ELECTRICAL TEAM TRAINING

TASK 10

Individual Tasks
Task 10.1: Data Analysis of the Titanic

The **Autobots** have intercepted the Titanic's passenger data to uncover survival patterns. Optimus Prime wants you to lead the **data analysis mission** to reveal hidden truths that may help humanity.

Requirements
You are required to write both your code and your analysis report inside Google Colab. After each data analysis step, include a conclusion or interpretation explaining your findings in your own words. using ai tools to generate your report is highly discouraged and if caught ,you will receive a penalty

- **Dataset Overview**
  - Load the Titanic dataset, link for the dataset
  - Display the **first 10 rows**.

- Summarize the dataset using info and describe.
- Report the number of **missing values** in each column.
- **Demographics Analysis**
  - Find the **average age** of passengers.
  - Find the **age distribution** (histogram).
  - Find the youngest and oldest passenger.
- **Survival Analysis**
  - Calculate the **overall survival rate**.
  - Compare survival rate by:
    - **Gender**
    - **Passenger class (Pclass)**
    - **Embarkation port (Embarked)**
  - Show these comparisons using **bar plots or pie charts**.
- **Family and Fare Analysis**
  - Analyze survival rate for passengers traveling **alone vs with family**
  - Explore how **Fare** relates to survival (e.g., average fare of survivors vs non-survivors).
  - Plot the **fare distribution** (histogram/boxplot).

- **Age Group Analysis**
  - Create age groups:
    - Children (0–12)
    - Teenagers (13–19)
    - Adults (20–59)
    - Seniors (60+)
  - Compare survival rates across these groups.
  - Plot with a **bar chart**.
- **Conclusions**
  - Write a **short summary** of key insights (e.g., "Women and

children had higher survival rates of about xx%).

⚡ *Remember: Autobots value clarity and visuals, your analysis should include tables, plots, and concise insights.*

Task 10.2 Heart Disease classification

You will explore the Heart Disease dataset and evaluate how well different machine learning classifiers can predict the presence of heart disease.

1. **Dataset Preparation**
   - Load the Heart Disease dataset , link for the dataset.
   - Perform **basic data cleaning** (check for missing values, encode categorical variables).
   - Split the dataset into **training (80%)** and **testing (20%)** sets.
   - Standardize/normalize features (especially for distance-based models like KNN).
2. **Model Training**
   Train both the logistic regression classifier and K-Nearest Neighbors (KNN) classifier
3. **Model Evaluation**
   - For each classifier, compute and report the following metrics:
     - **Accuracy**
     - **Precision, Recall, F1-score**
     - **Confusion Matrix**
   - **Resource for Model evaluation: Confusion matrix**
   - **BONUS:** Try to improve the logistic regression classifier using gradient descent, Link for ai's gradient descent session
4. **Visualization**
   - Plot confusion matrices as heatmaps for each classifier.
   - Create a bar chart comparing classifier accuracies.

5. **Analysis & Conclusion**
   ○ After each result, write a **short interpretation** of how the classifier performed.
   ○ In your final conclusion, answer:
      ■ Which classifier performed best overall?
      ■ Did different models excel in different metrics?

Submission

- You will submit your reports as **google colab** links for Task 10.1 and 10.2
- The Task's deadline is at monday, 1/9 11:59 PM.
- Submission link:https://forms.gle/GDYUQ4eUkfLLQjzs6