Phase-2 Submission Template

Student Name: SANJANA C

Register Number:412723104088

Institution: TAGORE ENGINEERING

COLLEGE

Department: CSE

Date of Submission: 08.02.2025

Github Repository Link: [Update the project source code to your Github Repository]

1. Problem Statement.

_ The objective of this project is to build a machine learning model that can predict the severity level of road accidents based on various environmental, vehicular, and driver-related factors. Accident severity can typically be categorized into different levels such as minor, moderate, and severe. This classification problem aims to help traffic authorities and public safety departments anticipate the seriousness of potential accidents and take appropriate preventive measures, such as improving infrastructure,regulating traffic during certainconditions, or deploying emergency services efficiently.

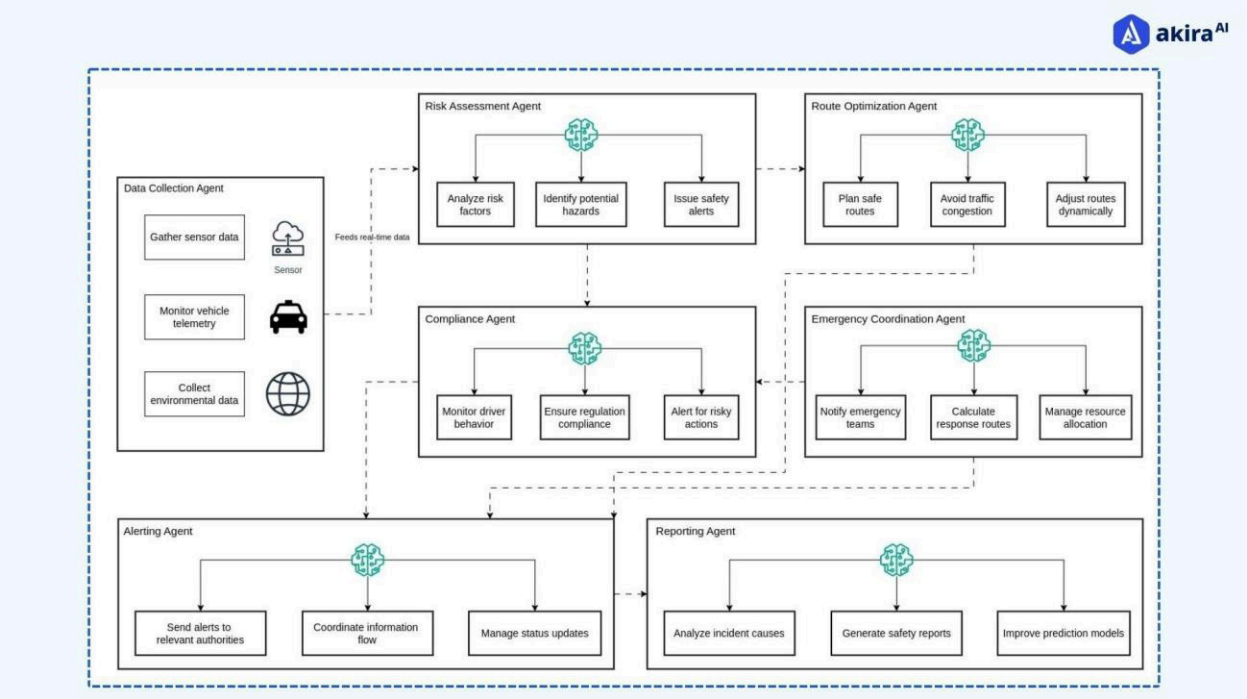
1. Project Objectives

- *To analyze the dataset and identify key features that influence accident severity.*
- *To apply appropriate preprocessing techniques to prepare the data for modeling.*
- *To implement and evaluate a Gaussian Naive Bayes classification model for predicting accident severity.*

- *To use various performance metrics like accuracy, precision, recall, and F1-score to assess the model.*
- *To visualize the outcomes and extract interpretable insights that could be useful in real-world applications.*

- *To establish a baseline performance which can be improved using advanced models in future phases.*

2. Flowchart of the Project Workflow



Flowchart Description:

The flowchart illustrates the step-by-step workflow of the machine learning project for predicting accident severity. It begins with loading the dataset, followed by data preprocessing which includes handling missing values, encoding categorical variables, and scaling numerical features. Next is exploratory data analysis to understand patterns in the data. The dataset is then split into training and testing sets. A Gaussian Naive Bayes classifier is trained on the training data, predictions are made on the test data, and the model is evaluated using performance metrics like accuracy and F1-score. Finally, results are visualized for better interpretation and insights.

3. Data Description

- *Target variable (if supervised learning).]*
- *Dataset Name:AccidentsBig.csv.zip*
- *Format: Structured tabular data in CSV format*
- *Total Records: [e.g., 20,000]*
- *Total Features: [e.g., 12]*
- *Data Type: Mix of categorical and numerical variables*
- *Nature: Static (not updating in real-time)*
- *Target Variable: Accident_Severity (multi-class: e.g., Low, Medium, High)*
- *The dataset includes information about weather conditions, road types, traffic density, driver behavior, and more*

4. Data Preprocessing

- *Numerical Features: Scaled using StandardScaler.*
- *categorical Features: Encoded using OneHotEncoder.*
- *Missing Values: Imputed or removed based on context.*
- *ColumnTransformer: Used for applying different transformations in*

a clean and organized manner.

5. Exploratory Data Analysis (EDA)

- *Univariate Analysis:*

Count plots and histograms explored the distribution of features like Road_Type,Accident_Severity, Time_of_Day

- *Bivariate/Multivariate Analysis:*

Bar plots and comparisons showed that severity is influenced by Driver_Alcohol,Weather, and Traffic_Density.

Key Insights:

Higher severity accidents occur more frequently during nighttime or poor weather.

Alcohol and inexperience are major contributing factors.

Intersections and highways tend to have more severe accidents.

6. Feature Engineering

No new features were added. Existing features were grouped as numerical and categorical for streamlined preprocessing. Future work may include travel duration, historical trends, or regional indicators.

7. Model Building

- *Algorithm: Gaussian Naive Bayes*
- *Justification: Simple, efficient, suitable for categorical data*
- *Data Split: 80-20 with random_state=42*
- *Evaluation Metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix*

8. Visualization of Results & Model Insights

- *Confusion matrix, ROC curve, feature importance plot, residual plots, etc.*

- *Used a seaborn confusion matrix heatmap to visualize prediction performance.*
- *Identified that the model performs better on majority classes but underperforms on rare severity cases.*
- *Observation: Suggests future use of advanced models like Random Forest or XGBoost for better handling of class imbalance.*

9. Tools and Technologies Used

- *Language: Python 3.x*
- *Environment: Jupyter Notebook / Google Colab*
- *Libraries: pandas, numpy, matplotlib, seaborn, sklearn*
- *Algorithm: GaussianNB from sklearn.naive_bayes*

10. Team Members and Contributions

[List names and responsibilities.

- *Clearly mention who worked on:*

Santhiya R: Data cleaning, feature engineering, EDA
Sanjana C: Model development, preprocessing, evaluation
Varshniarthi N: Documentation, result interpretation, visualization