Mini Project - Machine Learning

Domain: Stroke Prediction

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status.

Following is a description of each variable in the Dataset.

Variable	Definition
Id :	unique identifier.
Gender :	"Male", "Female" or "Other".
Age :	age of the patient.
Hypertension :	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
Heart_disease :	0 if the patient doesn't have any heart. diseases, 1 if the patient has a heart disease.
Ever_married :	"No" or "Yes".
Work_type :	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed".
Residence_type :	"Rural" or "Urban".
Avg_glucose_level :	average glucose level in blood.

- **Bmi :** body mass index.
- **Smoking_status :** “formerly smoked”, “never smoked”, “smokes” or “Unknown”.
- **Stroke :** 1 if the patient had a stroke or 0 if not.

Deliverables :

Data Preprocessing :

This step performs all pre-processing steps such as data manipulation, data filling,converting categorical into numeric, and all processes.

Exploratory Data Analysis :

- The data preprocessing stage extracts useful information statistically.
- Such as check outliers, skewness, compare the features by graph and many more. Do all the required steps as well.
- The EDA process involves performing.

1. **Univariate Analysis :** In this part, first check target features and start univariate analysis.
2. **Bivariate analysis :** This analysis involves studying two variables and their relationship, recalling some of the hypotheses that we generated earlier.

3. **Removing Missing values if any / Outlier treatment :** After exploring all the variables in our data, we can now impute the missing values and treat the outliers because missing data and outliers can have an adverse effect on the model performance and accuracy.
4. **Machine Learning :** Apply Appropriate machine learning algorithm to Predict the probability of a candidate will work for the company. and also check if the model is to be underfitting or overfitting if it has then solves this by using cross-validation technique, or perform hyperparameter tuning to improve model performance.

Deadline to submit the project : **One Week.**

ALL THE BEST