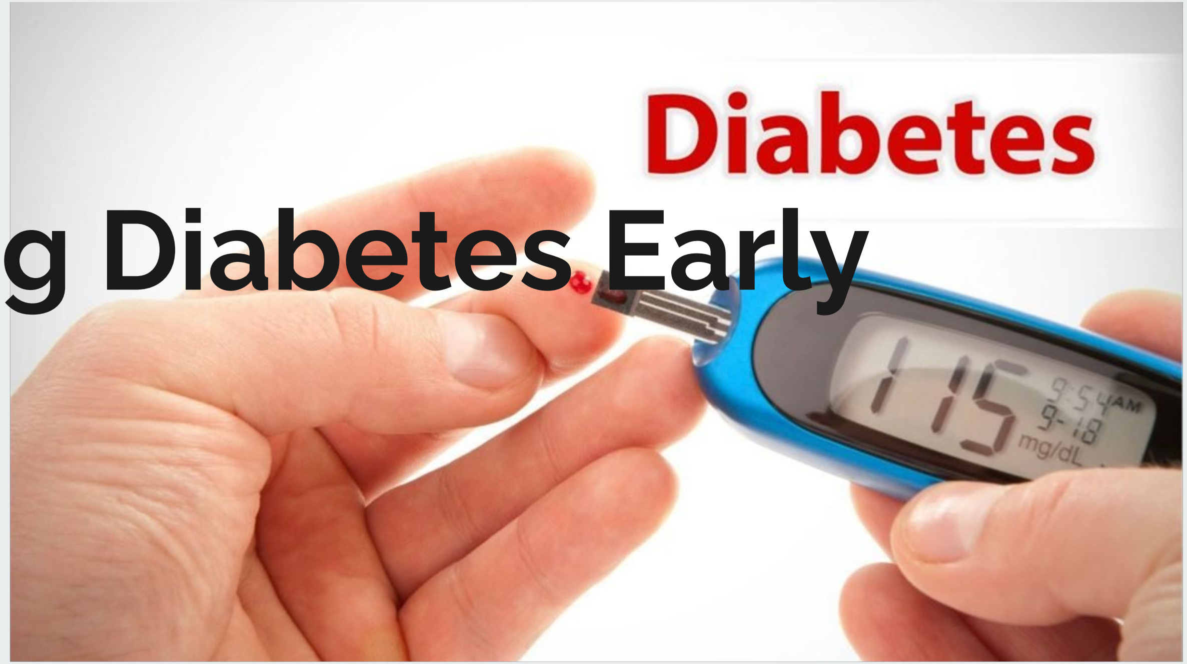
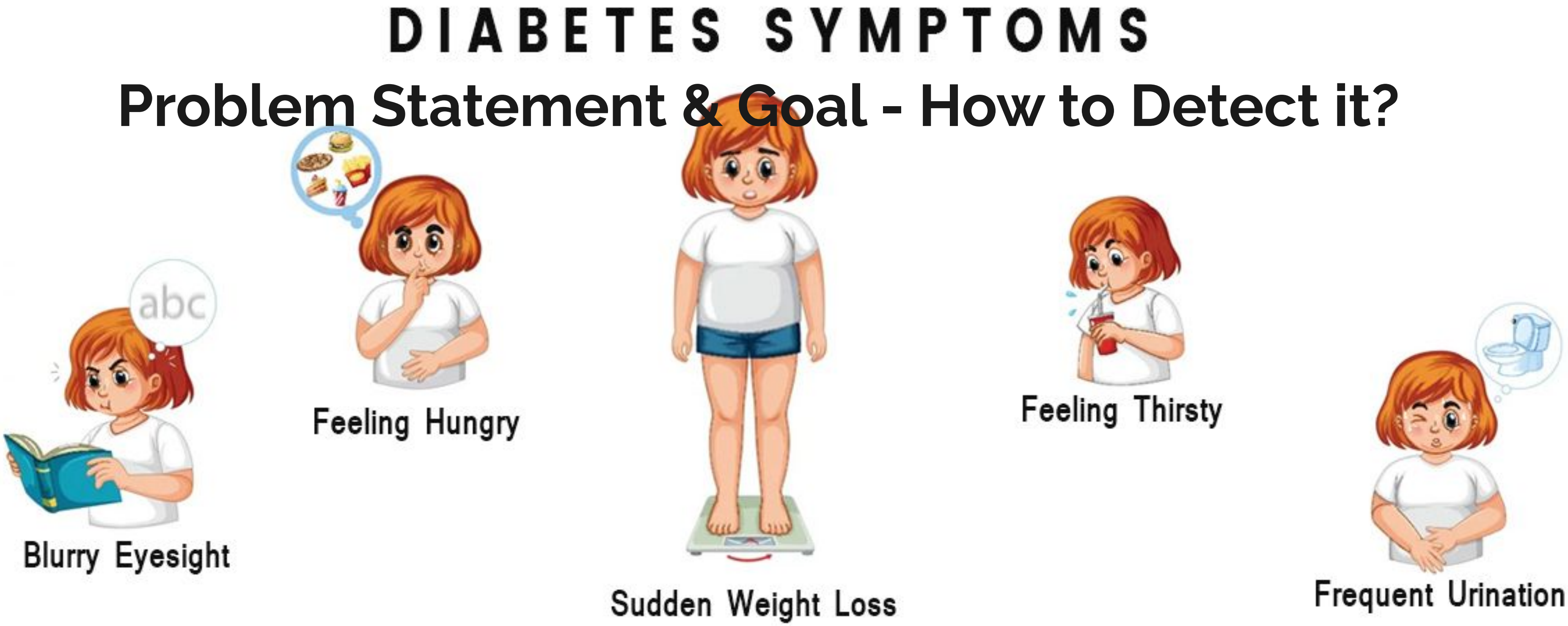
Detecting Diabetes Early



Detecting Diabetes Early

Daniel Kim

Problem Statement & Goal - How to Detect it?



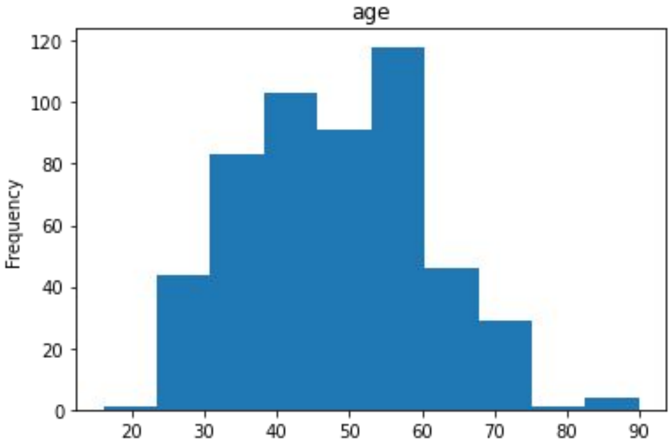
Data Wrangling

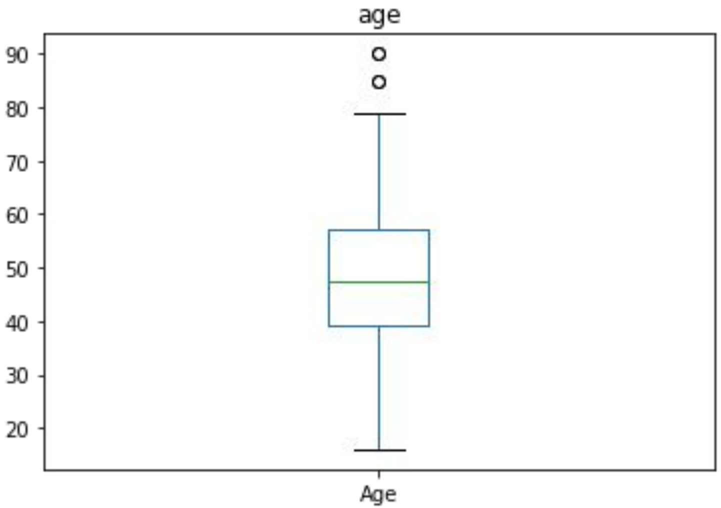
1. Detecting NaN & Redundancy
    a. There were no NaN
    b. Although there were redundancies, it is too early to decide to remove redundancies because the redundancies are not the same but different individuals having same symptoms.
    c. Except for Age, all other features are binary, so it was needed to do one hot encoding to make a machine learning understand the data better.

2. One Hot Encoding

   After one -hot encoding, all the other binary columns ending with such as '_Female', '_Negative', or '_No' were removed.

Exploratory Data Analysis - Histogram & Boxplot

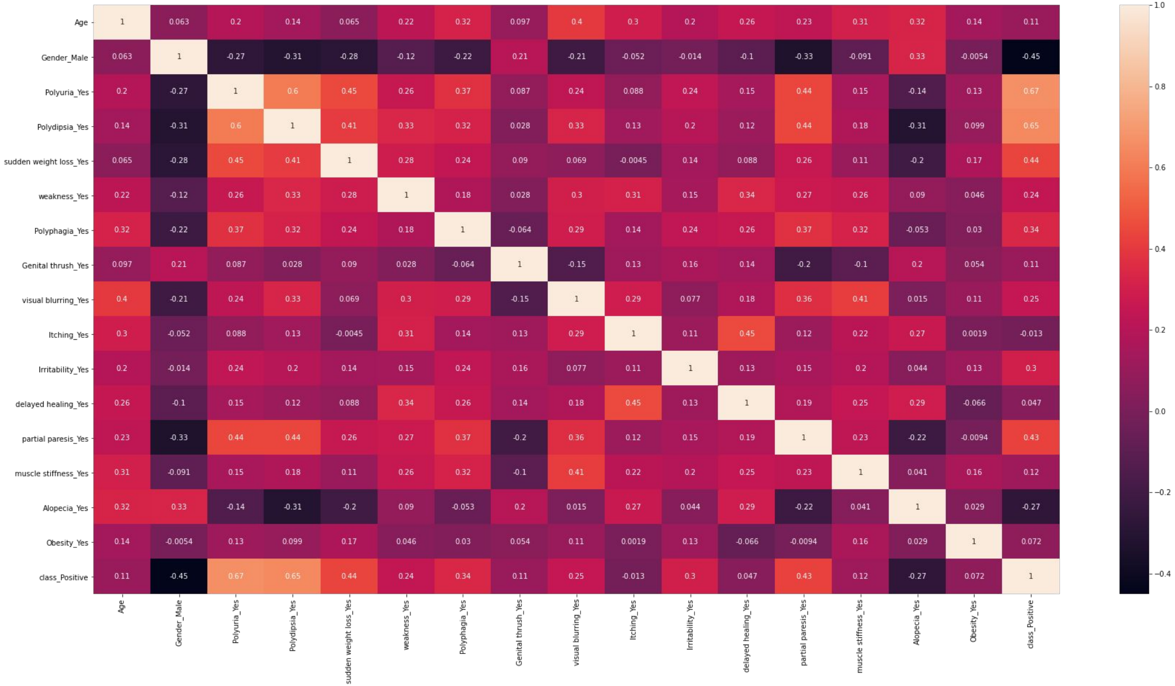




‘Age’ is the only feature showing normal distribution with a little right-skewness because all other features are binary. By its Boxplot, ‘Age’ seems having 2 outliers.

Exploratory Data Analysis - Heat Map

Too many features seem correlated to each other, so it is needed to simplify their correlation by feature importance of several machine learning algorithms.



	Age	Gender_Male	Polyuria_Yes	Polydipsia_Yes	sudden weight loss_Yes	weakness_Yes	Polyphagia_Yes	Genital thrush_Yes	visual blurring_Yes	Itching_Yes	Irritability_Yes	delayed healing_Yes	partial paresis_Yes	muscle stiffness_Yes	Alopecia_Yes	Obesity_Yes	class_Positive
Age	1	0.063	0.2	0.14	0.065	0.22	0.32	0.097	0.4	0.3	0.2	0.26	0.23	0.31	0.32	0.14	0.11
Gender_Male	0.063	1	-0.27	-0.31	-0.28	-0.12	-0.22	0.21	-0.21	-0.052	-0.014	-0.1	-0.33	-0.091	0.33	-0.0054	-0.45
Polyuria_Yes	0.2	-0.27	1	0.6	0.45	0.26	0.37	0.087	0.24	0.088	0.24	0.15	0.44	0.15	-0.14	0.13	0.67
Polydipsia_Yes	0.14	-0.31	0.6	1	0.41	0.33	0.32	0.028	0.33	0.13	0.2	0.12	0.44	0.18	-0.31	0.099	0.65
sudden weight loss_Yes	0.065	-0.28	0.45	0.41	1	0.28	0.24	0.09	0.069	-0.0045	0.14	0.088	0.26	0.11	-0.2	0.17	0.44
weakness_Yes	0.22	-0.12	0.26	0.33	0.28	1	0.18	0.028	0.3	0.31	0.15	0.34	0.27	0.26	0.09	0.046	0.24
Polyphagia_Yes	0.32	-0.22	0.37	0.32	0.24	0.18	1	-0.064	0.29	0.14	0.24	0.26	0.37	0.32	-0.053	0.03	0.34
Genital thrush_Yes	0.097	0.21	0.087	0.028	0.09	0.028	-0.064	1	-0.15	0.13	0.16	0.14	-0.2	-0.1	0.2	0.054	0.11
visual blurring_Yes	0.4	-0.21	0.24	0.33	0.069	0.3	0.29	-0.15	1	0.29	0.077	0.18	0.36	0.41	0.015	0.11	0.25
Itching_Yes	0.3	-0.052	0.088	0.13	-0.0045	0.31	0.14	0.13	0.29	1	0.11	0.45	0.12	0.22	0.27	0.0019	-0.013
Irritability_Yes	0.2	-0.014	0.24	0.2	0.14	0.15	0.24	0.16	0.077	0.11	1	0.13	0.15	0.2	0.044	0.13	0.3
delayed healing_Yes	0.26	-0.1	0.15	0.12	0.088	0.34	0.26	0.14	0.18	0.45	0.13	1	0.19	0.25	0.29	-0.066	0.047
partial paresis_Yes	0.23	-0.33	0.44	0.44	0.26	0.27	0.37	-0.2	0.36	0.12	0.15	0.19	1	0.23	-0.22	-0.0094	0.43
muscle stiffness_Yes	0.31	-0.091	0.15	0.18	0.11	0.26	0.32	-0.1	0.41	0.22	0.2	0.25	0.23	1	0.041	0.16	0.12
Alopecia_Yes	0.32	0.33	-0.14	-0.31	-0.2	0.09	-0.053	0.2	0.015	0.27	0.044	0.29	-0.22	0.041	1	0.029	-0.27
Obesity_Yes	0.14	-0.0054	0.13	0.099	0.17	0.046	0.03	0.054	0.11	0.0019	0.13	-0.066	-0.0094	0.16	0.029	1	0.072
class_Positive	0.11	-0.45	0.67	0.65	0.44	0.24	0.34	0.11	0.25	-0.013	0.3	0.047	0.43	0.12	-0.27	0.072	1

Machine Learning - Feature Importance

‘Polyuria’ showed the strongest feature when it comes to identifying ‘class_positive’. So we can conclude that ‘Polyuria’ plays a key role in detect diabetes.

Besides ‘Polyuria’, there are the following common strong features to detect diabetes even if their importance rates vary: Age, Gender_Male, and Polydipsia.

Algorithms	Decision Tree Classifier	Random Forest Classifier	Gradient Classifier
Feature Importance %	Age: 12% Gender_Male: 10% **Polyuria: 42%** Polydipsia: 6% Irritability: 9% muscle stiffness: 8% Alopecia: 7%	Age: 10% Gender_Male: 9% **Polyuria: 21%** Polydipsia: 18% sudden weight loss: 6% partial paresis: 5%	Age: 8% Gender_Male: 11% **Polyuria: 31%** Polydipsia: 25% Irritability: 6% Alopecia: 7%

Machine Learning - Detection Accuracy

The Decision Tree Classifier showed the best score (97%). Other algorithms showed quite remarkable accuracy scores, but their scores are still lower than the Decision Tree Classifier. The reason for the difference seems due to ‘max_depth’ per each algorithm; the Decision Tree Classifier had more depths than others: 30 vs. 5.

Algorithms	**Decision Tree Classifier**	Random Forest Classifier	Gradient Boost Classifier
Accuracy (R-Squared)	**97%**	93%	95%
Best Parameters by Randomized Search CV	**'splitter': 'random', 'random_state': 42, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': 30, 'criterion': 'entropy', 'class_weight': 'balanced'**	'n_estimators': 1200, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': None, 'max_depth': 5, 'criterion': 'gini', 'class_weight': 'balanced'	'n_estimators': 500, 'min_samples_split': 15, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 5, 'loss': 'exponential', 'criterion': 'mae'

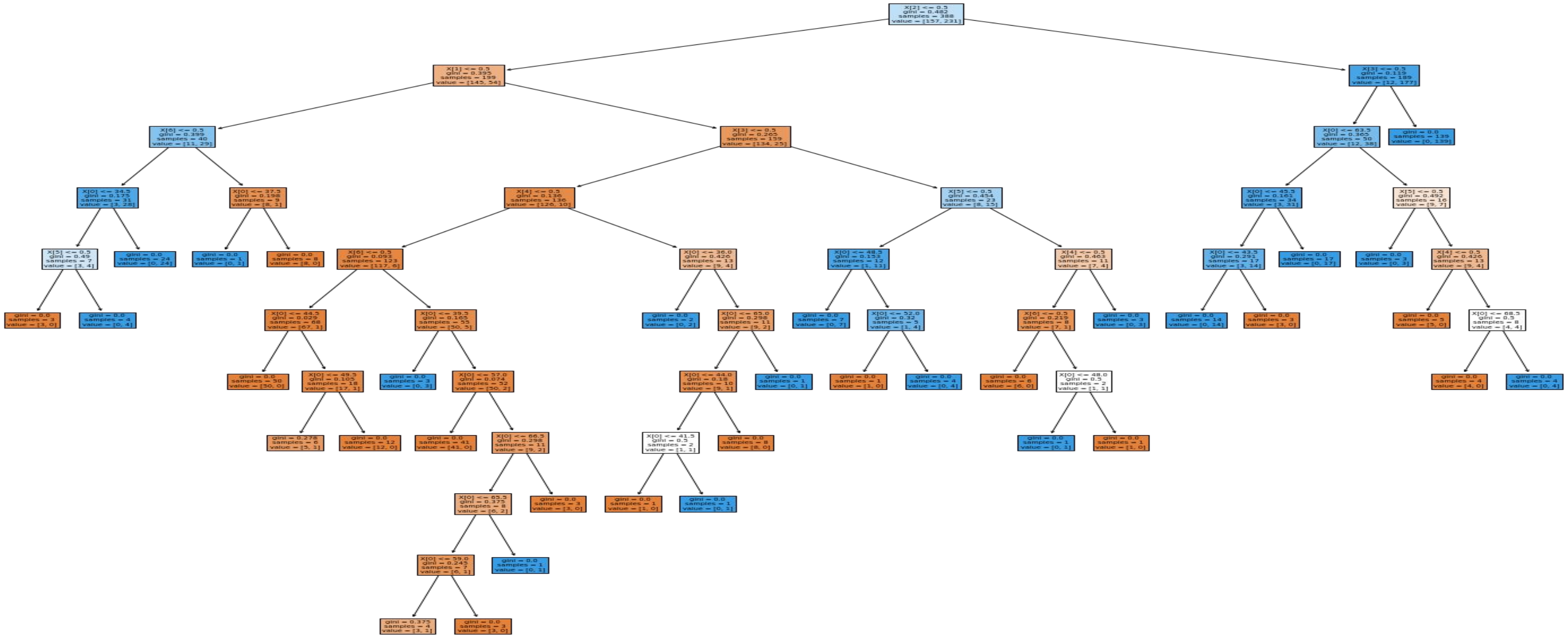
Recommendation - For Better Detection of Diabetes

In order to identify diabetes, you may conduct either of the following testing methods on a trade-off basis for both efficiency and effectiveness.

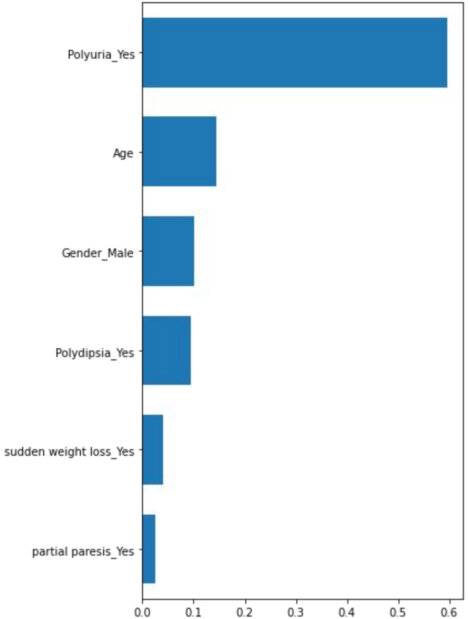
- Combinations of the above-mentioned features can give you optimal diabetes procedures. For example, a test of ‘Polyuria’ should be an independent variable for detecting diabetes. Such following features should be dependant variables with higher weights than others: ‘Age’, ‘Gender_Male’, and ‘Polydipsia’. And then, the other else features should be optional test variables with lower weights.

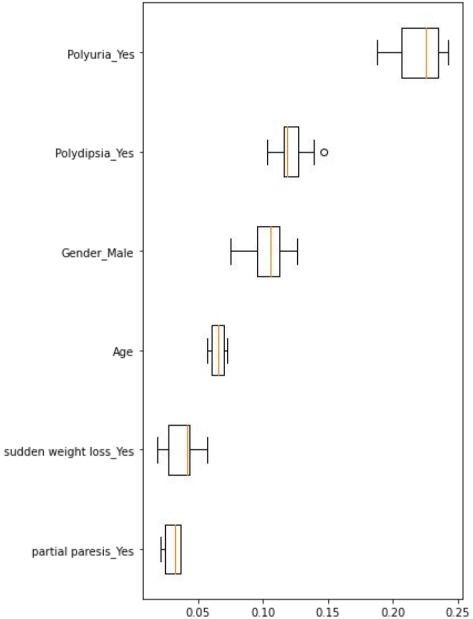
- You can just conduct a diabetes test by the Decision Tree Classifier method.

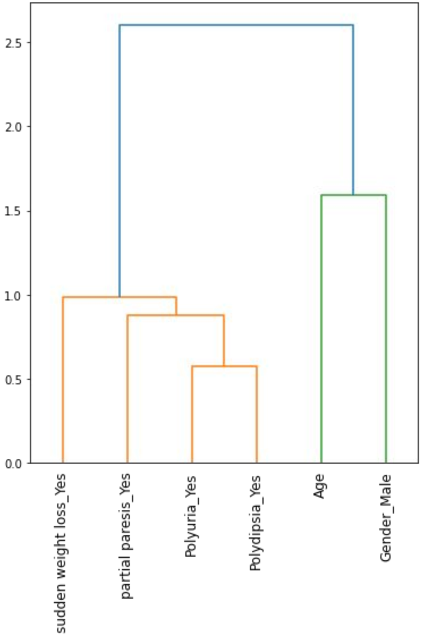
In-Depth Analysis - Decision Tree

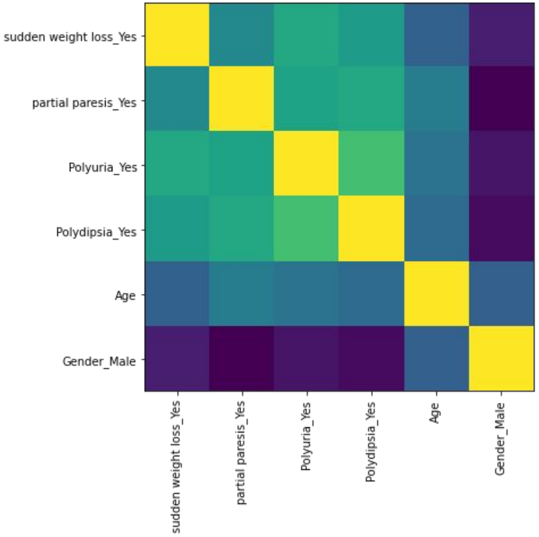

X[2] <= 0.5 gini = 0.482 samples = 388 value = [157, 231]
X[1] <= 0.5 gini = 0.395 samples = 199 value = [145, 54]
X[3] <= 0.5 gini = 0.119 samples = 189 value = [12, 177]
X[6] <= 0.5 gini = 0.399 samples = 40 value = [11, 29]
X[3] <= 0.5 gini = 0.265 samples = 159 value = [134, 25]
X[0] <= 63.5 gini = 0.365 samples = 50 value = [12, 38]
gini = 0.0 samples = 139 value = [0, 139]
X[0] <= 34.5 gini = 0.175 samples = 31 value = [3, 28]
X[0] <= 37.5 gini = 0.198 samples = 9 value = [8, 1]
X[4] <= 0.5 gini = 0.136 samples = 136 value = [126, 10]
X[5] <= 0.5 gini = 0.454 samples = 23 value = [8, 15]
X[0] <= 45.5 gini = 0.161 samples = 34 value = [3, 31]
X[5] <= 0.5 gini = 0.492 samples = 16 value = [9, 7]
X[5] <= 0.5 gini = 0.49 samples = 7 value = [3, 4]
gini = 0.0 samples = 24 value = [0, 24]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 8 value = [8, 0]
X[6] <= 0.5 gini = 0.093 samples = 123 value = [117, 6]
X[0] <= 36.0 gini = 0.426 samples = 13 value = [9, 4]
X[0] <= 48.5 gini = 0.153 samples = 12 value = [1, 11]
X[4] <= 0.5 gini = 0.463 samples = 11 value = [7, 4]
X[0] <= 43.5 gini = 0.291 samples = 17 value = [3, 14]
gini = 0.0 samples = 17 value = [0, 17]
gini = 0.0 samples = 3 value = [0, 3]
X[4] <= 0.5 gini = 0.426 samples = 13 value = [9, 4]
gini = 0.0 samples = 3 value = [3, 0]
gini = 0.0 samples = 4 value = [0, 4]
X[0] <= 44.5 gini = 0.029 samples = 68 value = [67, 1]
X[0] <= 39.5 gini = 0.165 samples = 55 value = [50, 5]
gini = 0.0 samples = 2 value = [0, 2]
X[0] <= 65.0 gini = 0.298 samples = 11 value = [9, 2]
gini = 0.0 samples = 7 value = [0, 7]
X[0] <= 52.0 gini = 0.32 samples = 5 value = [1, 4]
X[6] <= 0.5 gini = 0.219 samples = 8 value = [7, 1]
gini = 0.0 samples = 3 value = [0, 3]
gini = 0.0 samples = 14 value = [0, 14]
gini = 0.0 samples = 3 value = [3, 0]
gini = 0.0 samples = 5 value = [5, 0]
X[0] <= 68.5 gini = 0.5 samples = 8 value = [4, 4]
gini = 0.0 samples = 50 value = [50, 0]
X[0] <= 49.5 gini = 0.105 samples = 18 value = [17, 1]
gini = 0.0 samples = 3 value = [0, 3]
X[0] <= 57.0 gini = 0.074 samples = 52 value = [50, 2]
X[0] <= 44.0 gini = 0.18 samples = 10 value = [9, 1]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 1 value = [1, 0]
gini = 0.0 samples = 4 value = [0, 4]
gini = 0.0 samples = 6 value = [6, 0]
X[0] <= 48.0 gini = 0.5 samples = 2 value = [1, 1]
gini = 0.0 samples = 4 value = [4, 0]
gini = 0.0 samples = 4 value = [0, 4]
gini = 0.278 samples = 6 value = [5, 1]
gini = 0.0 samples = 12 value = [12, 0]
gini = 0.0 samples = 41 value = [41, 0]
X[0] <= 66.5 gini = 0.298 samples = 11 value = [9, 2]
X[0] <= 41.5 gini = 0.5 samples = 2 value = [1, 1]
gini = 0.0 samples = 8 value = [8, 0]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 1 value = [1, 0]
X[0] <= 65.5 gini = 0.375 samples = 8 value = [6, 2]
gini = 0.0 samples = 3 value = [3, 0]
gini = 0.0 samples = 1 value = [1, 0]
gini = 0.0 samples = 1 value = [0, 1]
X[0] <= 59.0 gini = 0.245 samples = 7 value = [6, 1]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.375 samples = 4 value = [3, 1]
gini = 0.0 samples = 3 value = [3, 0]

In-Depth Analysis - Random Forest Classifier

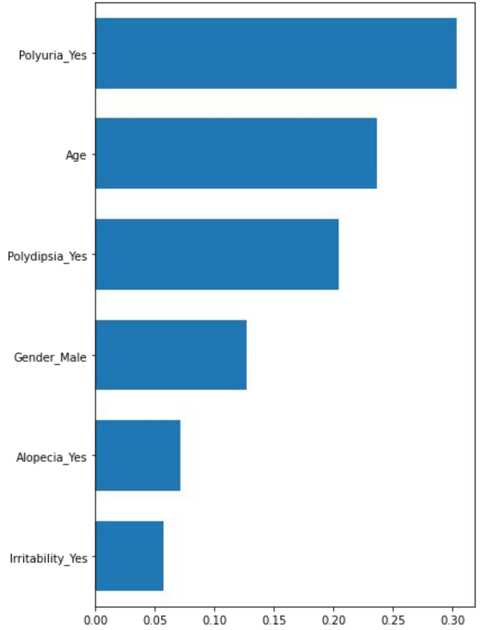


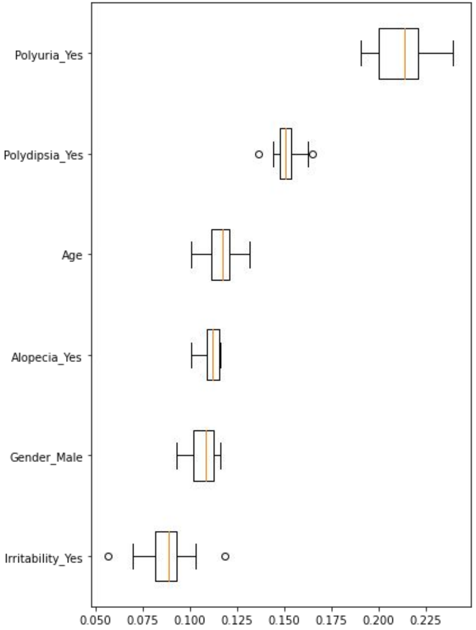


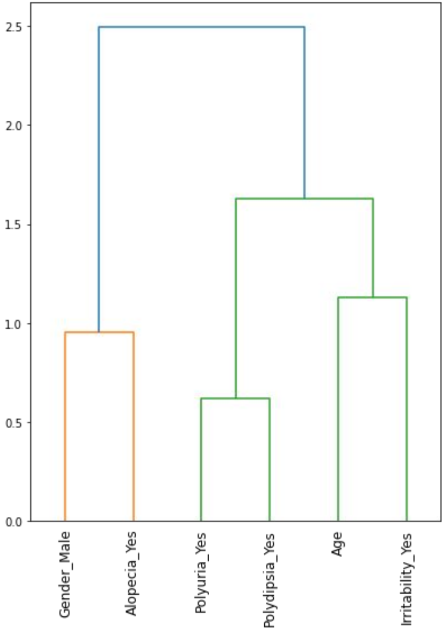


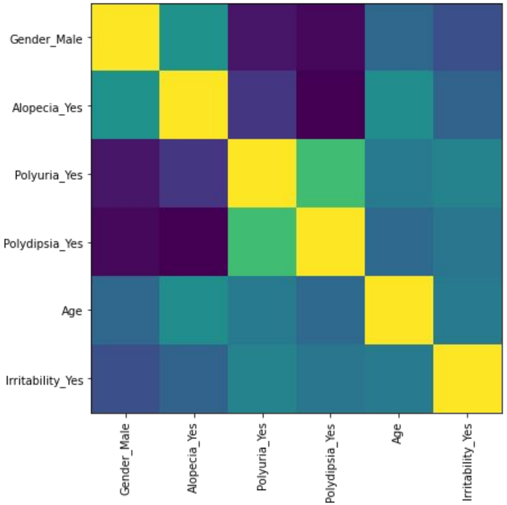


In-Depth Analysis - Gradient Boost Classifier


Polyuria_Yes
Age
Polydipsia_Yes
Gender_Male
Alopecia_Yes
Irritability_Yes


Polyuria_Yes
Polydipsia_Yes
Age
Alopecia_Yes
Gender_Male
Irritability_Yes


Gender_Male
Alopecia_Yes
Polyuria_Yes
Polydipsia_Yes
Age
Irritability_Yes


Gender_Male
Alopecia_Yes
Polyuria_Yes
Polydipsia_Yes
Age
Irritability_Yes
Gender_Male
Alopecia_Yes
Polyuria_Yes
Polydipsia_Yes
Age
Irritability_Yes