Capstone Project: Healthcare

Variables - Description

- Pregnancies - Number of times pregnant
- Glucose - Plasma glucose concentration in an oral glucose tolerance test
- BloodPressure - Diastolic blood pressure (mm Hg)
- SkinThickness - Triceps skinfold thickness (mm)
- Insulin - Two hour serum insulin
- BMI - Body Mass Index
- DiabetesPedigreeFunction - Diabetes pedigree function
- Age - Age in years
- Outcome - Class variable (either 0 or 1). 268 of 768 values are 1, and the others are 0

1. First I load necessary library that is numpy ,pandas,seaborn,matplotlib etc
2. I do some initial eda and found number of zeros in different columns
3. Replaced those zeros with null values
4. Then impute this null values with mean and median values
5. Then I compare outcomes and found there is imbalance
6. So I went for Synthetic minority oversampling techniques to balance the dataset
7. Also able to find correlation matrix heatmap
8. Then I went for modeling
9. Since this is a classification problem, I built all popular classification models for my training data and then compare performance of each model on test data accurately predict target variable (Outcome): These algorithims are mentioned below
10. Logistic Regression
11. Decision Tree
12. RandomForest Classifier
13. K-Nearest Neighbour (KNN)
14. Support Vector Machine (SVM)
15. Naive Bayes
16. Ensemble Learning -> Boosting -> Adaptive Boosting
17. Ensemble Learning -> Boosting -> Gradient Boosting (XGBClassifier)
18. I have also done hyperparametre tuning through Gridsearch CV
19. Also found out auc_score , ROC curve , Precision-Recall Curve ,accuracy ,f1 score ,average precision for all of the above algorithim
20. Then I compare all of these metrics and found out Random Forest gave best result for this dataset
21. Then I built my model with random forest algorithim

22. Then I also found out Confusion Matrix, TPR , FPR, Sensitivity , recall specificity etc
23. Then I went to made Dashboard with Tableau
24. I made a pie chart describing total number of diabetic and non diabetic patient
25. I also made scatter plot between different field by taking all the field through parameter
26. I also made Histogram for different field by tking all the field through parameter
27. Then I went for making of Bubble chart by creating age bins and different field as parameter
28. I also made a heatmap describing corealation between different fields
29. Finally I made dashboard by taking all of the above sheet
30. You can find it through my Tableau Public account, Pleas visit through below url
31. https://public.tableau.com/app/profile/tarun.kumar.mohapatra/viz/Healthcare_Dashboard_16701537352940/Dashboard1