Dataset Analysis Report – Task 1 (AI & ML Internship)

1. Dataset Overview

The purpose of this research is to explain the structure, data types, and suitability of machine learning for the Titanic dataset. The Titanic dataset contains data on individual passengers, such as demographic variables, class of travel, price, and survival. The primary goal of this research is to determine the readiness of the data before applying machine learning algorithms.

2. Data Structure and Types

The dataset includes numerical variables (e.g., Age, Fare), categorical variables (e.g., Sex, Embarked), an ordinal variable (Pclass, which represents the hierarchy of passenger classes), and a binary target variable (Survived). The diversity of variable types makes the dataset suitable for a supervised learning problem.

3. Missing Values and Data Quality

Preliminary analysis with df.info() revealed missing values in the Age and Cabin columns. Missing values in the data can affect the accuracy of the model and require appropriate treatment in the future. In addition, the dataset has a problem of class imbalance with respect to the target variable.

4. Statistical Summary

The descriptive statistics derived from df.describe() provide information on the distribution of the numerical attributes. There are some extreme values in Fare, indicating the existence of outliers, but the Age attribute shows a fairly good distribution of values among the passengers.

5. Target Variable and ML Suitability

The target variable, Survived, is a binary variable, making it suitable for classification problems. The number of samples in the dataset is sufficient for basic machine learning experiments. However, data preprocessing,

such as dealing with missing data and categorical variables, is necessary before building any models.

6. Key Observations - Presence of missing values - Class imbalance in survival outcomes - A mix of numeric and categorical variables - Overall appropriateness for introductory learning and exploratory machine learning tasks