

# EDA + Missing values and Outliers - Detection and Treatment + Model Building and MORE 😊

Dataset - [adult.csv](#)

Dataset Description - Google it 🤖

Perform below mentioned tasks:

**Step - 1** - Introduction -> Give a detailed data description and objective

**Step - 2** - Import the data and perform basic pandas operations

**Step - 3** - Univariate Analysis -> PDF, Histograms, Boxplots, Countplots, etc..

- Understand the probability and frequency distribution of each numerical column
- Understand the frequency distribution of each categorical Variable/Column
- Mention **observations** after each plot

**Step - 4** - Bivariate Analysis

- Discover the relationships between numerical columns using Scatter plots, hexbin plots, pair plots, etc..
- Identify the patterns between categorical and numerical columns using swarmplot, boxplot, barplot, etc..
- Mention **observations** after each plot.

**Step - 5** - In the above steps you might have encountered many missing values and outliers

- **Find and treat the outliers and missing values** in each column 🤖
- Read this Kaggle Notebook and understand various ways to detect and handle outliers. Try to implement the same. [Outlier!!! The Silent Killer](#)

**Step - 6** - Apply appropriate hypothesis tests to verify the below mentioned questions

- Is there a relationship between occupation and gender? (i.e. does the preference of occupation depend on the gender)
- Is there a relationship between gender and income?
- You are free to explore other tests also.

**Conclusion of EDA**

**NOTE:** Mention **observations and insights** clearly. 🤖

**Step - 7** - Split the data into train and test. After which you need to perform feature transformation:

- For Numerical Features -> Do Column Standardization
- For Categorical -> if more than 2 categories, use dummy variables. Otherwise convert the feature to Binary.
- You are free to explore other feature transformations.

**Step - 8** - Build various Machine Learning models considering 'income' as target variable. Also make sure to perform Hyperparameter tuning to avoid Overfitting of models.

**Step - 9** - Create a table to compare the performance of each of the ML Model

**Step - 10** - Read the research papers mentioned below & rethink the missing value treatment and feature engineering aspect. Try to document the things you are implementing from the research paper.

**Step - 11 - Research Paper Reading -**

[research\\_paper.pdf](#)

(Read this entire paper and try to perform some experiments and try to match the results)

[research\\_paper\\_2.pdf](#)

(From above research paper implements Extra Tree Classifier, Handling missing values, categorical variable encoding, gradient boosting for classification)

**Resources -**

[Basics of Missing Value Detection and Treatment](#)

[Outlier!!! The Silent Killer](#)