Dataset Analysis Report - Task 1

Titanic Dataset – Analysis

The Titanic dataset contains information about 891 passengers with 12 features describing personal, socio-economic, and travel-related details. The objective of this analysis is to understand the structure of the dataset, identify data types, and evaluate its suitability for machine learning.

The dataset includes both numerical and categorical variables. Numerical features such as Age, Fare, SibSp, and Parch represent measurable quantities. Categorical features include Name, Sex, Ticket, Cabin, and Embarked, which describe qualitative attributes. The Pclass feature is ordinal in nature as it represents ordered passenger classes (1st, 2nd, and 3rd). The Survived column is binary and serves as the target variable for prediction.

Using df.info(), it was observed that the Age column contains missing values, and the Cabin column has a significant number of missing entries. The Embarked column has a few missing values as well. These missing values indicate data quality issues that must be handled before applying machine learning algorithms.

The df.describe() output provides statistical summaries such as mean, standard deviation, minimum, and maximum values for numerical features, helping to understand data distribution and detect potential anomalies.

The dataset size is moderate and suitable for machine learning classification tasks. However, the target variable Survived shows class imbalance, with more passengers not surviving than surviving. This imbalance should be considered during model training.

Overall, the Titanic dataset is appropriate for machine learning after preprocessing steps such as handling missing values, encoding categorical variables, and addressing class imbalance.

Students Performance Dataset – Analysis

The Students Performance dataset contains records of students with features describing demographic information, parental background, and academic performance. The objective of this analysis is to understand the structure of the dataset, identify data types, and evaluate its suitability for machine learning applications.

The dataset includes both numerical and categorical variables. Numerical features such as math score, reading score, and writing score represent measurable academic performance. Categorical features include gender, race/ethnicity, parental level of education, lunch, and test preparation course, which describe qualitative attributes related to the students' background.

Using students_df.info(), it was observed that the dataset does not contain significant missing values, indicating good data quality. This reduces the need for extensive data cleaning and makes the dataset easier to work with during preprocessing.

The students_df.describe() output provides statistical summaries such as mean, standard deviation, minimum, and maximum values for the numerical score features, helping to understand score distribution and variation among students.

The dataset size is moderate and suitable for machine learning tasks such as regression and classification. The target variable can be chosen based on the objective, such as predicting math score or overall performance. Overall, the Students Performance dataset is well-suited for machine learning after preprocessing steps such as encoding categorical variables and feature scaling.