

EDA AND MACHINE LEARNING ASSIGNMENT

EDA Impact on Model Accuracy – A Case Study with Titanic Dataset

Learning Objectives:

- Understand the importance of thorough EDA.
- Perform feature engineering based on EDA insights.
- Train and evaluate models with and without EDA-driven preprocessing.
- Compare performance metrics and draw conclusions.
- Try different models to see which one gives better evaluation results.
- Explain why a certain model maybe better

Dataset:

Use the **Titanic - Machine Learning from Disaster** dataset from Kaggle:

<https://www.kaggle.com/c/titanic>

Task: Use machine learning to create a model that **predicts which passengers survived the Titanic shipwreck.**

Instructions:

1. **Download** the Titanic dataset (train.csv) and upload it to your Colab/Notebook.
2. **Start with baseline model:** Train a logistic regression or decision tree model on raw data with minimal preprocessing.
3. **Then perform detailed EDA** and preprocessing.
4. **Re-train** the model on the cleaned data and compare results.
5. Try different models and compare results and analyse
6. Compare the time it takes to train different models and compare results/present the time taken for each model in tabular format in the code itself (using dataframe)

A starter notebook is provided with the assignment, in the same folder

Submission:

A python notebook that contains the complete code and analysis of how EDA is impacting model evaluation results, and how choosing different models impacts evaluation results.