Questions arise before model creation and evaluation

Understanding the Data

1. What are the key attributes (features) in the dataset?
2. What is the target variable? (e.g., accident severity, likelihood of an accident, etc.)
3. Are there any missing values, duplicates, or inconsistencies in the data?
4. What types of variables are present (categorical, numerical, etc.)?
5. Do we need to perform data cleaning?

Univariate Analysis

1. Are there any outliers in the dataset?
2. What are the summary statistics (mean, median, mode, standard deviation, etc.)?
3. How do categorical variables distribute (e.g., count of accidents per weather condition, road type, etc.)?

Bivariate Analysis

1. How do two variables relate to each other? (e.g., accident severity vs. weather condition)
2. Are there any strong correlations between features?
3. How do different categories affect the target variable?
4. Do we need visualizations like scatter plots?

Feature Selection

1. Which features are the most relevant for predicting accidents?
2. Should we use statistical methods like ANOVA, Chi-square?
3. Do we need dimensionality reduction techniques like PCA?

Model Creation

1. What machine learning models are best suited for this problem? (e.g., logistic regression, decision trees, random forests, neural networks)
2. Should we use classification or regression models?
3. What performance metrics should be used? (e.g., accuracy, precision, recall, F1-score)
4. Do we need hyperparameter tuning (e.g., GridSearchCV, RandomizedSearchCV)?
5. Should we use techniques like cross-validation to improve model performance?

Final Model Preparation & Evaluation

1. Which model performed the best, and why?
2. How does the final model perform on a test dataset?
3. How can the model be deployed and used in real-world scenarios?