Dataset Understanding & ML Readiness Report

Titanic Dataset

The Titanic dataset consists of **891 passenger records** with **12 attributes** describing demographic and travel-related information. The **target variable is** Survived, which is binary in nature (0 = did not survive, 1 = survived). This makes the dataset suitable for **binary classification problems**.

Numerical features in the dataset include **Age, Fare, SibSp, and Parch**, which represent continuous or count-based values. **Sex and Embarked** are categorical features, while **Pclass** is an ordinal feature since it represents passenger class with an inherent order (1st, 2nd, 3rd class).

The dataset contains **missing values** in several columns. The Age column has moderate missing values, while Cabin has a very high proportion of null values, making it less useful without significant preprocessing. Embarked contains a small number of missing entries. Additionally, there is a **class imbalance** in the target variable, as the number of passengers who did not survive is higher than those who survived.

Before applying machine learning models, the dataset requires **data preprocessing**, including handling missing values, encoding categorical variables, dealing with class imbalance, and possibly scaling numerical features. Despite these challenges, the dataset is widely used for introductory machine learning tasks and is suitable for supervised learning after proper preparation.

Students Performance Dataset

The Students Performance dataset contains **1000 student records** with **8 attributes** related to academic performance and socio-educational background. Unlike the Titanic dataset, this dataset **does not contain any missing values**, making it relatively clean and easier to work with.

The features **math score, reading score, and writing score** are numerical and can be used as target variables for **regression tasks** or converted into categories for **classification problems**. Other features such as **gender, race/ethnicity, lunch type, parental level of education, and test preparation course** are categorical or binary in nature.

The dataset is fairly **well-balanced** and requires minimal preprocessing, mainly limited to encoding categorical variables. Due to its clean structure and meaningful features, it is highly suitable for machine learning models aimed at predicting academic performance.

Conclusion

Both datasets are appropriate for machine learning applications. The **Titanic dataset** requires substantial preprocessing due to missing values and class imbalance, making it useful for learning data cleaning and feature engineering techniques. In contrast, the **Students Performance dataset** is comparatively clean and ML-ready, making it ideal for building and evaluating models with minimal preprocessing.