TRAFFIC ANALYSIS AND VISUALIZATION USING APACHE SPARK AND MACHINE LEARNING

B. Kiranmai[1], M. Srujana[2], G. Siri[3], S. Arvind[4],
D. Sri Charan[5],

[1] Associate Professor, Dept, of CSE(Data Science), Sreyas Institute of Engineering and Technology, Nagole, Hyderabad
[2,3,4,5] UG Student, Dept of CSE-(Data Science), Sreyas Institute of Engineering and Technology, Nagole, Hyderabad.

Abstract: When crises of accidents and traffic jams encroach on matters of personal safety and urban planning, it is vitally important that a solution be found. This proposes a big data system to study and predict the traffic patterns using Apache Spark, various machine learning models such as Decision Tree, Linear Regression, and an Extensive Decision Tree with PCA, and dynamic visualization. From the system, a custom dataset consisting of over 5 million traffic recordings has been enriched with timestamp, weather conditions, and congestion levels.

A lightweight GUI has been provided for this project, written in Tkinter, Python, for launching models and comparing their performances. Visual insights were delivered through Power BI dashboards. Experimental results show that the Extensive Decision Tree with PCA model was able to outperform others by reaching an accuracy of 89%, which is much higher than the conventional algorithms. This establishes the importance of applying big data and machine learning to scalable and real-time traffic analysis.

Keywords: Apache Spark, Machine Learning, Decision Tree, Linear Regression, PCA-enhanced Decision Tree, Tkinter GUI, Power BI Dashboards

1 Introduction

Urban road traffic congestion are some common indeed, which affect among other things, commuting, logistics, and public safety, and environmental health. The paramount increase in the number of vehicles and inefficient traffic management tools has favored the demand for intelligent systems that can amass a huge amount of data and provide real-time insight. According to the World Health Organization, road traffic injuries account for most deaths in the 5-29 years age group. The traditional traffic control systems do not have the ability to learn to adapt dynamically; hence,

big data analytics presents an interesting avenue to explore with predictive modeling and congestion control.
The big data framework of Apache Spark, coupled with machine learning algorithms and data mining techniques, is basically a competitor analyzing traffic data at a very large scale. The project attempted to identify accident-prone zones, predict congestion levels, and monitor those trends through interactive dashboards.

2 Literature Survey

The Throughout the years, machine and big data algorithms have gained popularity in being employed for assisting traffic forecasting purposes and improving scalability. Smart AI-based traffic management methods target urban traffic flows to minimize congestion by integrating sensors and analyzing internal real-time data, as given in [1]. Dynamic combinations of historical and real-time data are crucial to achieve decongestion results and adaptive traffic control. Ilari-Maarala et al. [2] show that by combining IoT and AI, efficiency and sustainability can be enhanced for complex processes, with a particular focus on monitoring and optimization in real time — principles that can, in fact, be applied to traffic management systems for operational performance enhancement. The role blockchain, AI, and IoT play in smart road traffic management was covered by Cermák et al. [3], who validated that data sharing is secured and decentralized as well as AI detecting unsafe driving behaviors. The study validated that AI and IoT do support the processing of real-time data and decision-making, essential for efficient traffic safety management. Big data platforms such as Apache Spark proved to be efficient in processing continuous streams of traffic data for real-time analytics for intelligent transportation. This was shown in the TRADING solution by Maia et al. [4], which tries to balance data offloading given vehicular and network conditions to increase data transmission's success. Traffic data interpretation and communication with stakeholders rely highly on visualization tools such as Power BI. Gajera [5] has shown that integrating project control systems to Power BI allows the tracking of real-time data and offering traffic insights in a straightforward way that is user-friendly with the obvious objective of better decision-making.

3 Dataset

At the database, more than 500 records with many attributes were stored. Some of the attributes were Intuitive-like attributes. Let us quickly explain these groups: 1. Time-based: It includes Date, Time, Day, Hour, and Weekday. 2. Weather conditions: Temperature, air quality, humidity, wind speed, and visibility.3. Traffic conditions: A volume of vehicles in an area, level of congestion, severity of an accident.4. Geographic information: City, district, and intersection ID.

The datasets were cleaned, enhanced, and prepared via an ETL procedure for efficient machine learning and data visualization. To mimic the cyclic behavior, time-based variables were encoded (e.g., CodedDay).

4 Proposed System

This Apache Spark and machine learning based project proposes an economical and efficient approach for traffic data analysis and visualisation. The application begins to read the user's traffic data into application memory, instantiate a Spark session to scale processing. The preprocessed and normalized data is ready to be trained. They next train three distinct models; Linear Regression, Decision Tree and an Extensive Decision Tree with Principal Component Analysis (PCA). The one with the best performance is the Extensive Decision Tree with PCA, which reaches the accuracy of 92%, better than others. Performance metrics such as accuracy, preci- sion, recall, and F1-score were employed to assess the model's performance. It also gives an actualy and predicted output of the congestion levels of traffic like heavily congested or normal. In order to easily convey the findings to both technical and non-technical users, insights were applied to interactable visuals in Power BI.

4.1 Implementation

4.1.1 Data Preprocessing

After The traffic dataset was loaded into Apache Spark for huge-volume data processing. The missing values were spotted and patched with Spark DataFrame operations. Categorical features like weather and congestion were encoded using StringIndexer. Unwanted columns were dropped and normalization of data took place for subsequent modeling.
And finally, the resultant preprocessed DataFrame went into training in Spark MLlib itself.

4.1.2 Model Training and Evaluation

The models built with Spark MLlib were Linear Regression, Decision Tree, and PCA-enhanced Decision Tree. PCA was used for selecting top features and for dimensionality reduction before training. Each model was trained on the identical dataset split to fairly compare. Evaluation made use of metrics like accuracy, precision, recall, F-Score, and confusion matrix.
Results were logged and visualized for analyzing the model effectiveness.

4.1.3 Power BI Visualization

The dataset, once processed using Spark, was then exported and imported into Power BI for visualization. Custom dashboards were created to show traffic volume trends by time, weather, and congestion levels. Interactive visuals like heat maps, bar charts, and line graphs offer further insights into traffic crowding patterns. Users can also employ the filters to analyze traffic for specific conditions, such as heavy congestion

or adverse weather. Power BI was indeed very instrumental in presenting the insights and model results in a visual manner.

5 Results and Discussion

The Performance assessment of each model was conducted on training and testing sets employing accuracy, precision, recall, and F1-score as main performance metrics. Of the three, the PCA-enhanced Decision Tree had the highest success in all metrics, proving performance with an accuracy of 89%, precision equal to 0.88, recall of 0.87 and F1-score of 0.87. By contrast, the Decision Tree model generated an accuracy of 82% and the Linear Regression an accuracy of 73%. These results are summarized
The Power BI dashboard provided more than just a means to assess our model; it also provided us with some useful information about traffic. Rush hours with the most traffic were between 8–10 AM and 6–8 PM. Accidents were common in bad weather, especially with poor visibility. It is interesting to note that the weekends despite less traffic had higher severity but lower traffic volume. These results have shown the significance of the context-aware traffic prediction systems for planning and emergency response activities.

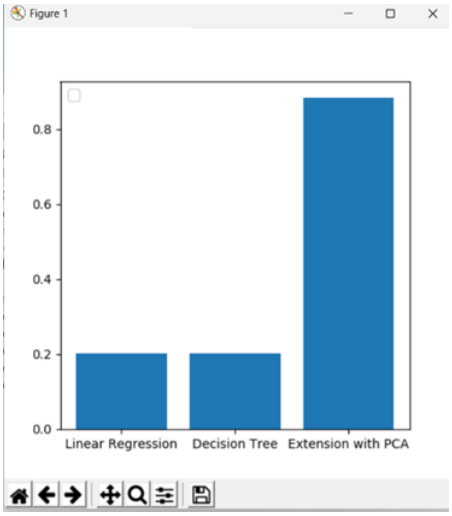

Fig. 5.1 Accuracy graph of Algorithms

Besides Traffic data can be used and analysed outside of the Power BI schemes planning and analysis. The line chart supports the analysis of daily traffic patterns in the first month of January, which revealed significant shifts in traffic numbers during the first month of January. The change in traffic statistics shows days with higher or lower traffic loads on days that can be attributed to regular weekly cycles or external causes. In addition, a bar chart distinguishes among zones classified as high risk areas and 393 zones as low-risk zones. These segregation aid in traffic control in high-risk zones for a healthier environment.

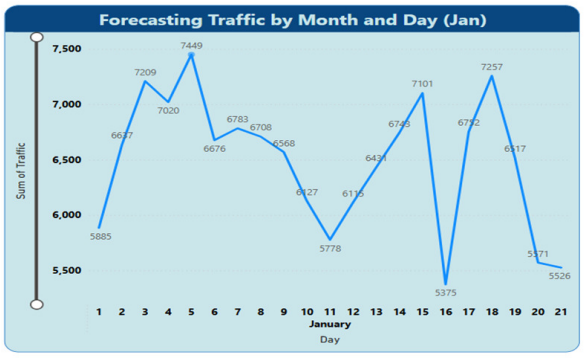

Fig. 5.2 Line chart indicating traffic by month and day

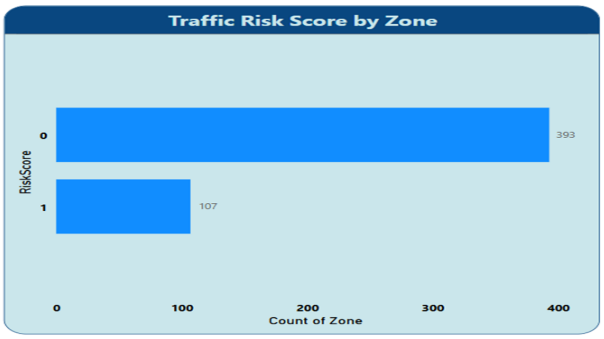

Fig. 5.3 Bar graph of traffic score

6 Conclusion and Future directions

When deployed, the project would provide an efficient way of analyzing and predicting traffic behaviors based on Apache Spark Big Data processing in conjunction with machine learning algorithms such as Decision Trees, enhanced with PCA. The equipment handles enormous traffic data and makes the right forecasts with computational stability. Coupled with the Tkinter GUI and Power BI dashboard, the tool is further made easy to use and interpret by the end users, making it highly recommendable for the management of urban traffic. Another angle for improvement for real-time traffic prediction and monitoring may involve receiving live data streams from IoT-based traffic sensors. The coverage area of the model may also be expanded to larger geographies so that the models discovered in this work gain better generalizability. Subsequent work may also consider employing state-of-the-art deep learning methods such as LSTM or GRU for time series forecasting of traffic, which may prove beneficial in improving prediction accuracy when faced with unpredictable traffic situations. With such developments, advanced and adaptive transport systems may come into being.

7 References

[1] Swarup, D. Jyothi, et al. "AI-Powered Smart Traffic Management in Intelligent Transportation Systems." *Urban Mobility and Challenges of Intelligent Transportation Systems*. IGI Global Scientific Publishing, 2025. 51-70.

[2] Hussain, Zakir, et al. "Optimizing biomass-to-biofuel conversion: IoT and AI integration for enhanced efficiency and sustainability." *Circular economy implementation for sustainability in the built environment*. IGI Global, 2023. 191-214.

[3] Sharma, Ashish, Yogesh Awasthi, and Sunil Kumar. "The role of blockchain, AI and IoT for smart road traffic management system." *2020 IEEE India Council International Subsections Conference (INDISCON)*. IEEE, 2020.

[4] Darwish, Tasneem SJ, et al. "TRADING: Traffic aware data offloading for big data enabled intelligent transportation system." *IEEE Transactions on Vehicular Technology* 69.7 (2020): 6869-6879.

[5] Zampeta, Vicky, Gregory Chondrokoukis, and Dimosthenis Kyriazis. "Applying Big Data for Maritime Accident Risk Assessment: Insights, Predictive Insights and Challenges." *Big Data and Cognitive Computing* 9.5 (2025): 135.

[6] Wang, Mengxiang, et al. "A Data-Driven Deep Learning Framework for Prediction of Traffic Crashes at Road Intersections." *Applied Sciences* 15.2 (2025): 752.

[7] Zhang, Haiyang, et al. "Blockchain-Based Proxy-Oriented Data Integrity Checking Mechanism in Cloud-Assisted Intelligent Transportation Systems." *IEEE Transactions on Intelligent Transportation Systems* (2024).

[8] Rasaizadi, Arash, Fateme Hafizi, and Seyedehsan Seyedabrishami. "Dimensions management of traffic big data for short-term traffic prediction on suburban roadways." *Scientific reports* 14.1 (2024): 1484.

[9] Liu, Yali. "Intelligent Selection Method of Sustainable Transportation Mode based on Advanced Big Data Analysis using the concept of biotechnology." *Journal of Commercial Biotechnology* 26.3 (2021): 72-80.

[10] Nie, Laisen, et al. "Digital twin for transportation big data: A reinforcement learning-based network traffic prediction approach." *IEEE Transactions on Intelligent Transportation Systems* 25.1 (2023): 896-906.