisions regarding train operations, scheduling, and infrastructure improvements.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
No, recommended data splits are not provided in the dataset. However, for training purposes, a data split can be created to predict the average delay at the final destination of each train. This split typically involves dividing the dataset into training, validation (or development), and testing sets.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
Given that the data has been manually verified due to its limited size, there are likely minimal errors, sources of noise, or redundancies in the dataset. Manual verification typically helps ensure data accuracy and reduces the likelihood of errors or noise.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources

and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset relies on external resources in the form of data scraped from the Indian government's Indian Railway website. Regarding guarantees of their existence and stability over time, it's essential to consider the maintenance and updates of the website by the Indian government. While there may not be explicit guarantees, government websites often strive for continuity and reliability in providing public information.

As for archival versions of the complete dataset, it's advisable to create snapshots or backups of the data to ensure access to the information as it existed at the time of scraping. This can help mitigate any potential changes or updates to the external resources.

Regarding restrictions associated with the external resources, it's crucial to adhere to any terms of use or licensing agreements specified by the Indian government for accessing and using data from their website. Users should review the terms of use to ensure compliance with any restrictions, licenses, or fees that may apply. Unfortunately, without specific links or access points to the Indian Railway website's data policies, it's challenging to provide detailed descriptions of any such restrictions.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, the dataset does not contain data that might be considered confidential. It consists of publicly available information scraped from the Indian government's Indian Railway website, which typically includes data related to train routes, delays, and cancellations. This data does not involve any confidential or sensitive information such as personal communications or legally privileged data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, the dataset does not contain data that might be offensive, insulting, threatening, or otherwise cause anxiety. It primarily consists of factual information related to train routes, delays, and cancellations, which are unlikely to contain content that could be considered offensive or anxiety-inducing.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No, the dataset does not directly relate to people. It primarily contains information about train routes, delays, and cancellations, without directly involving personal data or individual identities. **Therefore, the remaining questions in this section can be skipped.**

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Not Applicable.

Is it possible to identify individuals (i.e., one or more natural persons),

either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Not Applicable.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Not Applicable.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance was acquired through web scraping of the Indian government's Indian Railway website. The information was directly observable on the website, including details about train routes, station delays, and cancellation statuses. As the data was publicly available on the website, there was no need for validation or verification beyond ensuring the accuracy of the scraping process. However, manual verification have been conducted to confirm the integrity and accuracy of the scraped data, especially considering its importance for train scheduling and passenger information.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was collected using a software program specifically designed for web scraping (i.e. Selenium). This program automated the process of extracting information from the Indian government's Indian Railway website, including details about train routes, station delays, and cancellation statuses. The mechanisms and procedures used for data collection were validated through several steps:

1. **Testing**: The web scraping program was thoroughly tested to ensure its functionality and accuracy in extracting the desired information from the website.

2. **Validation Checks**: Post-scraping validation checks were performed to verify the integrity and consistency of the collected data. This involved comparing the scraped data against the original website content to ensure accurate extraction.

3. **Error Handling**: Mechanisms were implemented to handle errors and exceptions that may arise during the scraping process, such as network issues or changes in website layout.

4. **Iterative Improvement**: Continuous monitoring and refinement of the scraping process were conducted to address any issues or discrepancies encountered during data collection. This iterative approach helped enhance the reliability and effectiveness of the data collection procedures.

Overall, the validation of the mechanisms and procedures used for data collection involved rigorous testing, validation checks, error handling, and iterative improvement to ensure the accuracy and integrity of the collected dataset.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset consists of comprehensive listing of all express trains running between Guwahati and major metro cities in India, including both directions (Guwahati to metro cities and vice versa). Therefore, it represents a specific subset of train routes rather than a random or probabilistic sample. The sampling strategy can be considered deterministic, as it includes all relevant train routes connecting Guwahati to metro cities, albeit focusing on specific routes rather than encompassing all routes across India.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process was conducted by Ankita Anand as an individual initiative for submission to a course project at IIT Guwahati. Since it was a personal project, there were no external individuals involved in the data collection process, such as students, crowdworkers, or contractors. Therefore, no compensation was provided to any external parties.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The timeframe in which the data was collected corresponds to the duration of the web scraping process conducted by me for my university course project which is between January to April in 2024. Therefore, the data collection timeframe aligns with the creation timeframe of the dataset associated with the instances.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Since the dataset was created as an individual initiative for a university course project, it's unlikely that it underwent formal ethical review processes by an institutional review board (IRB). Typically, such review processes are required for research involving human subjects or sensitive data, particularly when conducted within an academic or institutional setting.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No, the dataset does not directly relate to people. **Therefore, the remaining questions in this section can be skipped.**

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
Not Applicable.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Not Applicable.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Not Applicable.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
Not Applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
Not Applicable.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
No, since the dataset was created through web scraping of publicly available information from the Indian government's Indian Railway website, only some basic preprocessing and cleaning were done during the data collection process.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Indian Railway Trains Delay Dataset this is the GitHub repository link where dataset has been uploaded and it publicly accessible.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Not Applicable.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No, the dataset has not been used for any tasks currently.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
Not Applicable.

What (other) tasks could the dataset be used for?
The dataset could be used for various tasks beyond its primary purpose. Some potential tasks include:

1. **Predictive Modeling**: Utilizing the dataset to build predictive models to forecast train delays or cancellations based on historical data, weather conditions, or other relevant factors.

2. **Performance Analysis**: Analyzing the performance of different train routes and stations to identify trends, patterns, and areas for improvement in terms of punctuality and reliability.

3. **Route Optimization**: Using the dataset to optimize train schedules and routes to minimize delays, improve efficiency, and enhance passenger satisfaction.

4. **Decision Support**: Providing insights from the dataset to support decision-making processes for railway authorities, operators, and policymakers regarding infrastructure investments, capacity planning, and service improvements.

5. **Customer Information**: Developing tools or applications that utilize the dataset to provide real-time information to passengers about train delays, cancellations, and alternative travel options.

6. **Research**: Serving as a valuable resource for academic research in transportation engineering, logistics management, and data analytics, enabling studies on factors influencing train delays and performance.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
One potential consideration for future users is the relevance of the dataset over time. Since the data is not continuously updated, there is a risk that it may become outdated, especially if there are changes in train schedules, routes, or performance metrics. Future users should be aware of this limitation and consider the temporal relevance of the dataset when using it for analysis or decision-making purposes.

To mitigate the risk of using outdated data, future users could periodically check for updates from reliable sources, such as official railway websites or government publications, to ensure that their analyses and conclusions are based on the most current information available. Additionally, they

could explore techniques such as data augmentation or simulation to extend the dataset's relevance and account for potential changes in train operations over time.

Are there tasks for which the dataset should not be used? If so, please provide a description.

While the dataset is valuable for analyzing train routes, delays, and cancellations, there are certain tasks for which it may not be suitable:

Real-time Decision Making: Due to potential delays in data updates and the static nature of the dataset, it may not be appropriate for real-time decision-making tasks such as operational scheduling or emergency response.

Generalization to Other Regions: The dataset focuses specifically on express trains connecting Guwahati to major metro cities in India. Therefore, it may not be representative of train operations in other regions or for different types of train services, limiting its generalizability for broader analyses.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Since the dataset was created by scraping details from different pages of the Indian Railways website as an individual initiative, it's unlikely that it will be distributed to third parties outside of the entity on behalf of which the dataset was created. Typically, datasets obtained through web scraping are used for personal or academic purposes rather than for distribution to external parties. Therefore, there may not be plans for distribution beyond the scope of the original project or initiative.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

As the dataset was created as an individual initiative for a university course project and there may not be plans for widespread distribution, it does not have a formal distribution method such as a tarball on a website, API, but it has a GitHub repository. Since the dataset may not be intended for formal publication or widespread distribution, it does not have a digital object identifier (DOI), which is typically assigned to datasets that are published or archived through formal channels.

When will the dataset be distributed?

Not Applicable.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No, since the dataset was created through web scraping of publicly available information from the Indian government's Indian Railway website, it's important to consider any applicable terms of use (ToU) or copyright restric-

tions associated with the website's content.

Typically, data obtained through web scraping may be subject to the website's terms of use, which may specify permissible uses of the data and any restrictions on redistribution or commercial use. Users should review the website's terms of use to ensure compliance with any applicable licensing terms or restrictions.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

As the dataset was created through web scraping of publicly available information from the Indian government's Indian Railway website, it's important to consider any restrictions imposed by the website's terms of use or other third-party agreements.

Typically, third-party websites may impose restrictions on the use of their data, including limitations on redistribution, commercial use, or modification. Users should review the terms of use specified by the Indian Railway website to understand any applicable restrictions or licensing terms associated with the data.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

As the dataset primarily consists of publicly available information scraped from the Indian government's Indian Railway website, it's unlikely to be subject to export controls or regulatory restrictions. However, users should be aware of any applicable laws or regulations governing the use and redistribution of data obtained through web scraping, particularly if they intend to export the data to jurisdictions with different legal frameworks. Since the dataset may not be subject to specific export controls or regulatory restrictions, there is no supporting documentation or links to provide.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

No one will be maintaining it.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Email : guptankitaanand@gmail.com

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

No.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of

time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Not Applicable.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

No.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

No.