smartknower
STAY AHEAD OF THE CURVE

Major Project

- **Project Name:**

Machine Learning June Major Project

- **Project Description:**

Problem statement: Create a classification model to predict whether a person makes over $50k a year

Context: This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

Dataset :

https://drive.google.com/file/d/1E_IaMMGqP8qDA3O9VW1rzhrXeaq2dY1S/view?usp=sharing

Details of features:

The columns are described as follows:

1) Age
2) Workclass
3) Fnlwgt
4) Education
5) education_num
6) marital_status
7) occupation
8) relationship
9) race
10) sex
11) capital_gain
12) capital_loss
13) hours_per_week
14) native_country
15) income

Steps to consider:

1)Rename the columns.

2)Remove handle null values (if any).

3)Split data into training and test data.

4)Apply the following models on the training dataset and generate the predicted value for the test dataset

   a. Decision Tree
   b. Random Forest Classifier
   c. Logistic Regression
   d. KNN Classifier
   e. SVC Classifier (with linear kernel)

5)Predict the income for test data

6)Compute Confusion matrix and classification report for each of these models.

7)Validate the result for Precision, Recall, F1-score and Accuracy for each model based on values from confusion_matrix and classification_report

8)Generate the percentage of misclassification in each of these models.

9)Report the model with the best accuracy.