CAPSTONE PROJECT - 1
PGA_09

PROJECT TITLE : Diabetes Prediction

Abstract :

Diabetes prediction is a critical task for early intervention and management of the disease. This study focuses on the application of classification models for predicting diabetes. A comprehensive dataset comprising various clinical and demographic features related to diabetic patients is utilized. Preprocessing techniques are applied to prepare the data, followed by the implementation of several classification algorithms for better understandings. Performance evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score. The results highlight the effectiveness of classification models in accurately predicting diabetes, facilitating timely intervention and personalized treatment plans. The findings of this study hold significant implications for healthcare professionals, enabling them to make informed decisions for diabetes prevention and management, thereby improving patient outcomes and quality of life.

Data-set Information :

The data-set consists of several predictor values and one target values. The target values is “diabetes” consists of 0 and 1. where 0 indicates no diabetes and 1 indicates patient has diabetes.

Column Name	Description
gender	Mentioned genders are male, female and others.
age	Age ranging from 2 to 80 and has average of 41.
hypertension	Hypertension has numerical values 1 and 0 to identify whether the patient has hypertension or not.
heart_disease	Heart_disease has numerical values 1 and 0 to identify whether the patient has hypertension or not.
smoking_history	Smoking_history consists of different categorical values to identify whether the patient is a smoker or not.
bmi	Bmi has various numerical values approximately ranging from 15 to 50
HbA1c_level	HbA1c_level i.e., Glycated Hemoglobin level is consists of numerical values approximately from 4 to 8.
blood_glucose_level	blood_glucose_level is a numerical value ranging from 80 to 220
diabetes	diabetes is the target variable consists of numerical values 1 and 0 which represent the patient has diabetes or not respectively.

Introduction :

Diabetes, also known as diabetes mellitus, is a chronic metabolic disorder characterized by high blood glucose (blood sugar) levels. It occurs when the body either doesn't produce enough insulin or cannot effectively use the insulin it produces. Insulin is a hormone produced by the pancreas that helps regulate blood sugar levels by facilitating the absorption of glucose into cells for energy.

The data-set consists of different predictor variable and one target variable. Total observation given in this data-set is 100000. In these data-set we have several duplicates and missing values.

#	Column	Non-Null Count	Dtype
0	gender	100000 non-null	object
1	age	100000 non-null	float64
2	hypertension	100000 non-null	int64
3	heart_disease	100000 non-null	int64
4	smoking_history	100000 non-null	object
5	bmi	100000 non-null	float64
6	HbA1c_level	100000 non-null	float64
7	blood_glucose_level	100000 non-null	int64
8	diabetes	100000 non-null	int64

The data-set consists of no missing values. The data-set has 2 object datatype “gender” and “smoking_history”, three float value “age”, “bmi” and “HbA1c_level”, and four integer value “hypertension”, “heart_disease”, “blood_glucose_level” and “diabetes”.

EDA and Pre processing :

Exploratory data analysis made on different variables in order to understand the data better.where we found that the blood glucose level, HbA1c value has more influence in the diabetic disorder than the other variable like smoking history, gender, bmi etc…

Model Building :

The model building we use in this case study are

1. **Decision Tree,**
2. **Random Forest,**
3. **Support Vector Machine,**
4. **Bagging method,**
5. **Hyperparameter tuning for Decision tree model.**

Results :

	Model	AUC Score	Precision Score	Recall Score	Accuracy Score	Kappa Score	f1-score
0	Decision Tree	0.876181	0.490295	0.836059	0.908007	0.569816	0.618109
1	Random Forest	0.971922	0.593750	0.836059	0.934464	0.658846	0.694373
2	Support Vector Machine	0.961826	0.566731	0.805296	0.927841	0.626200	0.665273
3	Bagging Model(ensembling)	0.963064	0.598521	0.819704	0.934984	0.656508	0.691865
4	D_tree with tuned Parameters	0.974755	0.660075	0.818925	0.946323	0.701538	0.730970

Conclusion :

After performing different types of models using the given data-set we can come into a conclusion that the Decision tree model after hyper parameter tuning has the higher accuracy when compared with other model. So this model can help the medical industries to predict who has prone to the diabetic disorder.

References :

The data set which I've taken here for diabetic prediction was fetched from www.kaggle.com

Some basic information about the diabetes was elaborated in my documentation with the help of https://chat.openai.com/