DATA MINING & STATISTICAL LEARNING
ISYE 7406
COURSE PROJECT

FARHAD BATMANGHELICH

11/20/2021

OUTLINE

- The dataset that is used in this project is “Diabetes Prediction” and is obtained from the following URL:
  - https://data.world/informatics-edu/diabetes-prediction
- This is real patient data collected from several hundred rural African American patients and is originally from biostatistics program at Vanderbilt.
- The dataset contains 390 labeled datapoints where each datapoint is a vector of 14 features describing a given patient and the label is binary “Diabetes/No Diabetes”

source:

FEATURES

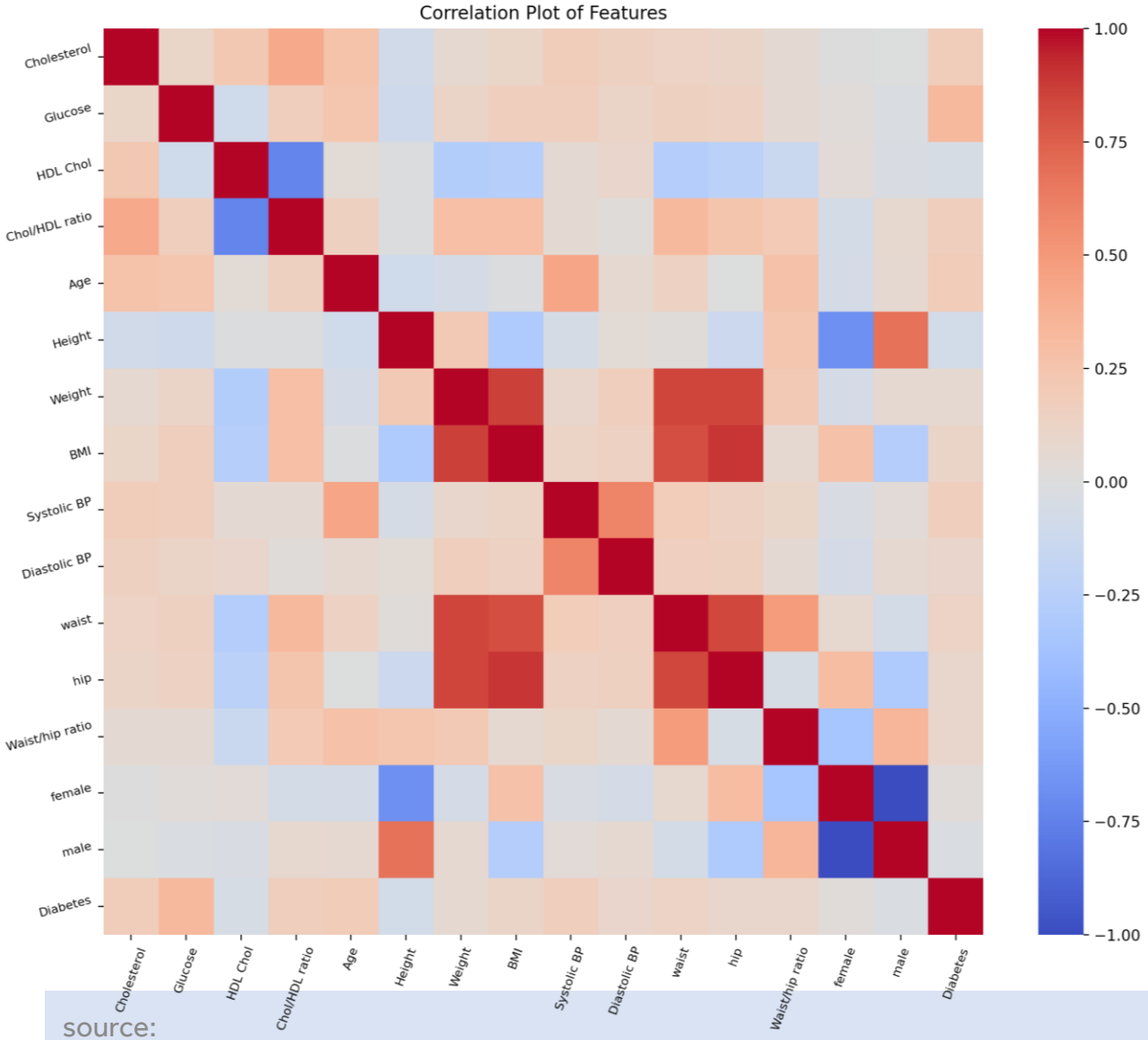
- The following table captures input features as well as the binary label

  (Diabetes/No Diabetes)

Column attribute	Description
Patient number	Identifies patients by number
Cholesterol	Total cholesterol
Glucose	Fasting blood sugar
HDL	HDL or good cholesterol
Chol/HDL	Ratio of total cholesterol to good cholesterol. Desirable result is < 5
Age	All adult African Americans
Gender	162 males, 228 females
Height	In inches
Weight	In pounds (lbs)
BMI	703 x weight (lbs)/ [height(inches]2
Systolic BP	The upper number of blood pressure
Diastolic BP	The lower number of blood pressure
Waist	Measured in inches
Hip	Measured in inches
Waist/hip	Ratio is possibly a stronger risk factor for heart disease than BMI
Diabetes	Yes (60), No (330)

EXPLORATORY DATA ANALYSIS

- Correlation plot

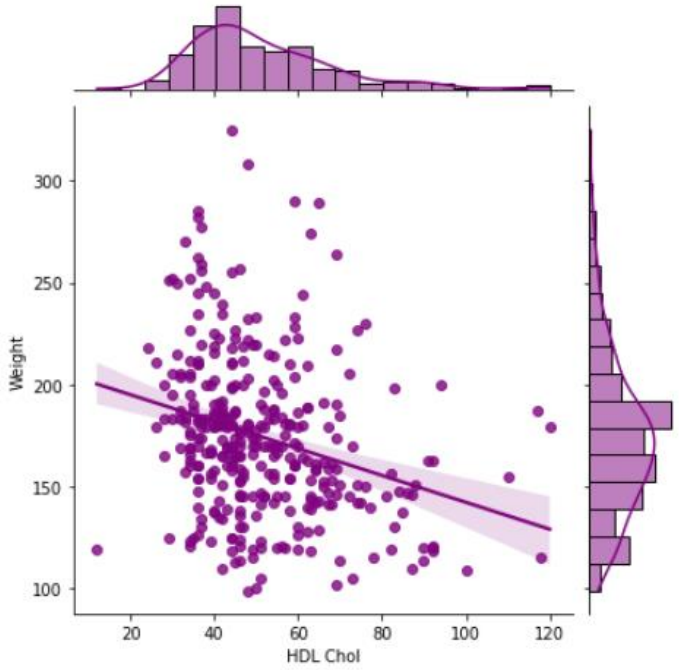


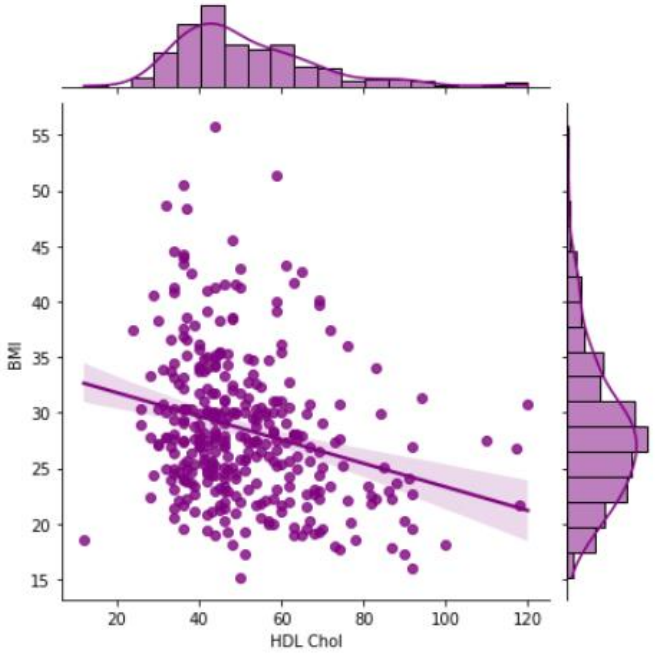
- HDL cholesterol has a negative correlation with body mass features such as Weight, BMI, waist etc. HDL cholesterol is the so called "good" cholesterol because it helps removing other forms of cholesterol, thus higher HDL is better.
- HDL cholesterol and Height have negative correlation with having diabetes
- Glucose, cholesterol, age, Systolic BP have positive correlation with having diabetes

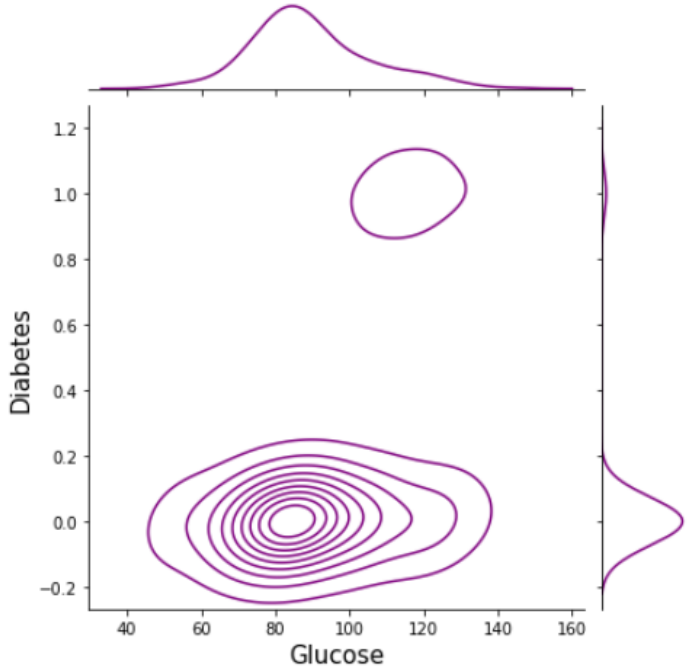
source:

EXPLORATORY DATA ANALYSIS

- HDL cholesterol has a negative correlation (linear correlation) with body mass features such as Weight and BMI
- There are two relatively distinct distributions at low and high values of Glucose, this shows that the Glucose feature in this dataset is promising in prediction of diabetes.



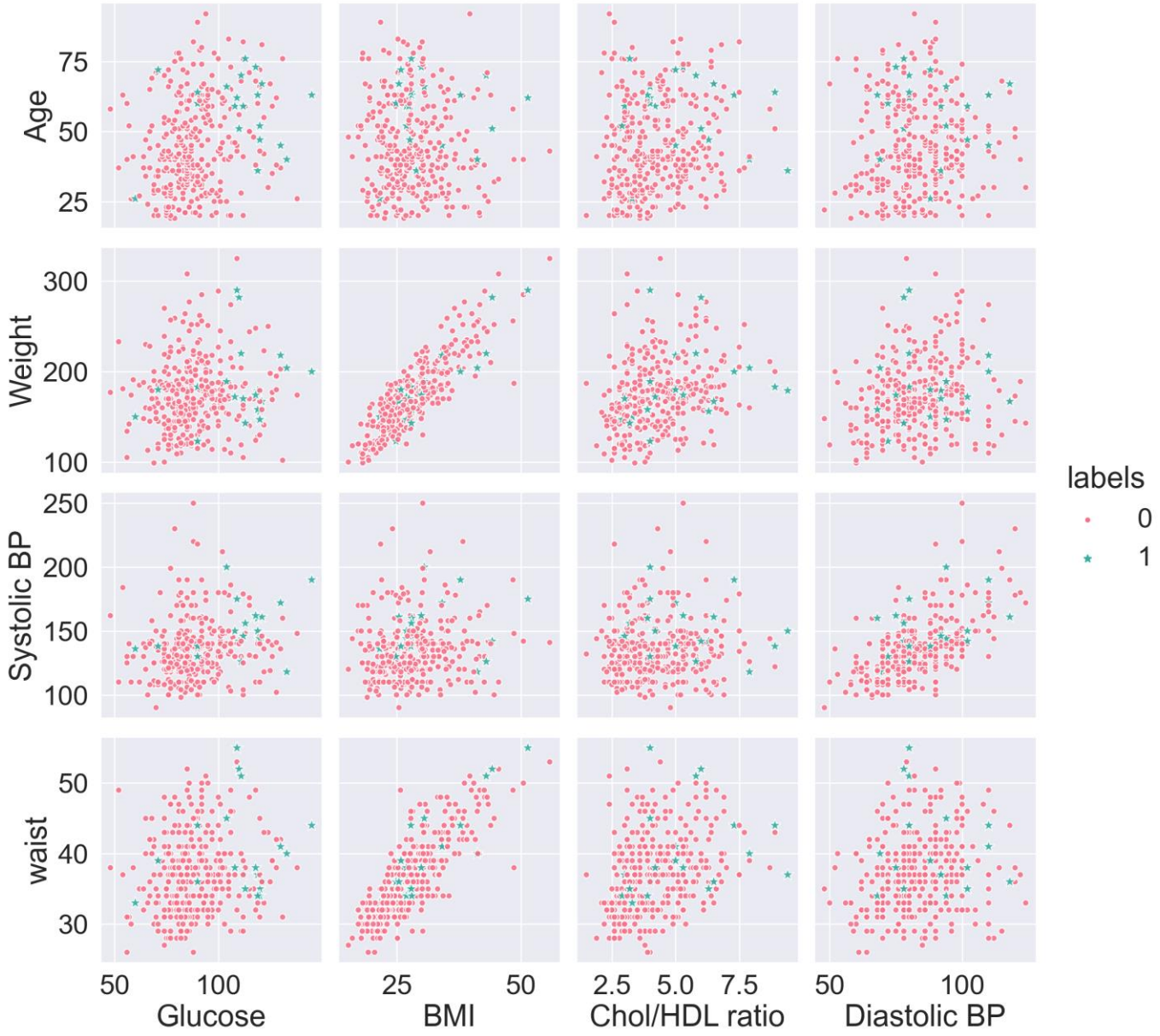




source:

MODELING

- Plotted few of the variables against each other with the label (0: no diabetes, 1: diabetes)
- Not a clear linear separation between the two classes
- Will plot PCAs to determine if there is a distinction there

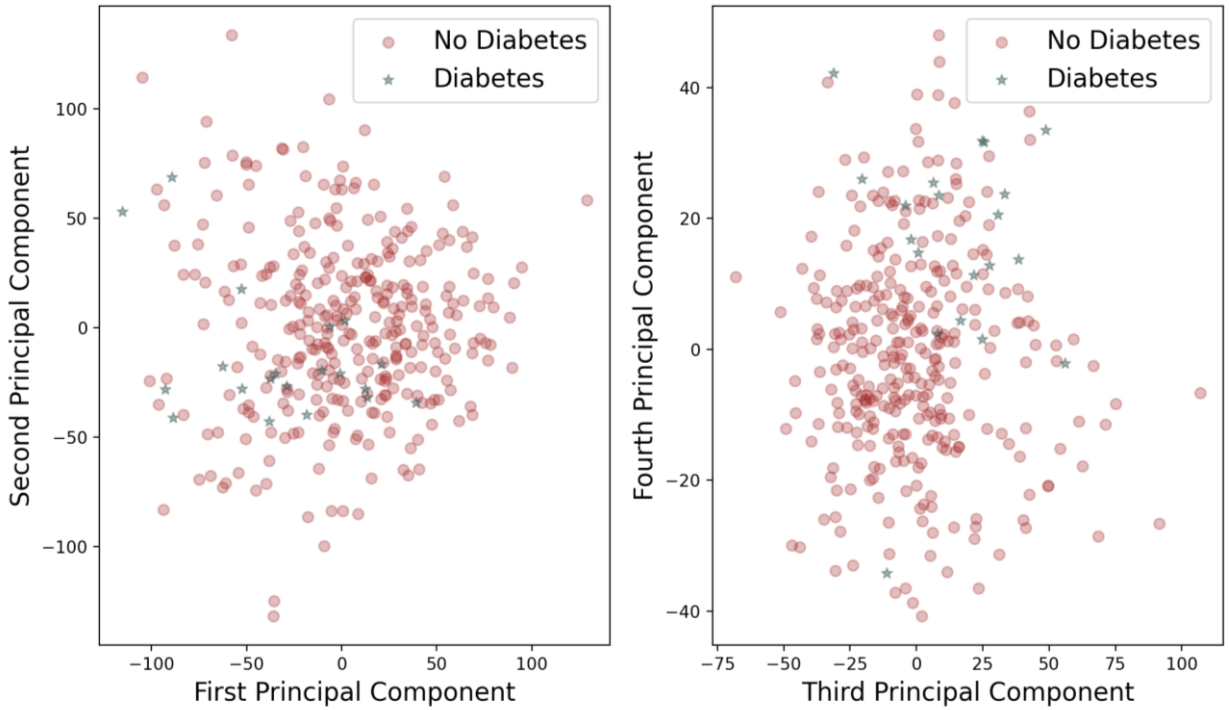


source:

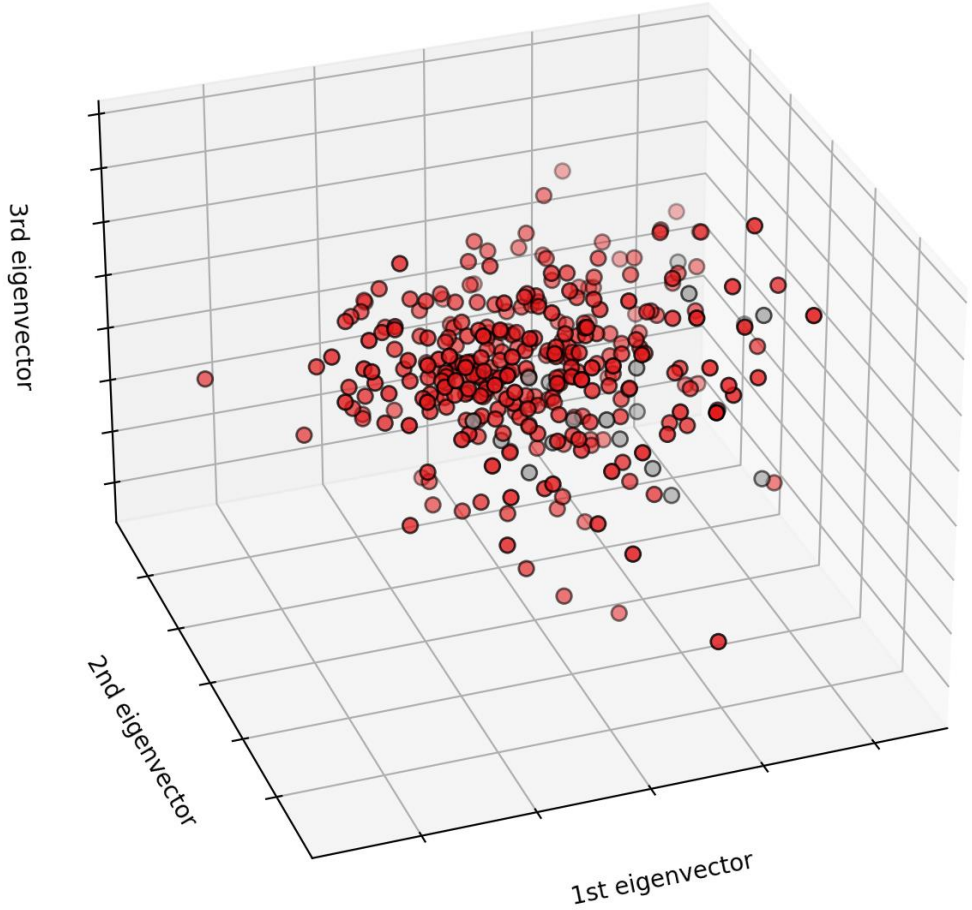
MODELING

- Seems that in the principal component space, there is still no linear boundary between classes

PCA 2D



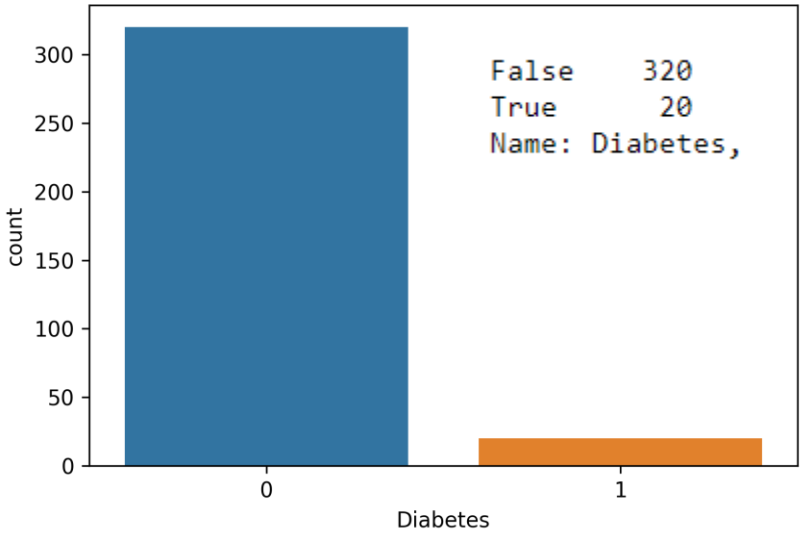
PCA 3D



source:

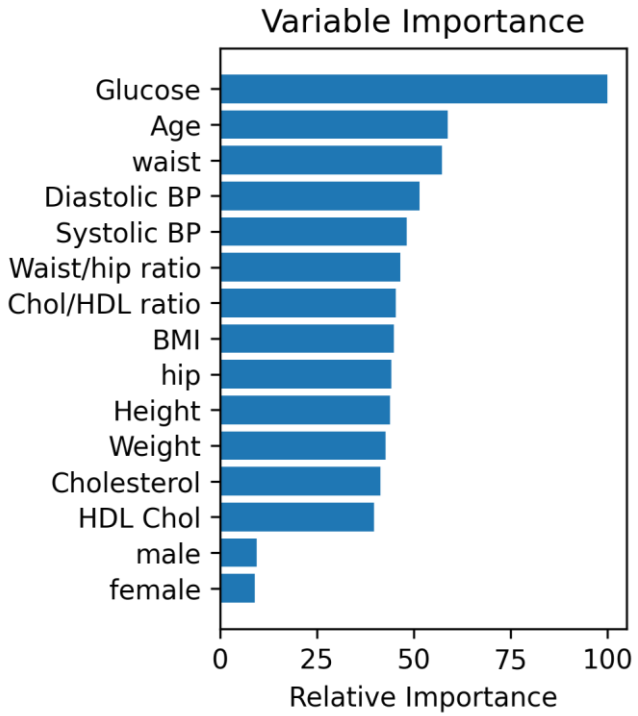
MODELING

- As can be seen from the previous results:
  - This data is highly imbalanced



  - The two classes are not linearly separable

- Seems that glucose, age and waist are important in predicting diabetes



source:

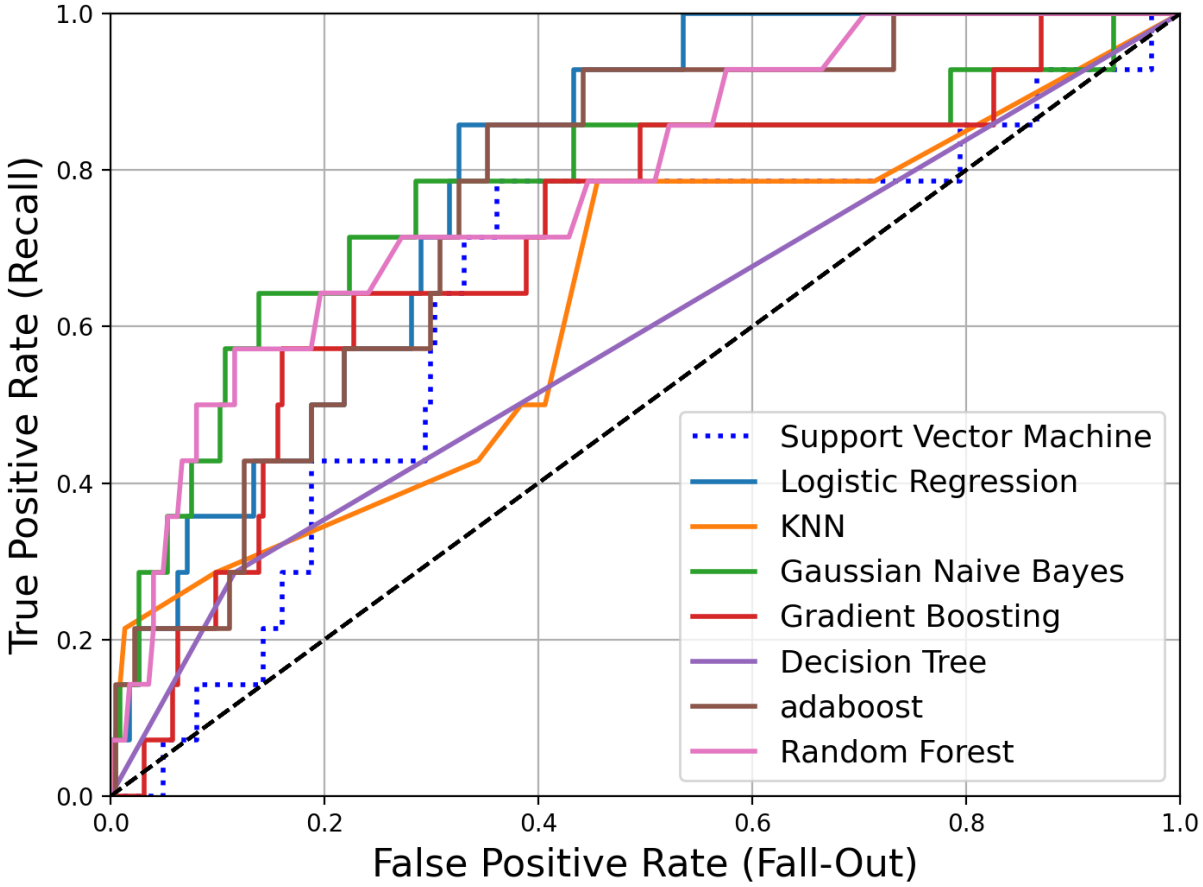
MODELING-dealing with imbalance in classes

- I tried “imblearn” library to deal with scarcity of class “1” i.e. Diabetes
- These methods are based on:
  - Under-sampling
  - Over-sampling
  - Combination of the above
- I used the **RepeatedEditedNearestNeighbours** class which is an under-sampling method on majority class i.e. No Diabetes:
  - Under-sample based on the repeated edited nearest neighbor method

https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RepeatedEditedNearestNeighbours.html

source:

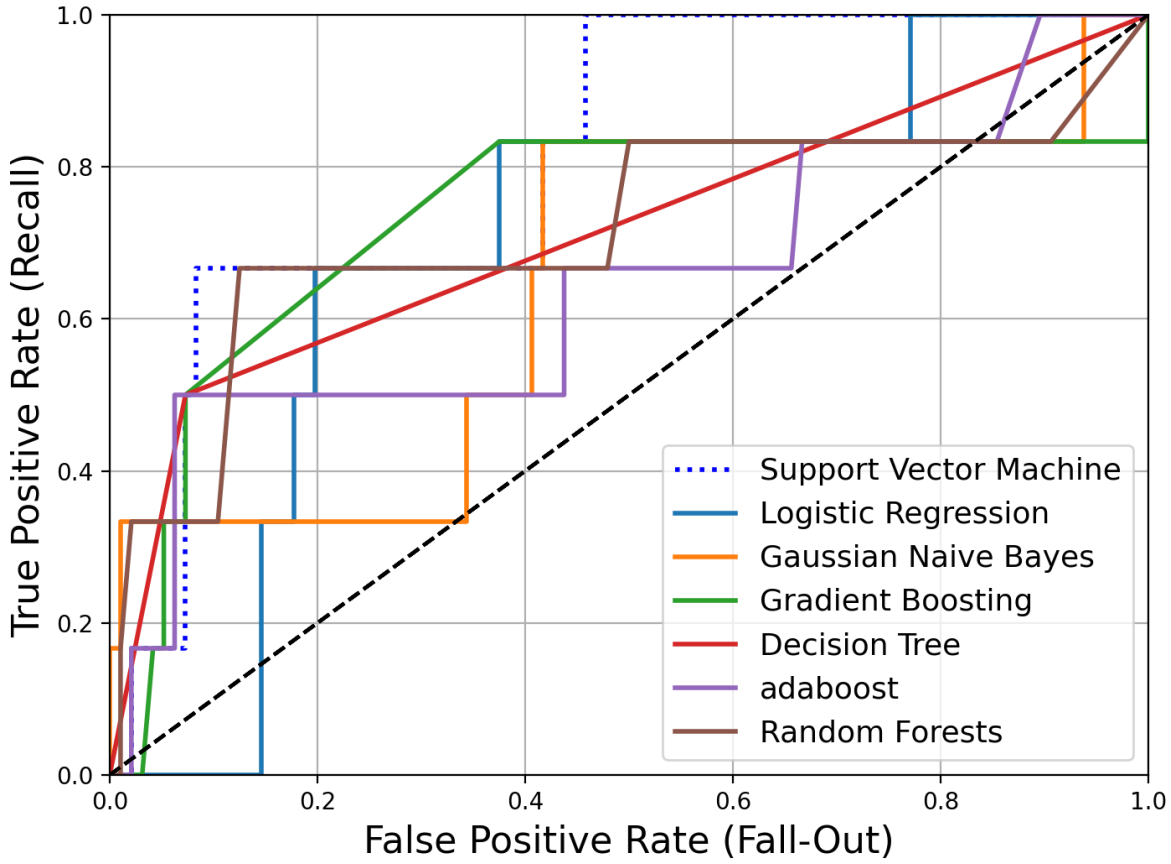
MODEL PERFORMANCE ON TRANING DATA



Model	Classification Report
Support Vector Machine	Confusion Matrix: [[224 0] [14 0]] precision recall f1-score support 0 0.94 1.00 0.97 224 1 0.00 0.00 0.00 14
Logistic Regression	Confusion Matrix: [[179 45] [2 12]] precision recall f1-score support 0 0.99 0.80 0.88 224 1 0.21 0.86 0.34 14
KNN	Confusion Matrix: [[202 22] [5 9]] precision recall f1-score support 0 0.98 0.90 0.94 224 1 0.29 0.64 0.40 14
Gaussian Naïve Bayes	Confusion Matrix: [[211 13] [6 8]] precision recall f1-score support 0 0.97 0.94 0.96 224 1 0.38 0.57 0.46 14
Gradient Boosting	Confusion Matrix: [[223 1] [0 14]] precision recall f1-score support 0 1.00 1.00 1.00 224 1 0.93 1.00 0.97 14
Decision Tree	Confusion Matrix: [[215 9] [0 14]] precision recall f1-score support 0 1.00 0.96 0.98 224 1 0.61 1.00 0.76 14
adaboost	Confusion Matrix: [[205 19] [0 14]] precision recall f1-score support 0 1.00 0.92 0.96 224 1 0.42 1.00 0.60 14
Random Forest	Confusion Matrix: [[204 20] [0 14]] precision recall f1-score support 0 1.00 0.91 0.95 224 1 0.41 1.00 0.58 14

source:

MODEL PERFORMANCE ON TEST DATA

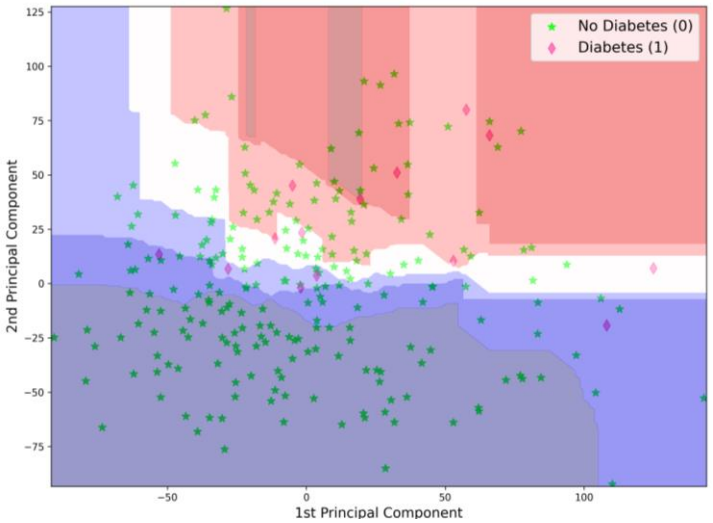


Model	Classification Report
Support Vector Machine	Confusion Matrix: [[96 0] [6 0]] precision recall f1-score support 0 0.94 1.00 0.97 96 1 0.00 0.00 0.00 6
Logistic Regression	Confusion Matrix: [[73 23] [1 5]] precision recall f1-score support 0 0.99 0.76 0.86 96 1 0.18 0.83 0.29 6
KNN	Confusion Matrix: [[85 11] [2 4]] precision recall f1-score support 0 0.98 0.89 0.93 96 1 0.27 0.67 0.38 6
Gaussian Naïve Bayes	Confusion Matrix: [[89 7] [3 3]] precision recall f1-score support 0 0.97 0.93 0.95 96 1 0.30 0.50 0.37 6
Gradient Boosting	Confusion Matrix: [[95 1] [5 1]] precision recall f1-score support 0 0.95 0.99 0.97 96 1 0.50 0.17 0.25 6
Decision Tree	Confusion Matrix: [[91 5] [2 4]] precision recall f1-score support 0 0.98 0.95 0.96 96 1 0.44 0.67 0.53 6
adaboost	Confusion Matrix: [[86 10] [3 3]] precision recall f1-score support 0 0.97 0.90 0.93 96 1 0.23 0.50 0.32 6
Random Forest	Confusion Matrix: [[89 7] [2 4]] precision recall f1-score support 0 0.98 0.93 0.95 96 1 0.36 0.67 0.47 6

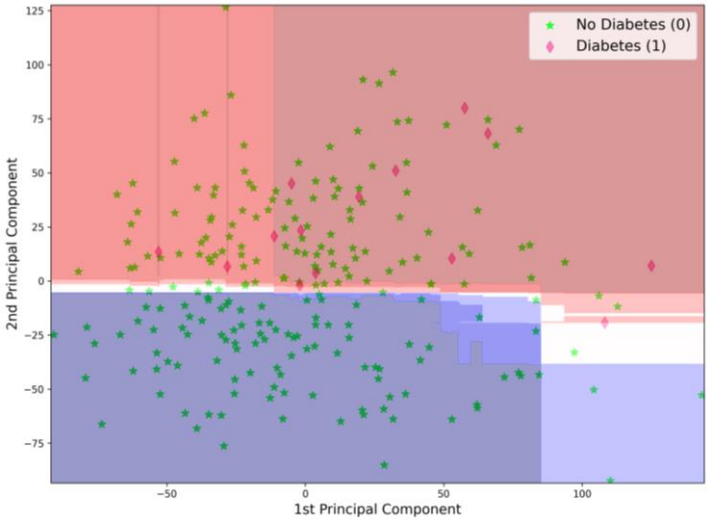
source:

DECISION BOUNDARIES LEARNED ON PCA SPACE FOR TRAINING

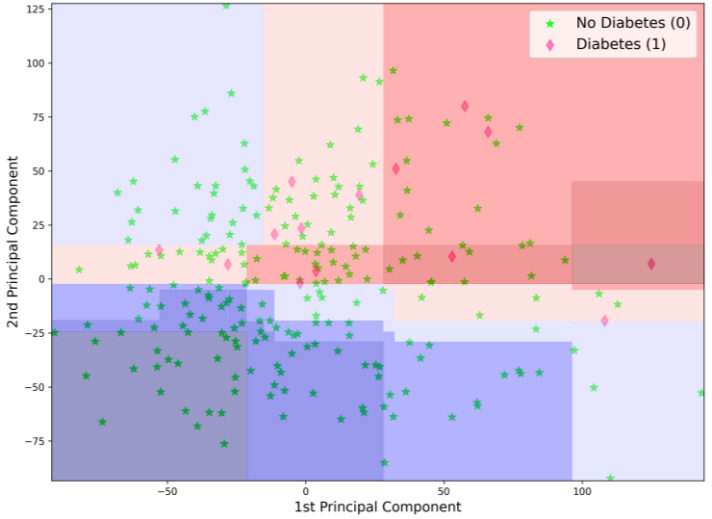
KNN



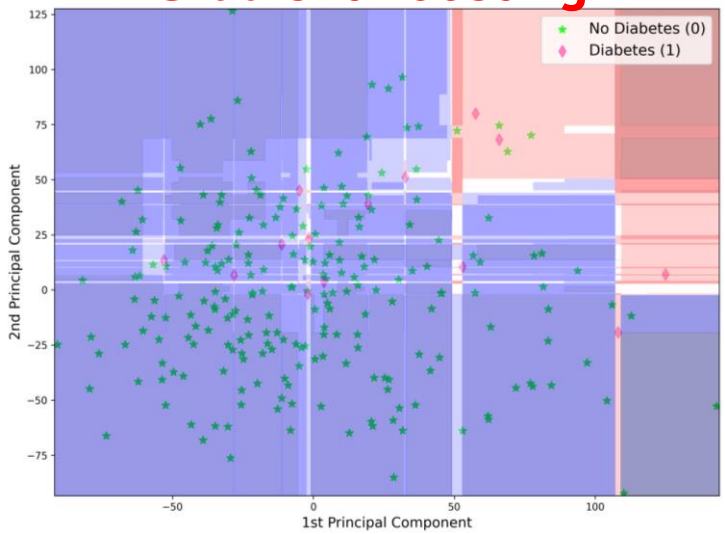
Random forests



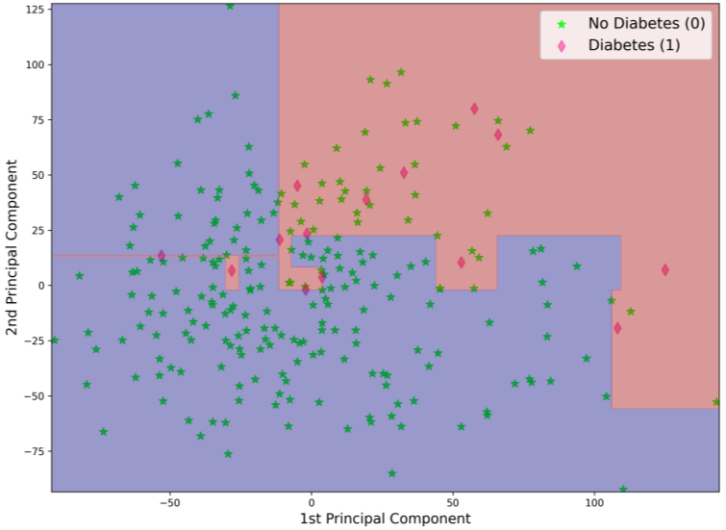
Adaboost



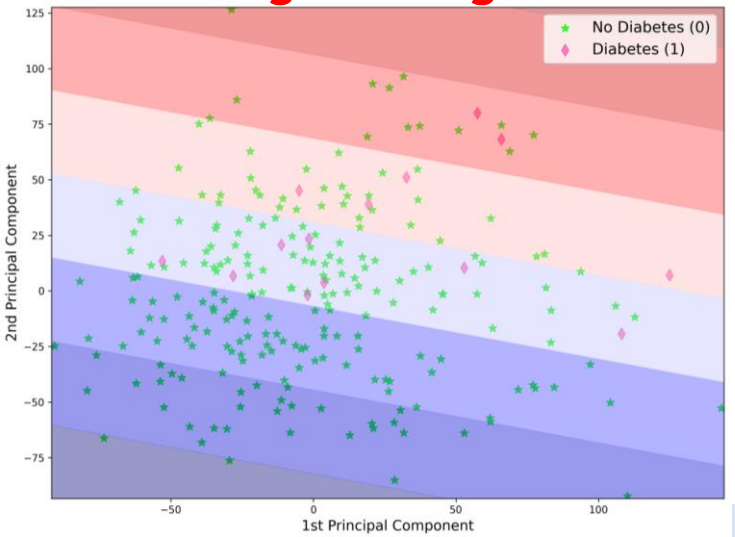
Gradient Boosting



Decision Tree



Logistic Regression



source:

CONCLUSIONS

- The imbalance in distribution of positive class is a major issue in fitting a classification model to diagnostics datasets such as diabetes, cancer etc.
- Model performance can be boosted with under-sampling, over-sampling or both on either majority or minority class
- Classes seemed to not be linearly separable
- Ensemble models such as random forests and adaboost seemed to be able to distinguish between classes given their precision, recall and f1-scores.
- Imbalance in dataset should be address if one of these models were to be deployed to be used by physicians.

source: