Road Trafic Analysis high or low risk based Machine Learning with SparkMLib

ELADNANI Rihab
Faculty of sciences , UM5
Rabat
rihab_eladnani@um5r.ac.ma

ZYAT Ismail
Faculty of sciences , UM5
Rabat
ismail_zyat@um5r.ac.ma

ISFI Imrane
Faculty of sciences , UM5
Rabat
Imrane_isfi@um5r.ac.ma

ELBAHALI Aya
Faculty of sciences , UM5
Rabat
aya_elbahali@um5r.ac.ma

LAKHDAR Ayoub
Faculty of sciences , UM5
Rabat
lakhdar_ayoub@um5r.ac.ma

***Abstract*— This project focuses on leveraging Big Data analytics and Machine Learning methodologies, specifically utilizing SparkMLib, to conduct a comprehensive analysis of road traffic aimed at assessing and mitigating risks associated with urban transportation. By integrating diverse datasets encompassing geographical information, weather conditions, and temporal factors, the project aims to develop predictive models capable of distinguishing between low and high-risk road segments.**

The integration of geographical data enables spatial analysis of traffic patterns, facilitating the identification of congestion-prone areas and potential hazards. Weather conditions serve as critical determinants in the analysis, providing insights into the impact of meteorological phenomena on road safety. Additionally, temporal factors such as time of day and seasonal variations are considered to unveil temporal trends influencing traffic dynamics and risk levels. Through the utilization of advanced Machine Learning algorithms within SparkMLib, the project seeks to extract meaningful patterns from large-scale traffic data and develop predictive models for risk assessment. These models will enable stakeholders to anticipate and mitigate potential risks proactively, optimizing traffic management strategies and resource allocation.

Overall, this project aims to harness the power of Big Data and Machine Learning to enhance road traffic analysis and contribute to the development of safer and more efficient transportation systems in urban environments. Keywords— big data , traffic , forecast , longitude , latitude…

I. INTRODUCTION

In today's bustling urban landscapes, effective management of road traffic is paramount for ensuring efficient transportation systems and enhancing road safety. However, traditional approaches to traffic analysis often struggle to adapt to the dynamic and multifaceted nature of traffic patterns. To address this challenge, the integration of Big Data analytics and Machine Learning methodologies presents an unprecedented opportunity to revolutionize road traffic analysis.

This project focuses on harnessing the power of Big Data and Machine Learning, particularly leveraging the capabilities of SparkMLib, to conduct a comprehensive analysis of road traffic aimed at distinguishing between low and high-risk road segments. By assimilating diverse datasets encompassing geographical information, weather conditions, and temporal factors, the project endeavors to develop predictive models capable of assessing the safety levels of roads with precision.

At the core of this endeavor lies the utilization of geographical data to understand the spatial distribution of traffic patterns and identify areas prone to congestion or hazards. By correlating traffic flow data with geographical features such as longitude an latitude of each location ,the project seeks to uncover insights into traffic dynamics and risk factors associated with specific locations.

Furthermore, the inclusion of weather conditions as a key determinant in the analysis adds a layer of complexity and nuance to the predictive models. By integrating meteorological data such as temperature, precipitation, and visibility, the project aims to elucidate the impact of weather-related phenomena on road safety, enabling stakeholders to anticipate and mitigate potential risks proactively.

Temporal factors, including time of day, day of the week, and seasonal variations, also play a pivotal role in understanding traffic patterns and assessing risk levels. By analyzing temporal data alongside geographical and weather-related parameters, the project seeks to unveil temporal trends and patterns that influence road safety, facilitating the identification of high-risk periods and locations.

This project focuses to leverage the power of Big Data and Machine Learning, specifically utilizing SparkMLib, to conduct a comprehensive analysis of road traffic with a specific emphasis on assessing the risk levels associated with different segments of the road network. , the project seeks to develop predictive models capable of distinguishing between high and low-risk areas based on various factors such as geographical longitude , geographical latitude , weather conditions, and the time generated .

II. Description and Context of the Project

In today's fast-paced urban landscapes, effective management of road traffic is a critical component of urban infrastructure planning and safety enhancement. However, traditional methods of traffic analysis often struggle to adapt to the complexities of modern traffic dynamics. To address this challenge, this project focuses on harnessing the capabilities of Big Data analytics and Machine Learning, particularly through the use of SparkMLib, to conduct a thorough analysis of road traffic patterns with a specific emphasis on risk assessment.

At the heart of this project lies the integration of diverse datasets, including geographical information, weather conditions, and temporal factors. By associating traffic data with geographical attributes such as road types, intersections, and topography, the project aims to discern spatial traffic patterns and identify areas of high congestion or potential hazards. Additionally, incorporating weather data enables the project to uncover correlations between meteorological events and traffic incidents, providing valuable insights into the impact of weather conditions on road safety.

Temporal factors such as time of day, day of the week, and seasonal variations also play a significant role in traffic analysis. By analyzing temporal data alongside geographical and weather-related parameters, the project seeks to unveil temporal trends influencing traffic dynamics and risk levels. This comprehensive approach allows for the identification of high-risk periods and locations, empowering stakeholders to implement proactive measures to mitigate potential risks and enhance overall road safety.

Through the utilization of advanced Machine Learning algorithms within SparkMLib, the project aims to develop predictive models capable of distinguishing between low and high-risk road segments. These models will leverage the rich dataset encompassing geographical, weather, and temporal information to provide accurate risk assessments, enabling stakeholders to make informed decisions regarding traffic management strategies and resource allocation.

Overall, this project represents a significant step towards leveraging Big Data and Machine Learning technologies to revolutionize road traffic analysis and enhance road safety in urban environments. By integrating geographical, weather, and temporal data to distinguish between low and high-risk road segments, the project aims to provide actionable insights to stakeholders for proactive management of road traffic and the development of more efficient and safer transportation systems.

III. Tools and Models Used

For this project, a combination of advanced tools and models will be utilized to conduct a comprehensive analysis of road traffic and assess risk levels. The primary tools and models include:

A. Apache Spark

Apache Spark is an open-source distributed computing framework that provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. It was developed in the AMPLab at UC Berkeley and later contributed to the Apache Software Foundation.
Apache Spark will serve as the foundational framework for processing large-scale traffic data in parallel and enabling efficient data manipulation and analysis.

B. SparkMLib

Spark MLlib is the machine learning library provided by Apache Spark, designed to simplify and scale machine learning tasks across distributed computing clusters. It offers a wide range of algorithms and tools for building scalable machine learning pipelines.

SparkMLib, a machine learning library within Apache Spark, will be utilized for developing and training predictive models. This library provides a wide range of algorithms for classification, regression, and clustering, making it suitable for various aspects of traffic analysis.

C. Hadoop

Hadoop is an open-source distributed computing framework designed for processing and storing large datasets across clusters of commodity hardware. It provides a scalable and fault-tolerant solution for handling big data applications.
HDFS will serve as the storage system for storing and managing large-scale traffic data collected from various sources. It provides distributed storage across multiple nodes in a Hadoop cluster, enabling scalability and fault tolerance.

Hadoop complements SparkMLib in this project by providing a robust infrastructure for storing, preprocessing, and scaling data processing tasks. By leveraging Hadoop's distributed computing capabilities alongside SparkMLib's advanced machine learning algorithms, the project can analyze large-scale traffic data and develop predictive models for assessing road safety levels effectively.

D. Pandas

Pandas: is a popular open-source data manipulation and analysis library for Python. It provides data structures for efficiently storing and manipulating large datasets, along with tools for data cleaning, exploration, and analysis. Two primary data structures in pandas are:

Series: A one-dimensional array-like object that can hold any data type.

DataFrame: A two-dimensional table of data with rows and columns, similar to a spreadsheet or SQL table.

Pandas is widely used in data science, machine learning, and data analysis tasks. Some of the key features and functionalities of pandas include:

•Data Cleaning: Handling missing data, filtering, and filling operations.

•Data Exploration: Descriptive statistics, summary functions, and data visualization.

•Data Manipulation: Slicing, indexing, merging, and reshaping datasets.

•IO Tools: Reading and writing data from/to various file formats, such as CSV, Excel, SQL databases, and more.

•Time Series Analysis: Specialized tools for working with time-series data.

•Grouping and Aggregation: Grouping data based on some criteria and performing aggregate functions.

E. Matplotlib:

Matplotlib: is a popular 2D plotting library for the Python programming language. It provides a variety of functions for creating static, animated, and interactive visualizations in Python. Matplotlib is widely used in data science, scientific computing, and other fields where data visualization is essential.

Key features of Matplotlib include:

Wide Range of Plot Types: Matplotlib supports a variety of plot types, including line plots, scatter plots, bar plots, histogram plots, 3D plots, and more.

Customization: It allows extensive customization of plots, including control over colors, line styles, markers, labels, and other visual elements.

Publication-Quality Graphics: Matplotlib is designed to create high-quality, publication-ready graphics with fine control over every aspect of the plot.

Integration with Jupyter Notebooks: Matplotlib can be seamlessly integrated with Jupyter notebooks, making it a popular choice for interactive data exploration and visualization.

Multi-platform Support: Matplotlib works on various operating systems, including Windows, macOS, and Linux, and supports different backends for rendering plots.

F. Ploty:

Plotly: is a Python library that enables interactive and web-based data visualization. It supports a wide range of chart types and is particularly well-suited for creating interactive plots and dashboards. Plotly can be used both in offline environments and to generate plots that can be embedded in web applications.

Key features of Plotly include:

1. Interactive Visualizations: Plotly allows users to create interactive plots with features like zooming, panning, and hovering over data points to display additional information.

2. Wide Range of Chart Types: Plotly supports various chart types, including line charts, scatter plots, bar charts, histograms, box plots, 3D plots, heatmaps, and more.

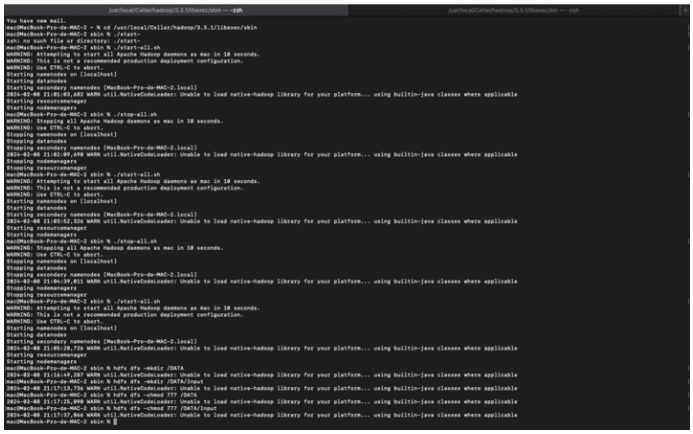
3. Dashboards and Web Applications: Plotly is often used in conjunction with Dash, a web application framework for building interactive dashboards. This combination enables the creation of dynamic and responsive data visualization applications.

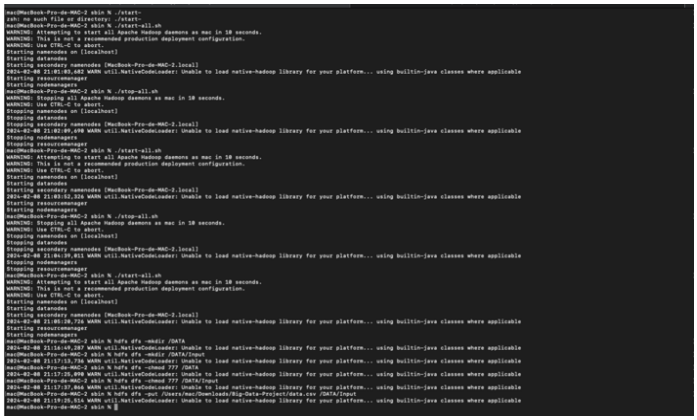
4. Support for Multiple Programming Languages: While Plotly has a Python library, it also provides APIs for other programming languages, including JavaScript, R, and Julia.

5. Offline and Online Usage: Plotly can be used offline, generating static images or HTML files for standalone use. It also provides an online platform, Plotly Chart Studio, where users can create, share, and collaborate on interactive plots.

IV. REALIZATION

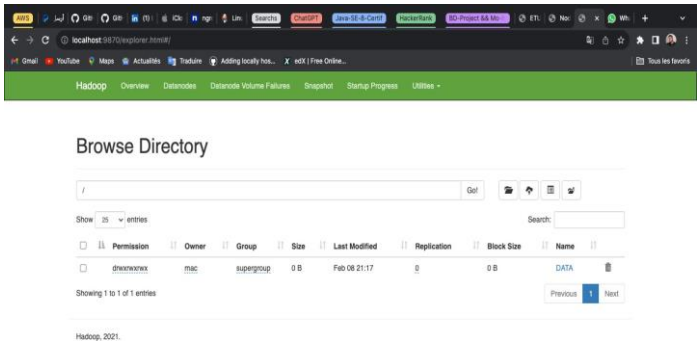
- Starting hadoop and creatingng our required directories in HDFS, adjusted permissions, and uploaded our dataset .CSV to the specified HDFS directory.



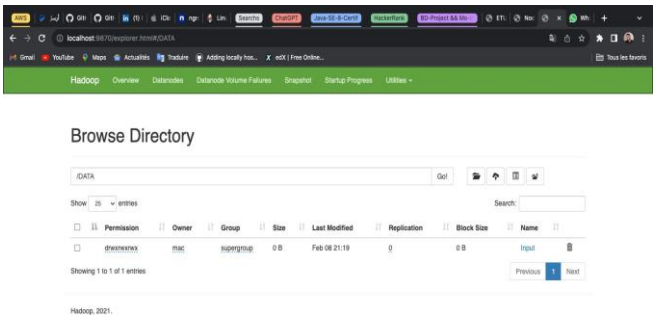


- Permissions:

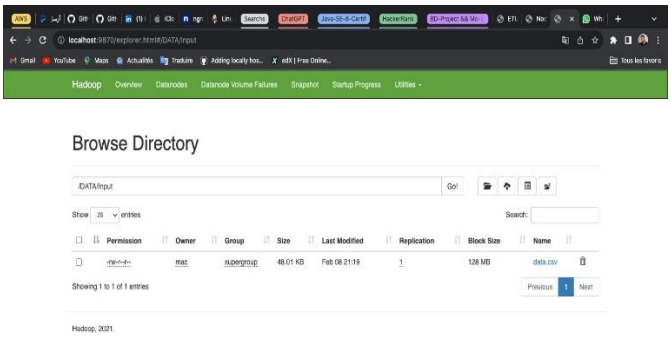
full permission for DATA directory:



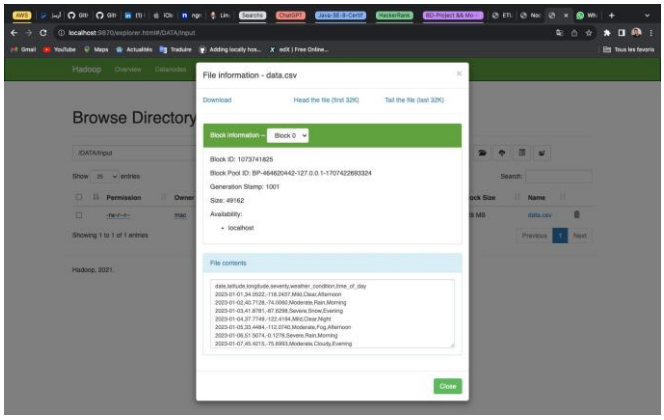
Full permisssions for the input directory:



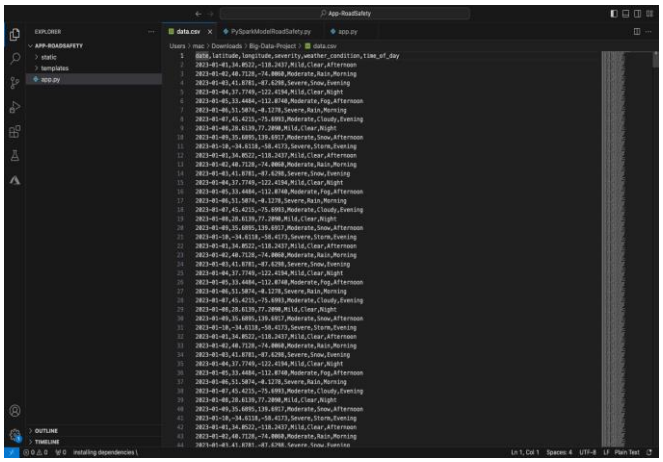
- For the dataset : read and write permissions for the owner, read-only permissions for the group, and read-only permissions for others



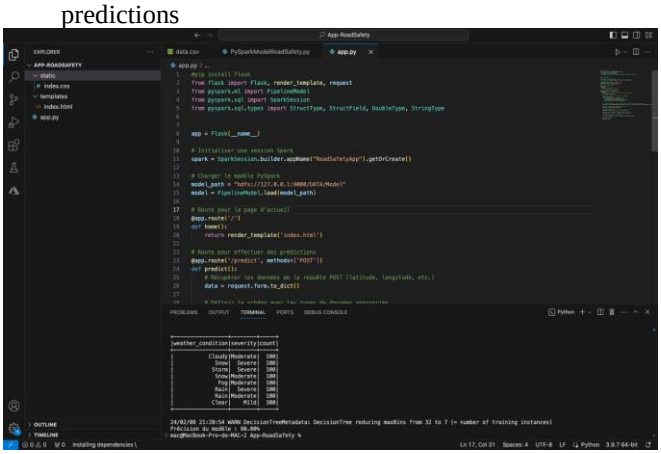
- Information preview for our dataset :

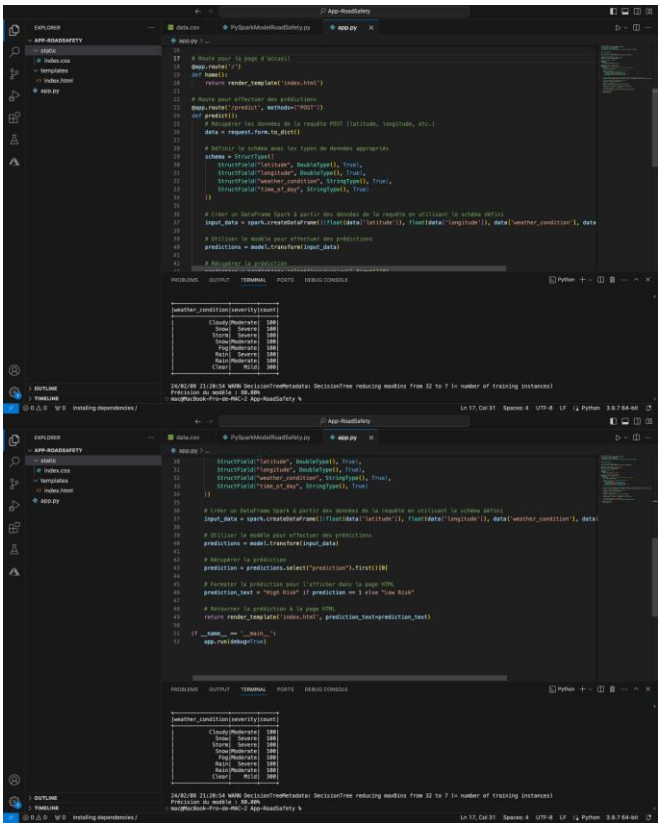


- Dataset content:

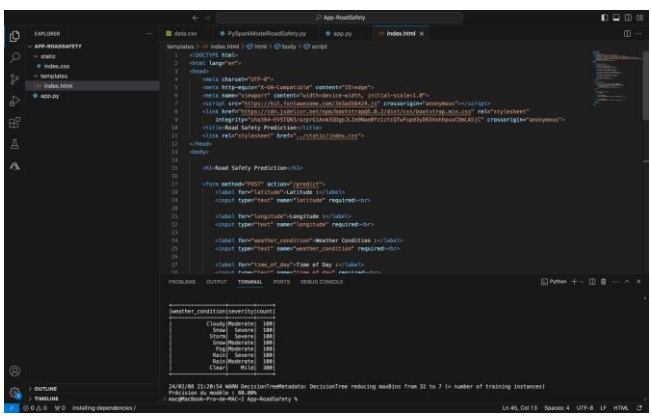


- Our Flask application script for the predictive analytics application using PySpark for roadsafety predictions

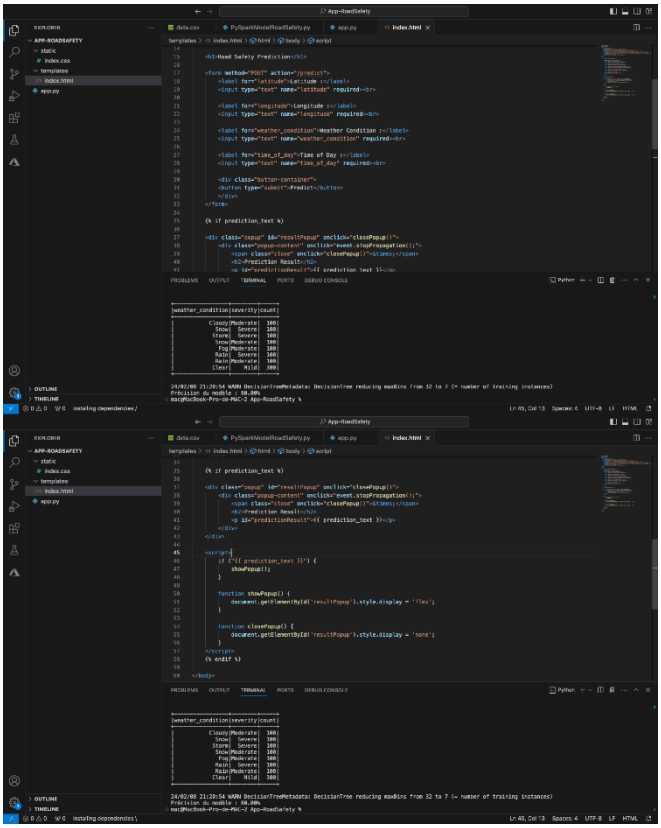


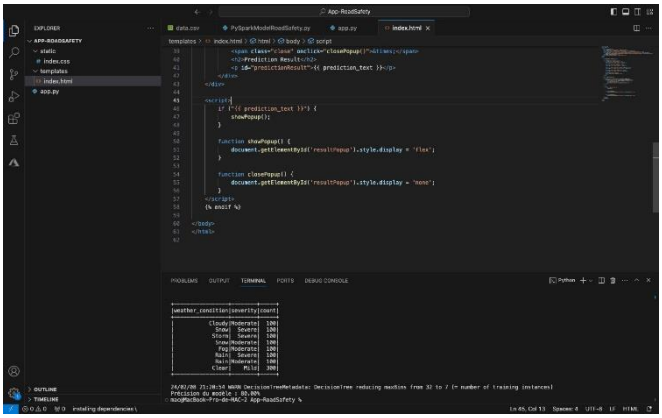


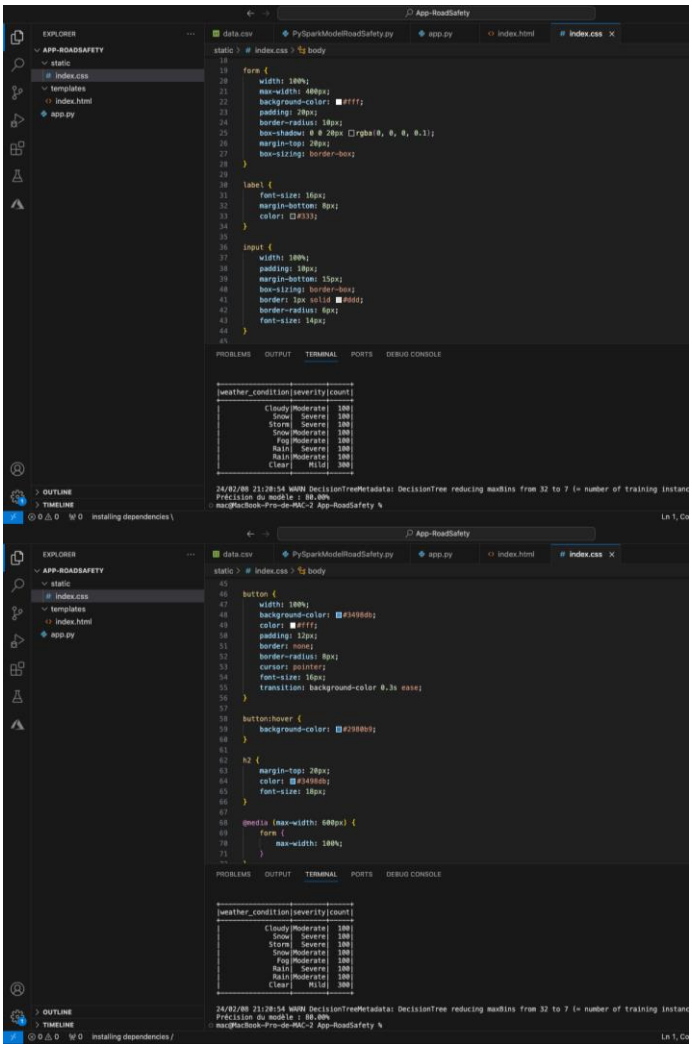
- HTML code that create our interface for predicting road safety based on input parameters such as latitude, longitude, weather condition, and time of



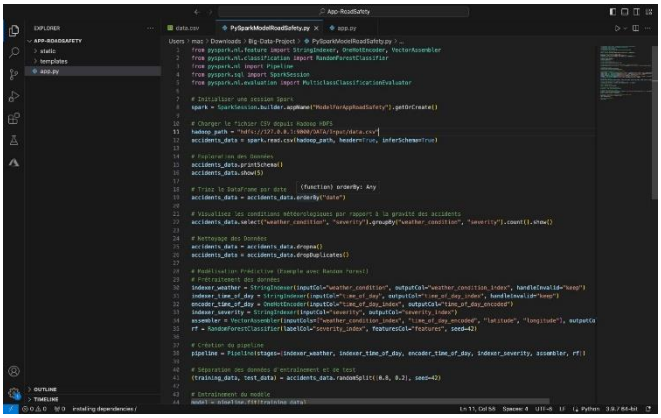
day, presenting the results in a visually appealing popup for enhanced user experience.



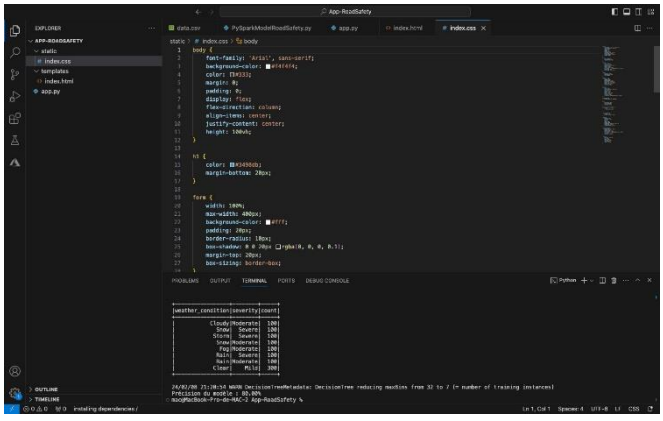


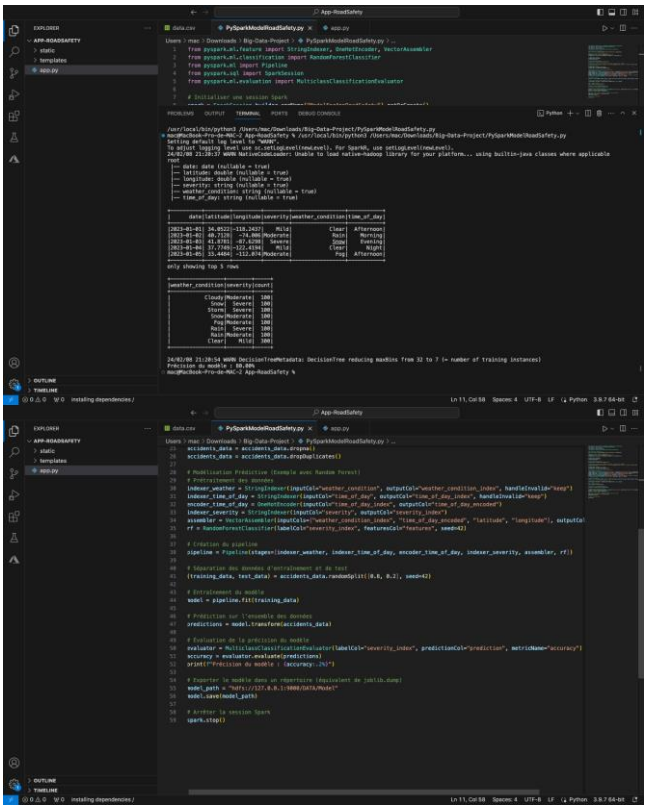


- Our Python script demonstrates the creation of a predictive model for road safetyusing PySpark. Key steps include data exploration, cleaning, and the construction of a Random Forest classification model. The model is trained, evaluated for accuracy, and saved for future use. The script showcases the power of PySpark for handling large-scale datasets and building machine learning pipelines in a distributed computing environment :

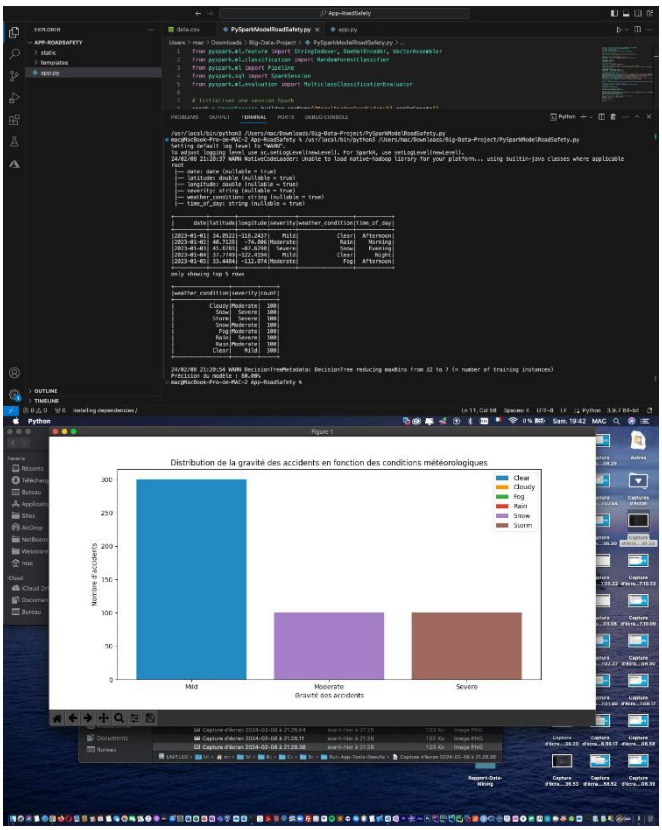


- CSS implementation for the index.html:

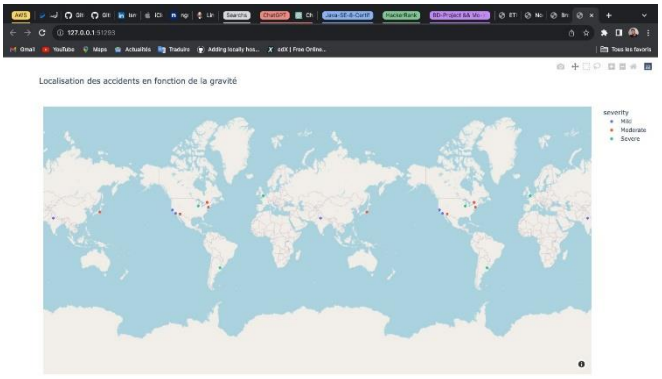


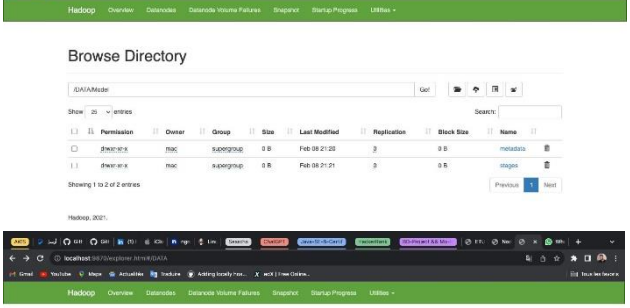


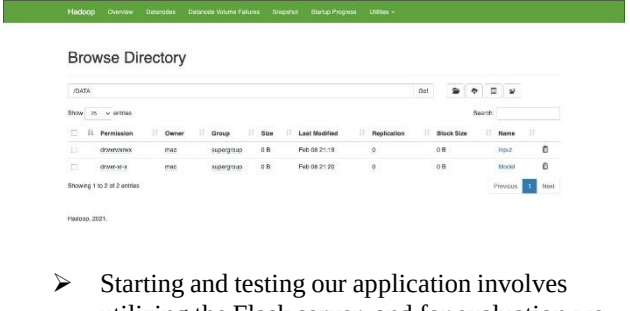
- The graph represents the distribution of accident severity based on meteorological conditions.



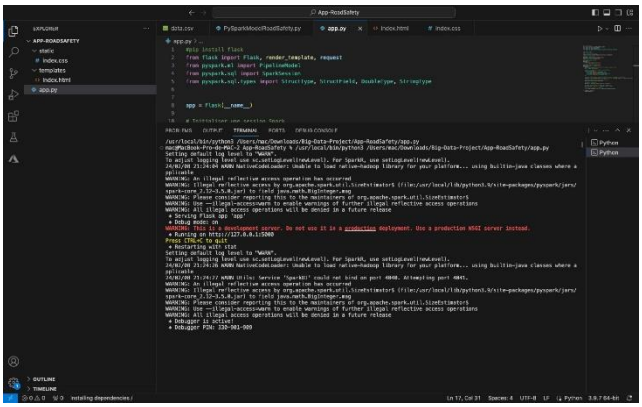
- The Map represents the location of accidents depending on the severity .

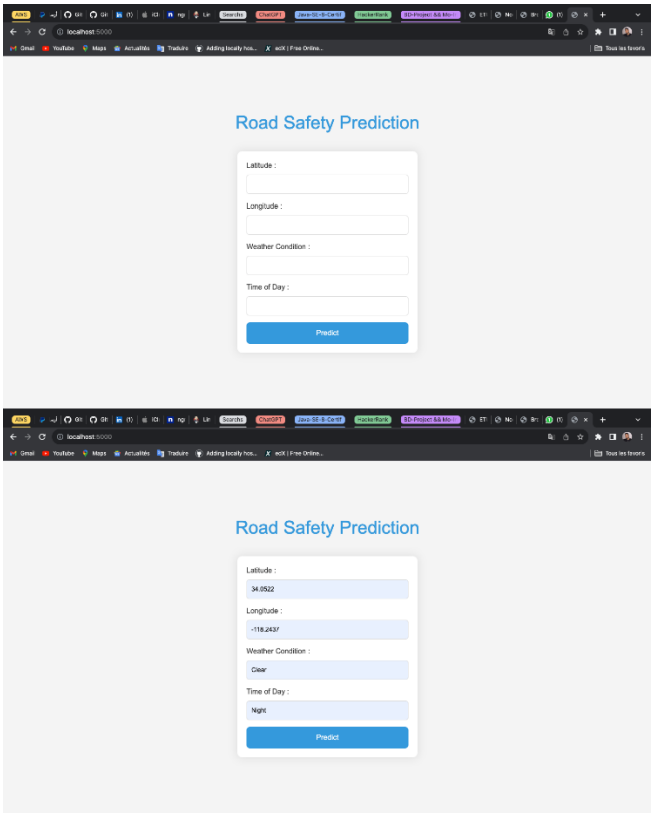




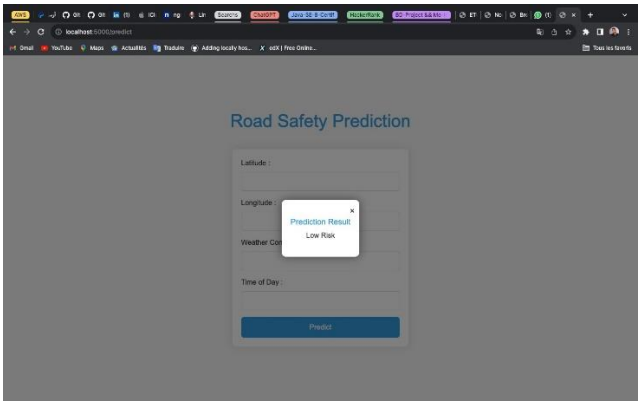


- Starting and testing our application involves utilizing the Flask server, and for evaluation,we provide two examples of coordinates for a road along with their corresponding prediction results.

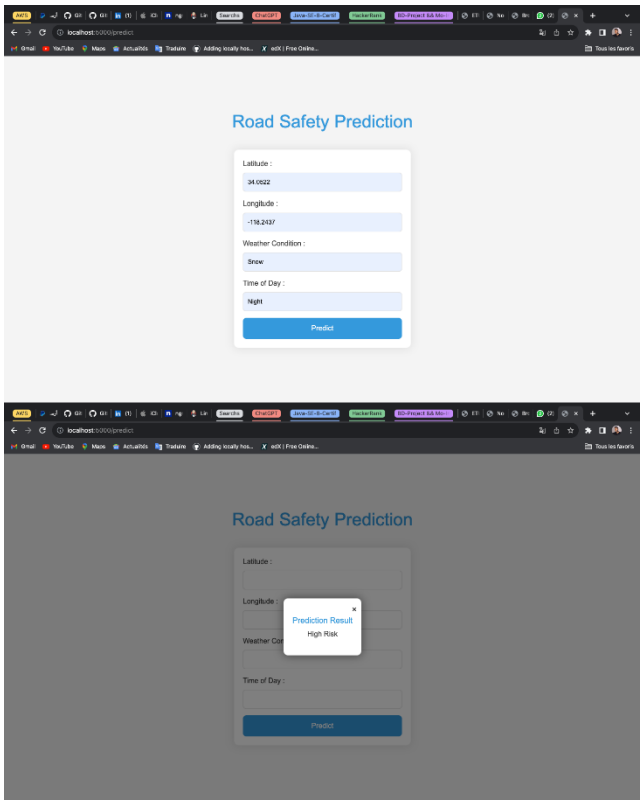




You enter the longitude, latitude, weather condition and the time of the day.



Once you enter the details it gives you a warning if the location is low risk or high risk in terms on accidents



V. CONCLUSION

In conclusion, this project represents a comprehensive and innovative approach to urban transportation analysis by leveraging the capabilities of Big Data analytics and Machine Learning methodologies, particularly utilizing SparkMLib. The integration of diverse datasets, including geographical information, weather conditions, and temporal factors, contributes to a holistic understanding of road traffic dynamics.

The project's primary objective is to assess and mitigate risks associated with urban transportation by developing predictive models capable of distinguishing between low and high-risk road segments. The spatial analysis of traffic patterns, facilitated by geographical data, aids in identifying congestion-prone areas and potential hazards. Weather conditions, serving as critical determinants, provide insights into the impact of meteorological phenomena on road safety. Additionally, considering temporal factors such as time of day and seasonal variations unveils temporal trends influencing traffic dynamics and risk levels.

The utilization of advanced Machine Learning algorithms within SparkMLib is a key aspect of this project, enabling the extraction of meaningful patterns from large-scale traffic data. The resulting predictive models are expected to empower stakeholders to anticipate and proactively mitigate potential risks. This proactive approach is aimed at optimizing traffic management strategies and resource allocation for safer and more efficient transportation systems in urban environments.

In essence, this project showcases the transformative potential of harnessing Big Data and Machine Learning to enhance road traffic analysis. By addressing the complexities of urban transportation and providing actionable insights, the project contributes to the development of smarter, safer, and more efficient urban transportation systems. The incorporation of keywords such as big data, traffic, forecast, longitude, and latitude underscores the multidimensional nature of the project, emphasizing its focus on utilizing diverse datasets and advanced analytics techniques for a holistic understanding of urban traffic dynamics.