Big Data Analytics for Analyzing and Predicting Transportation Incidents

***Abstract*—This study employs advanced data analytics to address hazardous material transportation incidents, integrating visualizations, machine learning, and geospatial techniques. Temporal analysis revealed a rising trend in incidents, with spikes in 2020 and 2021, and seasonal peaks during summer due to weather and transport activity. Geographic mapping identified California and Texas as leading states for incidents, influenced by industrial activity and infrastructure, while Memphis emerged as the city with the highest incidents. Analysis of transport modes showed that highways accounted for 90% of incidents, emphasizing their prominence in hazardous material movement. Failure analyses revealed closures like caps and plugs were the most frequent failure points, with leaking responsible for 50**

Machine learning models, including Decision Trees, Random Forests, and Logistic Regression, were applied to predict failure likelihood. An optimized Decision Tree model achieved 70% accuracy using GridSearchCV, with key predictors such as "What Failed," "Causes of Failure," and container capacity. Insights identified recurring issues, including loose closures responsible for 24,263 failures and forklift-related incidents accounting for 4,039 cases. These findings highlight areas for improvement, including enhanced handling procedures, stricter quality control, and targeted safety measures. By leveraging predictive analytics, this research offers practical recommendations, including collaboration with policymakers and transportation stakeholders to mitigate risks and bolster system resilience.

***Index Terms*—Data Analytics, Machine Learning, Decision Trees, Random Forest, Logistic Regression.**

I. Introduction

THE transportation industry is undergoing a significant transformation driven by the integration of big data analytics, providing powerful tools to improve safety, efficiency, and decision-making processes. With the increasing volume of data generated from sources like traffic sensors, vehicles, infrastructure, and human interactions, transportation systems are becoming more data-driven. This wealth of information, when properly analyzed, offers opportunities to enhance transportation safety, optimize traffic flow, and prevent accidents, particularly in high-risk sectors such as hazardous materials (hazmat) transportation. In particular, machine learning, predictive modeling, and geospatial analysis have emerged as central techniques in analyzing transportation incident data, ultimately aiming to reduce risks and improve operational safety.

Big data analytics in transportation primarily serves the purpose of providing actionable insights from large datasets that would otherwise remain underutilized. Machine learning, predictive models, and geospatial analysis are essential techniques that enable real-time risk assessment and proactive measures to mitigate incidents. These tools allow for the development of models that predict where accidents might occur, identify high-risk zones, and recommend interventions that can prevent or reduce the severity of accidents. As transportation systems continue to evolve, it is crucial to assess how these technologies can be applied to minimize incidents, particularly in the transportation of hazardous materials, which carry unique risks due to their potential environmental and societal impact.

II. Background

This section provides a comprehensive overview of the basic definitions and terminologies related to the approaches and methods employed in the research that focus on the incident datasets. Understanding these fundamental concepts is crucial for grasping the subsequent analyses and findings presented in the research. The dataset represents incidents containing information about transporting goods and materials. By exploring the definitions and terminology, readers will acquire a strong understanding of the research technique used, which will help them to understand the complexities of the ensuing analyses and findings.

A. Descriptive Statistics:

Descriptive statistics is a subfield that focuses on characterizing and summarising data. It involves demonstrating the key features of a dataset using graphical representations and numerical measurements. Data patterns can be seen with the aid of visualizations such as bar charts and histograms. In various fields, descriptive statistics are crucial for data exploration and decision-making. It facilitates comparisons and generalizations based on the data as well as the efficient sharing of study findings. The use of descriptive statistics as basic data can serve as the basis for further research by describing initial issues or highlighting crucial analyses in additional studies[1].

B. Random Forest

In order to generate predictions, this ensemble learning technique integrates several decision trees. Random Forest is renowned for its capacity to manage intricate information and generate reliable forecasts. In order to improve accuracy and decrease overfitting, multiple decision trees are created using random selections of data and features, and the results are combined. It handles complex relationships [2].

C. Logistic Regression:

Another statistical method using one or more independent variables is logistic regression, which is used to estimate the probability of a binary occurrence, in this context, of success or failure. It is a popular and widely used classification

algorithm in machine learning and statistics. Finding the best-fitting model that connects the independent variables to the likelihood of the binary outcome is the aim of logistic regression

D. Decision Tree

Decision trees are hierarchical models that provide a structured approach to problem-solving and decision-making. A group of instances can be turned into a decision tree by employing a divide-and-conquer strategy. In the event that every instance belongs to the same class, the tree is labeled as a leaf with that class. Otherwise, the instances are split based on the test's outcome, which is chosen if it produces different outcomes for at least two of the cases.

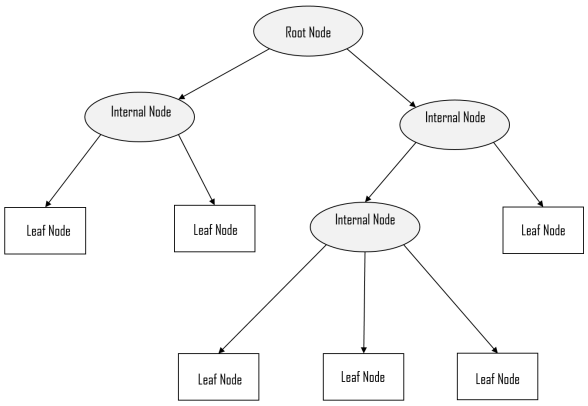

Fig. 1. Decision Tree Structure

Utilizing many algorithms, including entropy-based and information gain-based methods, to construct decision trees. Decision trees enable researchers and practitioners to comprehend intricate relationships within data, classify instances, and make predictions based on learned patterns [3]. The construction of decision trees using different algorithms, such as entropy-based and information gain-based approaches.

$$Entropy = -\sum_{n=1} pi\ log(pi) \tag{1}$$

$$GiniImpurity = 1 - \sum_{n=1} pi^2 \tag{2}$$

where:

- Entropy represents the entropy of the dataset.
- p(i) is the proportion of instances belonging to class i in the dataset.
- The summation is taken over all distinct classes in the dataset.

The entropy-based approach quantifies the impurity or disorder in a dataset using the concept of entropy. This allows for fair attribute comparison, considering the number of distinct values an attribute can take.

E. Hyperparameters

Hyperparameters are parameters set before the learning process begins and not learned from the data. They are essential in shaping the behavior and performance of a machine learning model. Grid search optimization involves evaluating all possible combinations of hyperparameter values within a predefined range, systematically exploring the hyperparameter space.

F. LIME (Local Interpretable Model-agnostic Explanations) Method

A widely-used interpretability technique, LIME (Local Interpretable Model-agnostic Explanations), offers a method to understand the factors contributing to a specific model prediction. LIME can provide insights into the key determinants of an incident, such as the factors that contribute to the outcome. LIME maintains interpretability by generating explanations based on a simplified, human-understandable data representation. [4] These explanations are simpler because they demonstrate a closer relationship between the input and prediction.

III. Literature Review

Big data analytics plays a pivotal role in the field of transportation, providing invaluable insights into improving safety, efficiency, and decision-making processes. With the transportation industry continuously generating large amounts of data from diverse sources such as traffic sensors, vehicles, infrastructure, and human interactions, advanced data analytics techniques are increasingly leveraged to analyze this data. In particular, the application of machine learning, predictive models, and geospatial analysis are central to the optimization of transportation systems. These techniques not only enable real-time risk assessment but also foster proactive measures to mitigate incidents before they occur. This literature review examines how big data analytics is utilized in the analysis of transportation incident data, with a focus on the development of models aimed at enhancing transportation safety.

A key application of big data analytics in transportation is predictive modeling. By analyzing historical incident data, machine learning algorithms such as Random Forest, Decision Trees, and Logistic Regression are employed to forecast potential risks and prevent accidents. For example, predictive models have been applied in urban accident forecasting, where data from traffic cameras, sensors, and past accident reports are used to predict where future accidents are likely to occur. These models consider various factors such as traffic flow, road conditions, and weather patterns. The effectiveness of these models in improving road safety by anticipating accident hotspots, allowing city planners and safety authorities to take preemptive actions, such as altering traffic signals or modifying road designs[5].

Geospatial analysis is another crucial technique utilized in the transportation sector. By combining geographical data with transportation incident records, this technique helps to identify accident-prone areas and assess the risks associated

with different routes. Geospatial algorithms, including spatial autocorrelation methods and Geographic Information Systems (GIS)-based tools, enable the visualization of incidents over geographic regions, thus pinpointing high-risk zones. This is particularly important in hazardous materials (HazMat) transportation, where route selection plays a critical role in minimizing risks. Research has shown that optimizing routes for hazardous materials transportation can significantly reduce the risk of accidents while minimizing environmental and societal impacts[6]. These findings underline the value of geospatial analysis in mitigating transportation risks, especially in the case of hazardous material shipments, which often require careful route planning to avoid populated areas and critical infrastructures.

The use of clustering algorithms in analyzing transportation data has also gained traction in identifying patterns and groups of similar incidents. Clustering techniques, such as K-Means and DBSCAN, are employed to categorize transportation incidents based on features like location, time, and cause, helping to identify recurring safety issues. Research has shown that rerouting and improving the design of tank cars can significantly reduce the risk associated with transporting hazardous materials by rail [7]. By clustering incidents, transportation analysts can pinpoint specific areas or situations where accidents are more likely, enabling targeted interventions such as infrastructure improvements or enhanced safety measures. Furthermore, clustering methods facilitate the discovery of hidden patterns that could be overlooked in traditional analysis, helping to enhance predictive capabilities and improve overall safety management.

Predictive maintenance, made possible by big data analytics, is another vital application in the transportation industry. In sectors such as aviation and rail, predictive maintenance models analyze data from vehicle sensors to identify potential failures before they occur. By analyzing large datasets, researchers have found that it's possible to predict when aircraft components may fail, allowing for proactive maintenance and reducing the risk of accidents [8]. Similarly, predictive analytics can be used to monitor the health of rail cars, enabling timely maintenance and reducing the likelihood of accidents caused by equipment failures [9]. These predictive models rely on advanced machine learning techniques to analyze patterns in sensor data and historical maintenance records, allowing for a more proactive approach to equipment management. The analysis of unstructured data is also a growing area of interest in transportation safety. Many transportation datasets include unstructured text, such as accident reports, safety audits, and driver behavior logs. Natural Language Processing (NLP) techniques, including text classification and topic modeling, are applied to extract meaningful insights from these unstructured data sources. For example, NLP algorithms can classify accident reports into categories such as mechanical failure, human error, or environmental factors, aiding in the identification of recurring themes or emerging risks. These techniques allow transportation agencies to perform root cause analysis and improve decision-making based on a comprehensive understanding of incident patterns. Natural language processing tools can analyze safety reports to identify potential hazards in hazardous materials transportation, leading to improved risk assessments and safety protocols [10].

Anomaly detection is another key technique employed in big data analytics for transportation risk management. Algorithms such as Isolation Forest, One-Class SVM, and autoencoders are used to detect unusual patterns or rare events that may indicate emerging risks or data errors. This is especially important in transportation incident analysis, where outliers or anomalies in the data could signal significant safety concerns or system failures. For instance, anomalous spikes in accident rates could point to issues with infrastructure or operational practices that need to be addressed promptly. The ability of anomaly detection algorithms to identify these rare occurrences ensures that even low-probability events are considered in risk assessments and mitigation strategies. Time series analysis is another method extensively used in the transportation industry to examine the temporal patterns of incidents over time. Time series forecasting models, such as ARIMA and Long Short-Term Memory (LSTM) networks, help identify trends, seasonal effects, and periodicity in incident data. For instance, certain accident patterns may emerge during specific seasons or times of day, prompting targeted interventions such as adjusting traffic control measures or deploying additional resources. By analyzing historical incident data over time, transportation agencies can develop predictive models that inform strategies to reduce accident frequency during peak times or adverse weather conditions.

The integration of big data analytics with simulation models has further enhanced transportation safety efforts. Simulation models, including Monte Carlo simulations and agent-based models, are used to assess the potential outcomes of various interventions and risk scenarios. Monte Carlo simulations, for instance, provide probabilistic risk assessments by simulating multiple scenarios based on different input parameters. These models are particularly useful in evaluating the risks associated with the transportation of hazardous materials, where the consequences of accidents can be severe. To improve transportation safety, researchers have employed various modeling techniques. Monte Carlo simulations can be used to evaluate the risks associated with specific scenarios, while agent-based models can simulate the behavior of individual actors to understand how their actions may influence overall safety [7]. As the transportation industry continues to embrace big data, it is important to address the challenges associated with data integration, privacy concerns, and computational efficiency. Data from diverse sources, such as traffic cameras, vehicle telematics, and environmental sensors, must be integrated and processed in real time to provide actionable insights. Moreover, ensuring the privacy and security of sensitive data, such as personal information or proprietary vehicle data, remains a significant concern. Computational challenges also arise due to the large volume of data, necessitating the development of efficient algorithms and systems capable of processing and analyzing big data in a timely manner. In conclusion, the integration of big data analytics in transportation provides significant benefits in terms of improving safety, optimizing operational efficiency, and reducing risks associated with incidents. Predictive models, geospatial analysis, clustering,

anomaly detection, and simulation models have proven to be invaluable tools in transportation incident analysis. By leveraging these techniques, transportation agencies can proactively identify risks, develop effective safety measures, and enhance the overall performance of transportation systems. However, addressing the challenges of data integration, privacy, and computational efficiency will be essential to fully realize the potential of big data analytics in the transportation sector.

IV. METHODOLOGY

A. *Methodology Overview*

This study utilizes a systematic approach to analyze transportation incident data, specifically focusing on hazardous material (hazmat) incidents, by leveraging big data analytics. The methodology is designed to uncover trends, build predictive models, and generate actionable insights to improve transportation safety. The analysis is structured into several key stages: data collection, data preprocessing, exploratory data analysis (EDA), predictive modeling, statistical testing, and visualization of results.

The primary dataset used in this analysis is the Hazmat Incident Database provided by the Pipeline and Hazardous Materials Safety Administration (PHMSA), which contains historical records of hazmat transportation incidents [11]. This dataset includes detailed information on each incident, such as time, location, material type, and cause. To enrich the dataset and provide additional context, supplementary data is incorporated, including weather reports, traffic flow patterns, and geographic features.

B. *Dataset*

Data preprocessing is a critical step to ensure the integrity and consistency of the dataset. Missing values in both categorical and numerical features are addressed through imputation or removal strategies. Specifically, missing values in categorical columns (e.g., "What Failed" and "Causes of Failure") are filled using corresponding values from related columns. For numerical features, such as "Quantity Released," missing values are replaced with the mean or median value of the respective column. Furthermore, variables such as geospatial coordinates and date formats are standardized to ensure consistency and readiness for analysis. Data preprocessing is a critical step to ensure the integrity and consistency of the dataset. Missing values in both categorical and numerical features are addressed through imputation or removal strategies. Specifically, missing values in categorical columns (e.g., "What Failed" and "Causes of Failure") are filled using corresponding values from related columns. For numerical features, such as "Quantity Released," missing values are replaced with the mean or median value of the respective column. Outliers in numerical data are identified using the Interquartile Range (IQR) method, and these outliers are replaced with the median value of the corresponding column. Furthermore, variables such as geospatial coordinates and date formats are standardized to ensure consistency and readiness for analysis.

- Features
- Missing Values
- Duplicates

Column	Null Count	Total Count
Incident City	3	100000
Incident State	491	100000
Date Of Incident	0	100000
Incident Time	275	100000
Quantity Released	0	100000
Commodity Long Name	835	100000
Hazardous Class	753	100000
Total Hazmat Fatalities	0	100000
Total Damages	0	100000
Shipper Name	16	100000
Mode Of Transportation	0	100000
Cont1 Packaging Type	35210	100000
Cont1 Package Capacity	0	100000
What Failed	17565	100000
How Failed	22354	100000
Causes of Failure	19211	100000
Incident Result	1	100000

TABLE I
DATA SUMMARY FOR VARIOUS COLUMNS

The data summary table reveals that while most features have minimal missing data, columns like "Cont1 Packaging Type" "What Failed" "How Failed" and "Causes of Failure" exhibit significant null values. This highlights the need for careful data cleaning and potential imputation or removal of these features based on the specific analysis goals and the impact of missing data on the results. The data was cleaned and preprocessed to address missing values and inconsistencies. Missing values in the "What Failed" and "Causes of Failure" columns were imputed using information from the other column. For instance, if one column was missing, it was filled with the value from the other column. The "How Failed" column was imputed with the most frequent value to maintain consistency. Categorical variables like "Incident City" and "Commodity Long Name" were filled with "mode" and "Unknown" to indicate missing information. Missing values in the numerical columns "Quantity Released" and "Cont1 Package Capacity" were addressed using imputation techniques. The mean was used to fill missing values in "Quantity Released," assuming missingness was random. For "Cont1 Package Capacity," the median was used to mitigate the impact of potential outliers. These imputation strategies ensured that the numerical data was complete and ready for further analysis and modeling. These data cleaning steps ensured the data was ready for further analysis.

C. *Exploratory Data Analysis (EDA) and Statistical Analysis*

Column	Mean	Std	25th	Median	75th	Max
Qty Rlsd	386.18	921.03	0.00	0.06	0.47	2.00
Total Haz	0.00	0.03	0.00	0.00	0.00	0.00
Total Dmgs	669.72	163.73	0.00	0.00	0.00	0.00
Cont1 Pck	153.90	954.62	0.00	0.00	4.00	10.00

TABLE II
DESCRIPTIVE STATISTICS FOR VARIOUS COLUMNS

The descriptive statistics table summarizes the key features of the data. While some variables like ”Quantity Released” show a wide range of values, others like ”Total Hazmat” and ”Total Damages” are highly concentrated around zero. This suggests potential data imbalances.

D. *Goodness of fit test*

This will typically be done by some goodness-of-fit statistical techniques that calculate a match between the model and real data to come up with whether such differences between the expected and observed values are statistically significant or otherwise. One of the techniques in conducting this kind of analysis is called the Chi-squared test. This technique will develop an understanding of the model on applicability and reliability aspects. In the context of this project, such datasets have been analyzed using statistical approaches that included Chi-squared tests. Results are shown below, including variable names and the critical values providing the decision point for each variable.

Variable	Chi-square Statistic	p-value
Mode of Transportation	30.91	0.0000886
Incident Result	644.31	0.0001203
Incident City	8.01	1.0
Incident State	109.25	0.00018807
Commodity Long Name	33.58	1.0
Shipper Name	1.11	1.0
Causes of Failure	87.07	1.0
Cont1 Packaging Type	46.95	0.000005714

TABLE III
CHI-SQUARED TEST RESULTS FOR VARIOUS VARIABLES

The Chi-squared test results reveal that variables like “Mode of Transportation,” “Incident Result,” “Incident State,” and “Cont1 Packaging Type” show significant differences, as indicated by their low p-values, suggesting potential issues with the model’s fit. In contrast, variables such as “Incident City,” “Commodity Long Name,” “Shipper Name,” and “Causes of Failure” exhibit no significant differences.

E. *Data Visualization*

To gain a comprehensive understanding of the dataset, a series of data visualizations will be employed. Time series analysis will be conducted to identify temporal trends and seasonal patterns in incident occurrences. Additionally, visualizations will be created to examine the distribution of incidents across different shippers, highlighting those with a higher frequency of incidents. To delve deeper into the root causes of incidents, visualizations will be generated to analyze the frequency and types of failures, as well as the associated causes. Furthermore, geographic visualizations will be employed to map the spatial distribution of incidents across cities and states, identifying regions with higher incident rates and potential geographic clusters. Through these visualizations, underlying patterns and relationships within the data can be uncovered, informing the development of effective strategies for preventing and mitigating hazardous material transportation incidents.

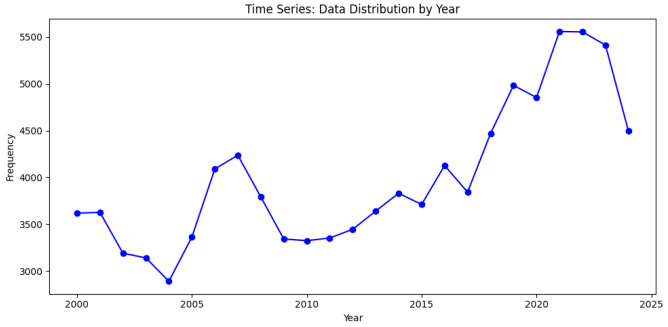

Fig. 2. Decision Tree Structure

The time series analysis reveals an increasing trend in incident frequency over the years, punctuated by significant fluctuations. The years 2020 and 2021 stand out as periods of peak activity.

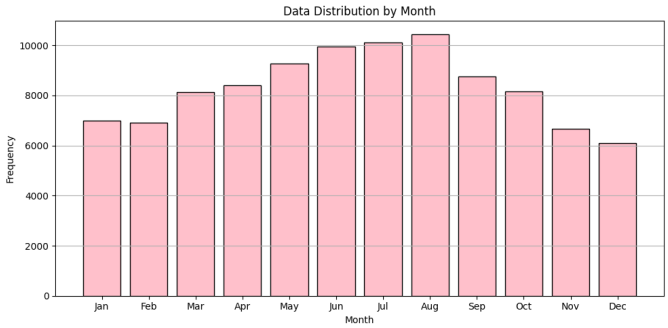

Fig. 3. Monthly Incident Frequency

The bar plot illustrates a clear seasonal pattern in incident frequency, with peaks during the summer months. This seasonal variation may be influenced by factors such as weather conditions, transportation patterns, and specific events.

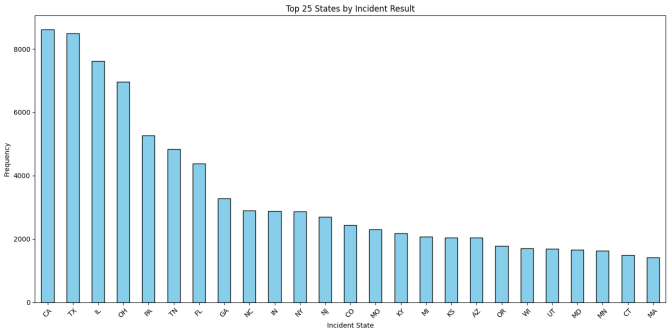

Fig. 4. Frequency of States Modes

The bar chart highlights the geographic distribution of incidents, with California and Texas experiencing the highest frequency. This visualization suggests a concentration of incidents in certain states, potentially influenced by factors such as population density, industrial activity, regulatory environment, and infrastructure.

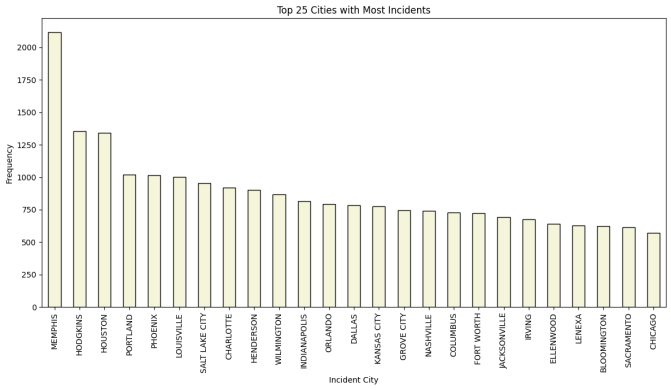

Fig. 5. Frequency of Cities

The bar chart reveals a concentration of incidents in specific cities, with Memphis leading the list. This visualization, combined with the temporal and geographic trends observed in the previous plots, suggests that certain cities may have unique factors contributing to their higher incident rates.

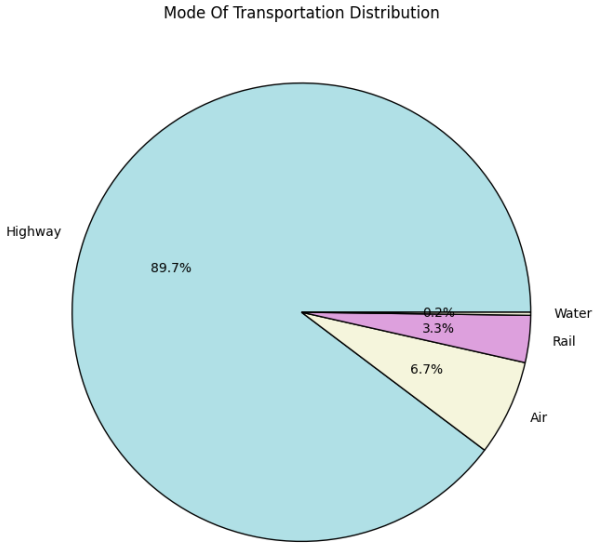

Fig. 6. Pie Chart for Transportation

The pie chart reveals that highway transportation is the dominant mode of transportation for hazardous materials, accounting for nearly 90'%' of incidents. To gain a deeper understanding of the factors contributing to these variations, further analysis was conducted, focusing on regulatory frameworks, infrastructure, and operational practices.

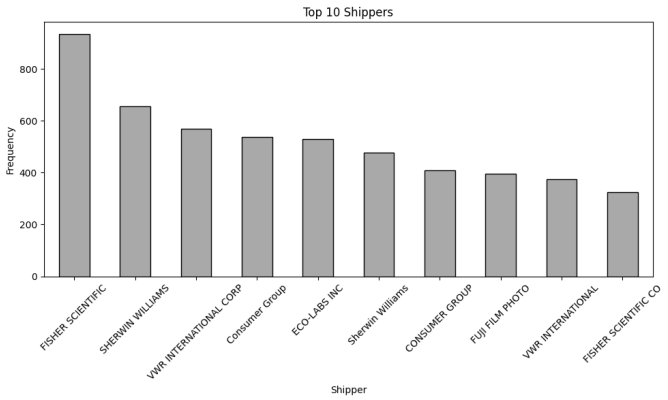

Fig. 7. Frequency of Incidents by Top 10 Shippers

The bar chart highlights the top 10 shippers with the highest number of incidents. Fisher Scientific emerges as the leading shipper in terms of incident frequency. This visualization suggests that certain shippers may have unique factors contributing to their higher incident rates.

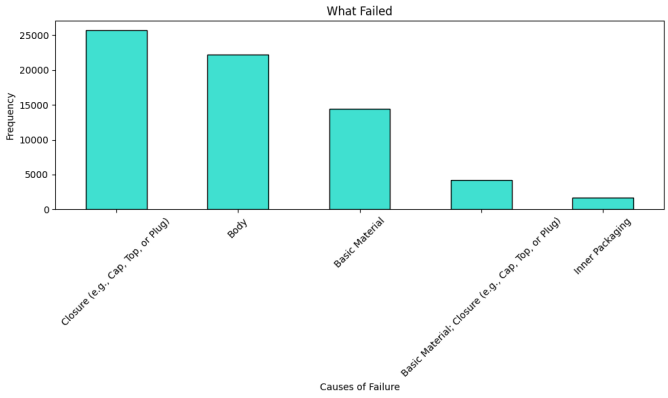

Fig. 8. Frequency of Failures by Component Type

The bar chart titled "What Failed" illustrates the frequency of incidents associated with different types of failures. "Closure (e.g., Cap, Top, or Plug)" emerges as the most common type of failure, followed by "Body" and "Basic Material." This visualization provides insights into the specific components that are most prone to failure.

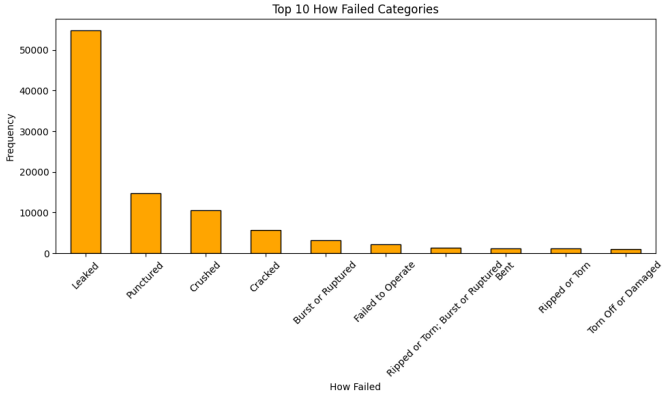

Fig. 9. Frequency of Failure Modes

The bar chart titled "Top 10 How Failed Categories" illustrates the frequency of different types of failures. "Leaked" emerges as the most common type of failure, followed by "Punctured" and "Crushed." This visualization provides insights into the specific ways in which components fail, highlighting areas where targeted preventive measures can be implemented.

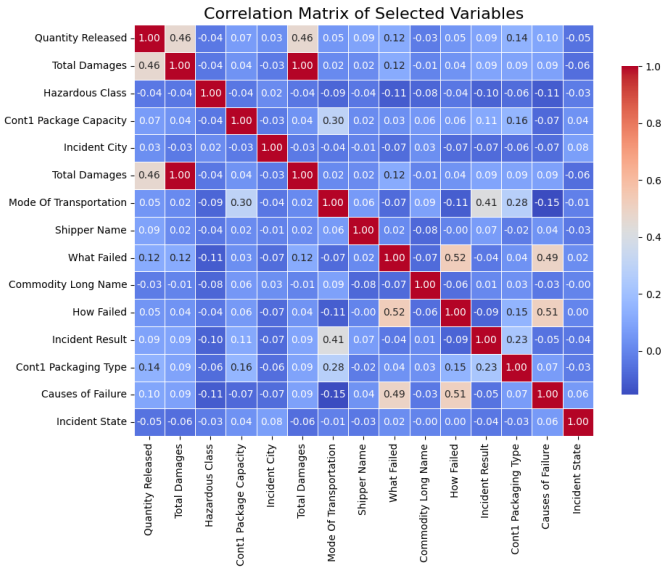

Fig. 10. Correlation Matrix of Selected Variables

The correlation matrix reveals several interesting relationships between variables, including strong positive correlations between container capacity and failure causes, as well as between mode of transportation and incident result. Additionally, strong negative correlations were observed between quantity released and total damages. These correlations suggest that certain factors, such as container size, mode of transportation, and quantity of hazardous material, may influence the likelihood and severity of incidents.

F. *Feature Engineering*

Feature engineering is a critical step in the data processing pipeline. It has to do with transforming raw data into meaningful features that can improve the performance of machine learning models. This process encompasses dimensionality reduction, feature selection, and feature extraction techniques. One key feature engineering application is predictive modeling, which helps refine the feature space and reduce modeling errors.

1) Feature Selection: Feature selection was performed using a combination of correlation analysis and statistical tests. Pearson correlation was employed to identify and remove highly correlated variables, reducing redundancy in the dataset. Additionally, F-tests were used to select numerical features most strongly associated with the target variable. For categorical features, techniques such as chi-squared tests or information gain could be utilized to identify significant variables for inclusion in machine learning.

G. *Dimentionality Reduction*

In order to gather a substantial amount of data, a threshold of 70% for the cumulative proportion of variance explained will be applied in this study as a criterion for choosing the main components. This threshold makes it possible to reconcile keeping enough information in the dataset with lowering its dimensionality and complexity. Principal components that account for at least 70% of the variance will be retained. Once the threshold is met and the next component does not contribute much, it will not be included because, in PCA, the fewer components, the better.

*1) **Principal Component Analysis (PCA)**:* From the dataset, the number of pca components correspond to the number of filtered variables. Each component describes the percentage of variance in the dataset. The cumulative proportion explains the total variance.

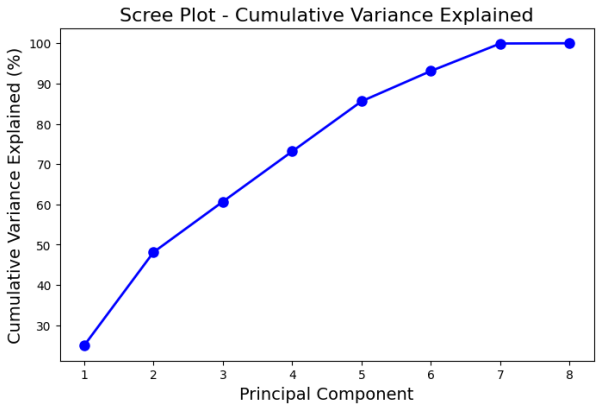

Fig. 11. Variance Explained by Principal Components

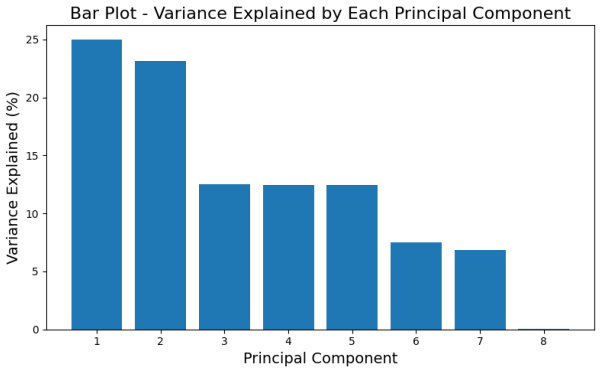

Fig. 12. Cumulative Variance Explained

The bar and scree plots illustrate the proportion of variance captured by each principal component in the dataset. The first two principal components explain approximately 45% of the total variance, with the first component alone capturing over 25%. The first four principal components cumulatively explain around 70% of the variance. Given this, retaining the first four principal components as features in further analysis is justified, as they collectively capture a significant portion of the data's variability.

H. Machine Learning Models

Predictive modeling is central to understanding future risks associated with transportation incidents. Several machine learning algorithms, including Logistic Regression, Random Forests, and Decision Trees, are applied to predict the likelihood of incidents based on features such as time of day, weather conditions, and material type. The dataset is split into training and testing subsets (80 for training and 20 for testing), and the models are evaluated based on performance metrics such as accuracy, precision, recall, and F1-score. Feature importance is assessed using techniques to identify which factors most significantly contribute to incident occurrences.

*1) **Exploring the Applicable Models** :* In this section, an in-depth analysis will be conducted based on accuracy, precision, recall, and F1 score of the logistic regression, decision tree, and random forest models. The performance of these models will be investigated in different domains, analyzing their strengths and weaknesses based on empirical results. This will help understand the suitability of each model for the problem in question. The prediction helps make informed decisions by choosing the most suitable machine learning approach.

- **Decision Trees:** By recursively dividing the data according to many criteria, the decision tree approach seeks to maximise information gain or decrease impurity at each split as it constructs the tree. The attributes with the highest discriminatory power are then selected as the decision nodes after a stopping condition is met, such as reaching a maximum depth or acquiring a minimum amount of data points in a leaf node. This was done across the dataset to assess the model's correctness.
- **Random Forest:** This is an ensemble learning method that combines multiple decision trees to make predictions. Random Forest also provides feature importance insights, making it a versatile and reliable model for complex datasets. It is highly suitable for the analysis due to its robustness and ability to handle complex relationships in the data.
- **Logistic Regression:** Logistic Regression is a statistical model for binary classification that predicts probabilities using a logistic function. It is efficient, interpretable, and works well with linearly separable data. While simple, it is suitable for the analysis as a baseline model to understand feature relationships and ensure interpretability.

I. Model Fine-tuning and Optimization

The Fine-tuning and Optimization module is a crucial part of the machine learning process. It focuses on adapting pre-trained models to specific tasks and refining model parameters to achieve optimal performance. This module delves into various techniques, including optimization algorithms and hyperparameter tuning. By mastering this module, we can significantly enhance the efficiency and effectiveness of machine learning models across diverse applications.

V. RESULTS AND DISCUSSIONS

This section reveals the outcomes of our comprehensive machine learning tests. The training set is used to train the model, while the testing set is used to evaluate its performance and generalization ability on unseen data. In this research, 'How Failed' is the target variable for prediction following comprehensive feature engineering and selecting relevant variables. The variable 'How Failed' is well-suited for machine learning as it encapsulates distinct failure modes, offering valuable insights for addressing and preventing system failures. Furthermore, its categorical nature makes it compatible with classification algorithms, facilitating the identification of patterns and predictive modeling based on historical data. The dataset was split into a training set and a testing set with a ratio of 80% training data and 20% testing data. Below are outputs of the three machine learning test. These metrics are the F1-score, precision, recall, and accuracy as shown in the equations.

1) Accuracy:

$$Accuracy = \frac{TP + TrueNegatives(TN)}{TP + TN + FP + FN} \tag{3}$$

2) Precision:

$$Precision = \frac{TruePositive(TP)}{TP + FalsePositive} \tag{4}$$

3) Recall:

$$Recall = \frac{TruePositive(TP)}{TP + FalseNegative(FN)} \tag{5}$$

4) Area under the Curve (AUC): where:

- TP - True Positives
- FP - False Positives
- TN - True Negatives
- FN - False Negative

A. Machine Learning performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.62	0.62	0.62	0.56
Decision Tree	0.64	0.64	0.68	0.66
Logistic Regression	0.62	0.59	0.60	0.61

TABLE IV
MODEL PERFORMANCE METRICS

The table presents the performance metrics of the machine learning models, the models were evaluated using accuracy, precision, recall, and F1-score.

Key Observations

- Decision Tree: Outperforms the other models in terms of accuracy, precision, recall, and F1-score.
- Random Forest: Shows comparable performance to the Decision Tree, especially in terms of accuracy and precision.
- Logistic Regression: While it achieves a relatively high accuracy, its precision and recall are lower compared to the other models.

Based on these results, the Decision Tree model will be chosen for further analysis and optimization. While it currently performs best, further fine-tuning hyperparameters and exploring additional techniques may lead to even better results.

B. *Model Fine-tuning and Optimization*

This section will detail machine learning optimization and fine-tuning principles and procedures for the Decision tree. How different hyperparameters, optimization, and fine-tuning strategies affect model performance will be investigated.

1) ***Decision Tree****:* Hyperparameter tuning through GridSearchCV significantly improves the model's performance, increasing accuracy from 64% to 70%. GridSearchCV works by systematically testing different combinations of hyperparameters to identify the best-performing set. It does this by performing an exhaustive search over a specified parameter grid and evaluating the model's performance using cross-validation. This process helps to optimize the model, making it more robust and generalized, as it is tested across various configurations before settling on the best one. The advantage of using GridSearchCV is that it automates the process of hyperparameter selection, saving time and effort while ensuring the model performs at its best. The Feature Importance Plot identifies the top 7 features that contribute the most to the model's predictions, providing valuable insights into which aspects of the data are most influential. Additionally, the Learning Curve visually represents how the model's performance improves as more training data is used, offering a deeper understanding of the model's learning process and its potential for further enhancement.

Hyperparameter	Average Values Used
max_depth	10
min_samples_split	5
min_samples_leaf	2
max_features	sqrt

TABLE V
AVERAGE HYPERPARAMETER VALUES USED IN GRIDSEARCHCV

Tuned Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.70	0.64	0.69	0.65

TABLE VI
TUNED MODEL PERFORMANCE METRICS

The improvement of the tuned model is a significant achievement, as it enhances the model's ability to make more accurate predictions. Through the use of GridSearchCV, the model's hyperparameters were optimized, leading to a notable increase in accuracy from 64% to 70%.

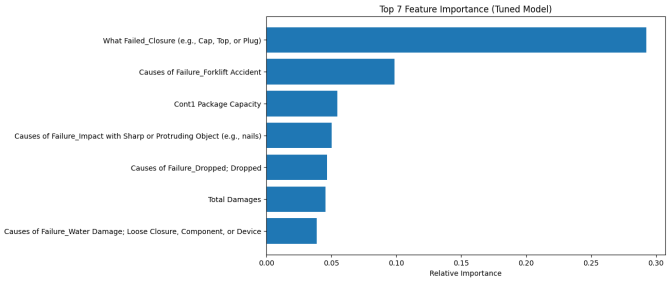

Fig. 13. Decision Tree Feature Importance

The feature importance plot highlights the top 7 features that significantly influence the model's predictions. The most influential feature is What Failed, indicating that the type of closure failure plays a crucial role in determining the overall failure mode. Other important features include "Causes of Failure", "Cont1 Package Capacity", and various causes related to impact, dropping, and water damage. These findings suggest that factors such as packaging capacity, handling practices, and environmental conditions contribute significantly to the occurrence of specific failure modes.

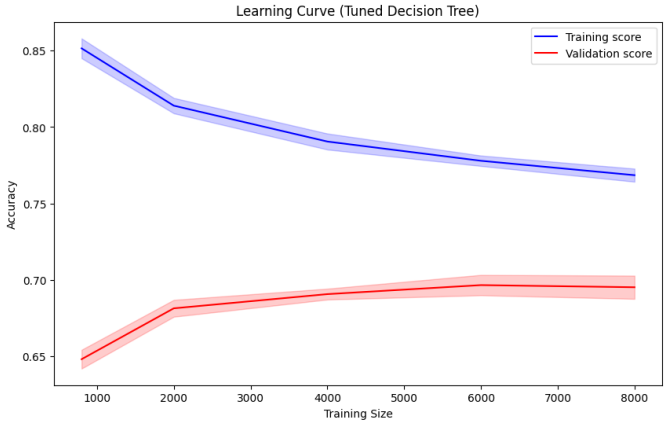

Fig. 14. Tuned Decision Tree learning curve

The learning curve illustrates the performance of the tuned decision tree model as the training size increases. Training accuracy starts high 85% and gradually decreases, while validation accuracy improves steadily and stabilizes around 70%. This indicates that the model generalizes better with more data, though a slight gap between training and validation scores suggests minor overfitting. The narrow shaded regions highlight low variance across folds, demonstrating consistent and reliable model performance. Overall, the model shows assertive learning behavior and benefits from additional data.

*2) **Sample Instances And LIME**:*

Fig. 15. Local Interpretable Model-Agnostic Explanations Output

The Local Interpretable Model-Agnostic Explanations (LIME) framework was used to interpret the predictions of a decision tree model, providing transparency into why specific outcomes were assigned to individual data points. LIME works by approximating the decision boundaries of a complex model locally, identifying the most influential features for a given prediction. This method is beneficial for understanding black-box models, as it highlights the contributions of individual features to specific outcomes.The results offer insights into the decision tree model's predictions for specific instances. The model assigns probabilities to different failure modes, with the highest probability indicating the most likely outcome.

For example, in the first instance, "Leaked" is predicted with a probability of 0.50, suggesting it is the most probable failure mode. The model's predictions are influenced by factors such as shipper name, incident city, quantity released, and container package capacity. The analysis highlights that the Shipper Name was consistently a critical feature across predictions. The most frequent outcome was leaked in two instances, suggesting it may be a dominant failure mode in the dataset. LIME effectively demonstrated how the decision tree model utilized key variables, making it easier to understand its predictions and offering actionable insights for mitigating risks in shipment management. This information can be leveraged to develop strategies for preventing future incidents and improving overall system reliability.

*3) **Analysis of Common Failure Scenarios**:* This section highlights the most frequent causes of operational failures, their associated components, modes of failure, and impact in terms of frequency. The data underscores the need for improvements in handling, transportation, and quality assurance practices to mitigate these recurring issues.

Causes of Failure	What Failed
Loose Closure, Component, or Device	Closure
Forklift Accident	Body
Human Error	Closure
Improper Preparation for Transportation	Body
Water Damage; Loose Closure, Component, or Device	Basic Material
Dropped	Closure
Impact with Sharp or Protruding Object (e.g., Nail)	Body
Defective Component or Device	Closure

TABLE VII
CAUSES OF FAILURE AND WHAT FAILED

The analysis highlights several key failure scenarios affecting operations. The most frequent issue, with 24,263 occurrences, was due to loose closures, components, or devices, where closures like caps or plugs failed, leading to significant leaks. Another critical problem attributed to forklift accidents (4,039 cases) involved the puncturing of container bodies during handling. Improper preparation for transportation resulted in 2,531 failures, where inadequate stacking or bracing caused containers to be crushed. Environmental factors also played a role, as water damage combined with loose closures or components led to 2,464 leak incidents, exacerbating container vulnerabilities. Physical impacts with sharp or protruding objects, such as nails, caused 1,958 punctures, highlighting a need for enhanced protective measures. Human error-related failures, which accounted for 782 incidents, were mainly related to poor handling or preparation, which led to leaks that further compromised the container's integrity. Finally, 1,892 failures were caused by defective devices or components, including closures, highlighting the significance of strict quality control procedures. These results highlight the need for better handling, shipping, training, and quality control procedures.

VI. CONCLUSION

This research adopts big data analytics to analyze accidents in hazardous material transportation to enhance safety and operational efficiency. Among these predictive models that were experimented with, it was observed that, after optimization through GridSearchCV, the highest accuracy of 70% belonged to the Decision Tree. Logistic Regression and Random Forest provided strong baselines, whereas the Decision Tree outperformed them regarding recall and F1-score, proving effective. Other critical incident factors present were "What Failed," "Causes of Failure," and "Container Package Capacity," indicating the importance of resolving the issues on loose closures and defective components to reduce risks.

Temporal analysis showed a trend of incidents that were on the rise, peaking during the summer months. Geographic clustering in states like California and Texas underlined the influence of industrial density on incident rates. Providing transparency to actionable insights, interpretable models such

as LIME were used to enable deep insight into the findings for operational improvement and proactive risk mitigation strategies.

The different solutions offered to enhance hazardous material transportation safety need to address these challenges. Improving transportation protocols through better packaging standards and periodic inspections is fundamental to reducing failures. Comprehensive training programs are also fundamental, as they reduce human error and improve the preparedness of personnel handling such operations. Other actions include upgrading container designs with protective features to prevent failure by environmental factors or physical impacts. At the same time, implementing a real-time monitoring system plays the most important role in quickly finding abnormalities and taking immediate measures.

Finally, collaboration with policymakers is important for targeted intervention in higher-risk regions and seasons. These targeted interventions should focus on improved infrastructure and revised regulations that address the identified geographic and temporal challenges. Together, these measures can reduce incidence rates, minimize risks, and provide for safer transport operations.

A. *Future Work*

Future work will consider the increase in the robustness and accuracy of the prediction by trying more advanced machine learning models such as Gradient Boosting or Neural Networks and ensemble methods. Augmentation of feature sets using expansion of the dataset with real-time sensor data, weather conditions, and traffic dynamics will increase predictive power. The insights drawn from this research will inform regulatory frameworks about high-risk regions and seasonal variations, while standardized training programs could reduce human-error-induced failures. Real-time anomaly detection systems using techniques such as Isolation Forests or autoencoders will be implemented to identify risks in advance. Scaling the analysis pipeline to cloud-based platforms will ensure the system can handle larger datasets and provide real-time predictions. Methodologies could also be extended to other modes of transportation, like aviation or rail, where similar patterns and risks in data exist. These initiatives will refine the current findings, develop safety measures, and establish a proactive framework for dealing with transportation incidents effectively.

REFERENCES

[1] Emily Johnson Brian Conner. Descriptive statistics, 2017.
[2] Julián Chaparro-Peláez Ángel Hernández García. Predicting teamwork group assessment using log data-based learning analytics, 2018.
[3] J. R. QUINLA. Learning decision tree classifiers, 1996.
[4] Simon Dixon Saumitra Mishra, Bob L. Sturm. Local interpretable model-agnostic explanations for music content analysis, 2017.
[5] Ahmed, S., Rahman, A. (2022). Predictive models for urban accident forecasting using machine learning. Transportation Research Part C: Emerging Technologies.
[6] Kawprasert, A., Barkan, C. P. L. (2008). Effects of route rationalization on hazardous materials transportation risk. Journal of Transportation Research.
[7] Rapik Saat, M., Barkan, C. P. L. (2011). The effect of rerouting and tank car safety design on the risk of rail transport of hazardous materials. Transportation Research Record.
[8] Rajasekaran, K., et al. (2021). Enhancing aircraft safety through predictive maintenance using big data. Aerospace Engineering Review.
[9] Liu, X., Saat, M. R., Barkan, C. P. L. (2020). Big data analytics for railway safety risk assessment. Journal of Transportation Safety and Security.
[10] Liu, X., et al. (2013). Railroad hazardous materials transportation risk analysis: A review of current methods and emerging directions. Journal of Hazardous Materials.
[11] U.S. Department of Transportation. (n.d.). Hazardous materials regulations. Retrieved from PHMSA.