Department of Applied Data Science

DATA 230

DataVisualization-230-26

Instructor: Guannan Liu

PROJECT REPORT

Diabetes Prediction in Pima Indian women

Group 8

Group Members

Neha Thakur

Sakshi Jain

Veena Ramesh Beknal

INDEX

Abstract

Pima Indians are better known worldwide for their diabetes than for their culture and history. Pima is typically referred to as “River People” a group of native North American Indians living in Arizona, USA. More than half of all Pima Indians over age 35 have diabetes, a condition arising from a body's decreased ability to metabolize glucose (Knowler et al. 1990). Adult-onset, non-insulin-dependent diabetes, now called type 2 Diabetes, is the most common form of diabetes among Pima Indians due to several historical and sociological factors. To understand diabetes better, we are looking at the Pima Indian Diabetes Database dataset which includes females from age 21 – 81 from the Pima Indian heritage. The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The dataset consists of various biological and lifestyle parameters such as Pregnancy, glucose, blood pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and outcome for female Pima Indians. We propose using 1 or more supervised machine learning models such as Logistic Regression, Decision Trees, Random Forest, etc. to predict the ‘Outcome’ (target variable in our dataset) along with exploratory data analyses and feature importance to understand further the nuances of relationships between the biological parameters and the prevalence of diabetes. The real-world implications among the Pima Indian population are crucial for healthcare professionals, aiding in identifying high-risk individuals.

Data overview

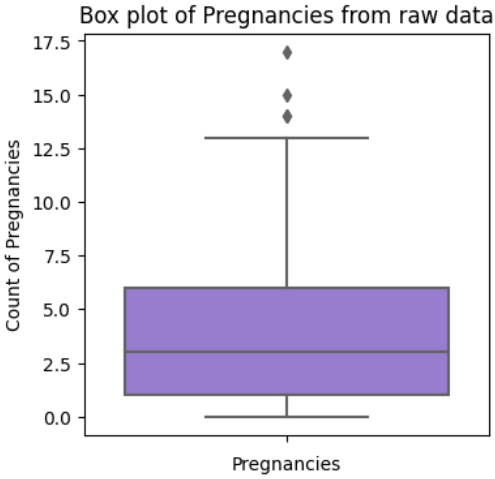
There are 8 feature variables in our dataset, pregnancies, glucose, blood pressure, insulin, skin thickness, BMI, diabetes pedigree function, age, and the target variable is outcome. The data dictionary is shown in the table below:

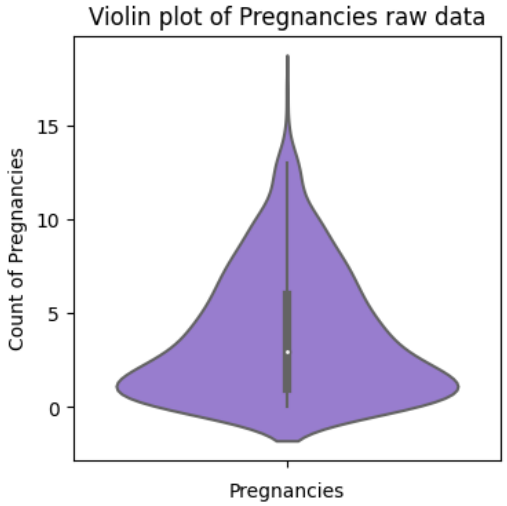
Column	Definition	# missing values	# 0 values	Is 0 valid?
Pregnancies	# of pregnancies	0	111	Yes
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	0	5	No
BloodPressure	Diastolic blood pressure (mm Hg)	0	35	No
SkinThickness	Triceps skin fold thickness (mm)	0	227	No
Insulin	2-Hour serum insulin (mu U/ml)	0	374	No
BMI	Body mass index (weight in kg/(height in m)^2)	0	11	No
DiabetesPedigreeFunction	Diabetes pedigree function	0	0	No
Age	Age in years	0	0	No
Outcome	1 = diabetic, 0 = not diabetic	0	NA	NA

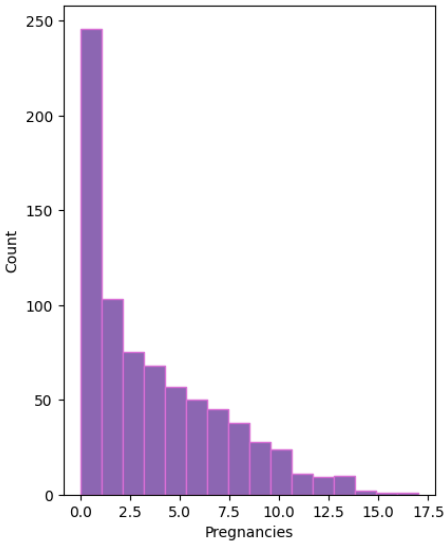
EDA – Univariate analyses and data treatment

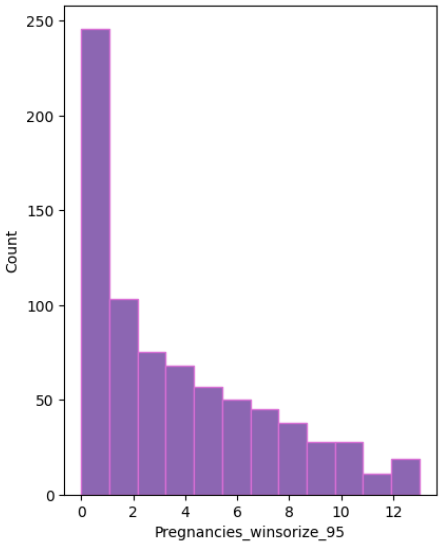
Pregnancies

Box plot, violin plot and histogram for Pregnancies are shown below:









Observations on Pregnancies and Treatment:

- 0 is a valid value for Pregnancies and is not considered a missing value, 14% of the records are with 0 Pregnancies
- The distribution of the data is somewhat right-skewed i.e. positive skewness with most values being on the lower side

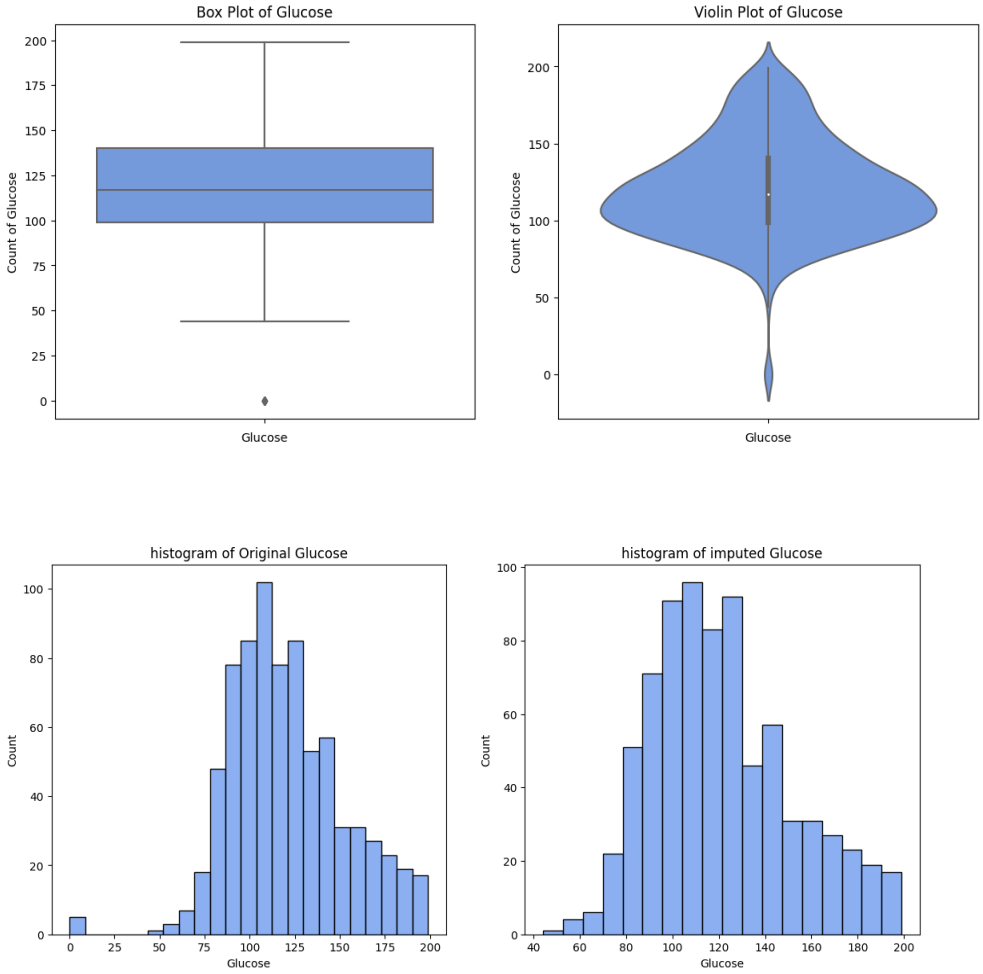
- The upper limit (whisker) for Pregnancies is 13.5 which when rounded down gives us 13;

values beyond 13 are considered outliers

- In the case of Pregnancies, we decided to replace outliers values with the 95th percentile value - this is called winsorization and the effect is noticeable in the pre and post histograms

Glucose

Box plot, violin plot, and histogram for Glucose are shown below:

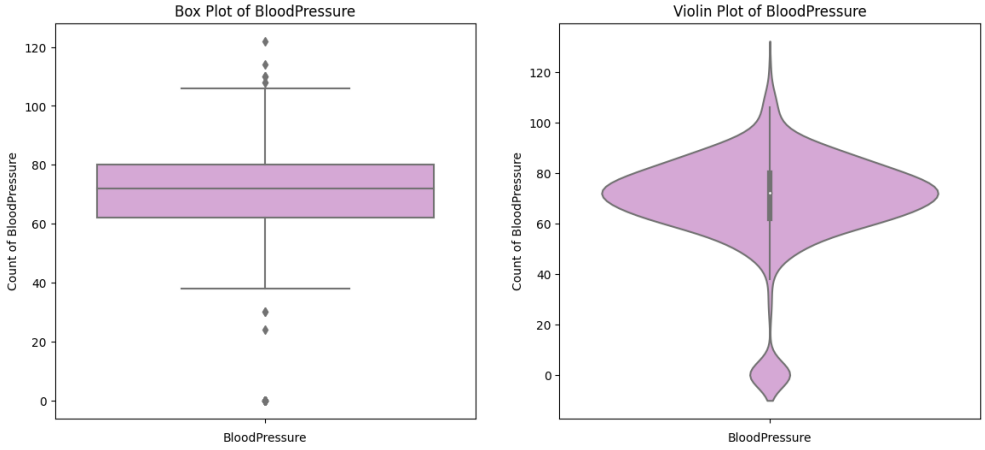


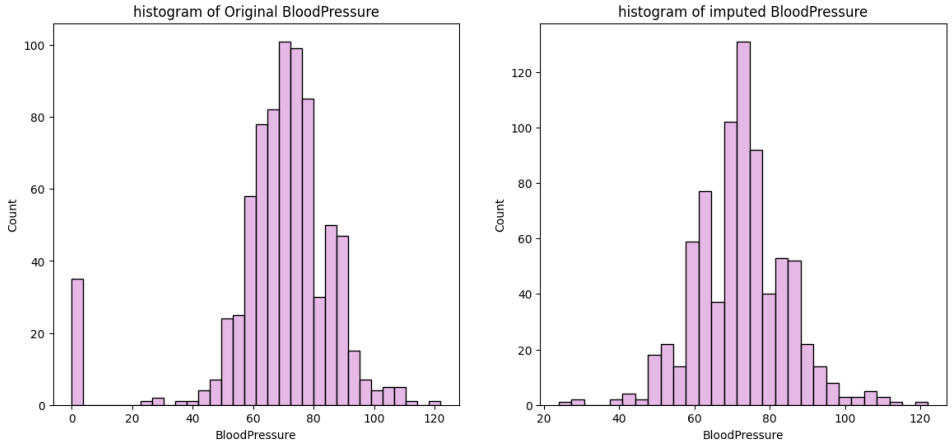
<u>Observations on Glucose and treatment</u>:

- 0 is not a valid value for Glucose and hence 0 values have been considered missing values
- 5 values were missing in the dataset and were imputed with the median value (from non-zero observations)
- The upper limit (whisker) and lower limit (whisker) for Glucose are 201 and 39 respectively with values beyond being considered outliers; no outliers were observed for Glucose and hence no outlier treatment is required.

Blood Pressure

Box plot, violin plot, and histogram for Blood Pressure are shown below:



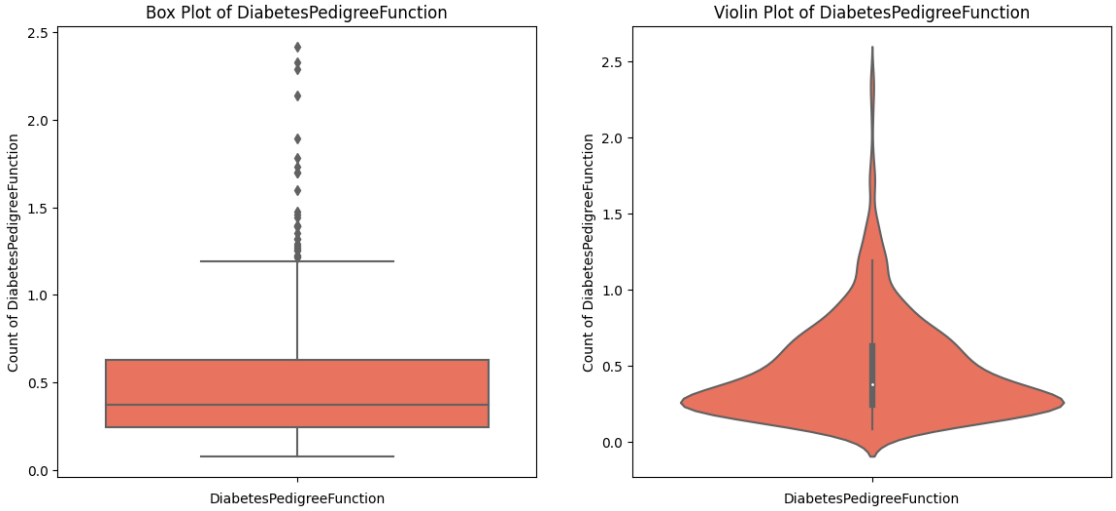


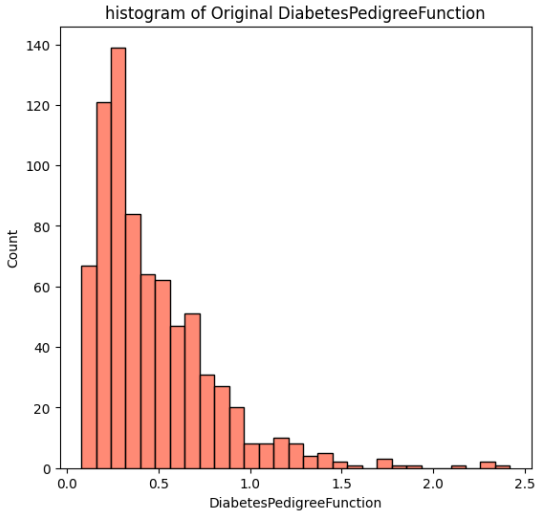
Observations on Blood Pressure and treatment:

- 0 is not a valid value for Blood pressure and hence 0 values have been considered as missing values
- 35 values were missing in the dataset and were imputed with the mean since the distribution is somewhat normal from visual inspection – the effect of imputation is noticeable in the pre and post-imputation histograms
- Outliers are values > 104 (upper limit) and < 40 (lower limit) but this is influenced by the 35 missing values (0), hence we're not removing any outliers based on the distribution observed in the raw data

Diabetes Pedigree Function

Box plot, violin plot and histogram for Diabetes Pedigree Function are shown below:



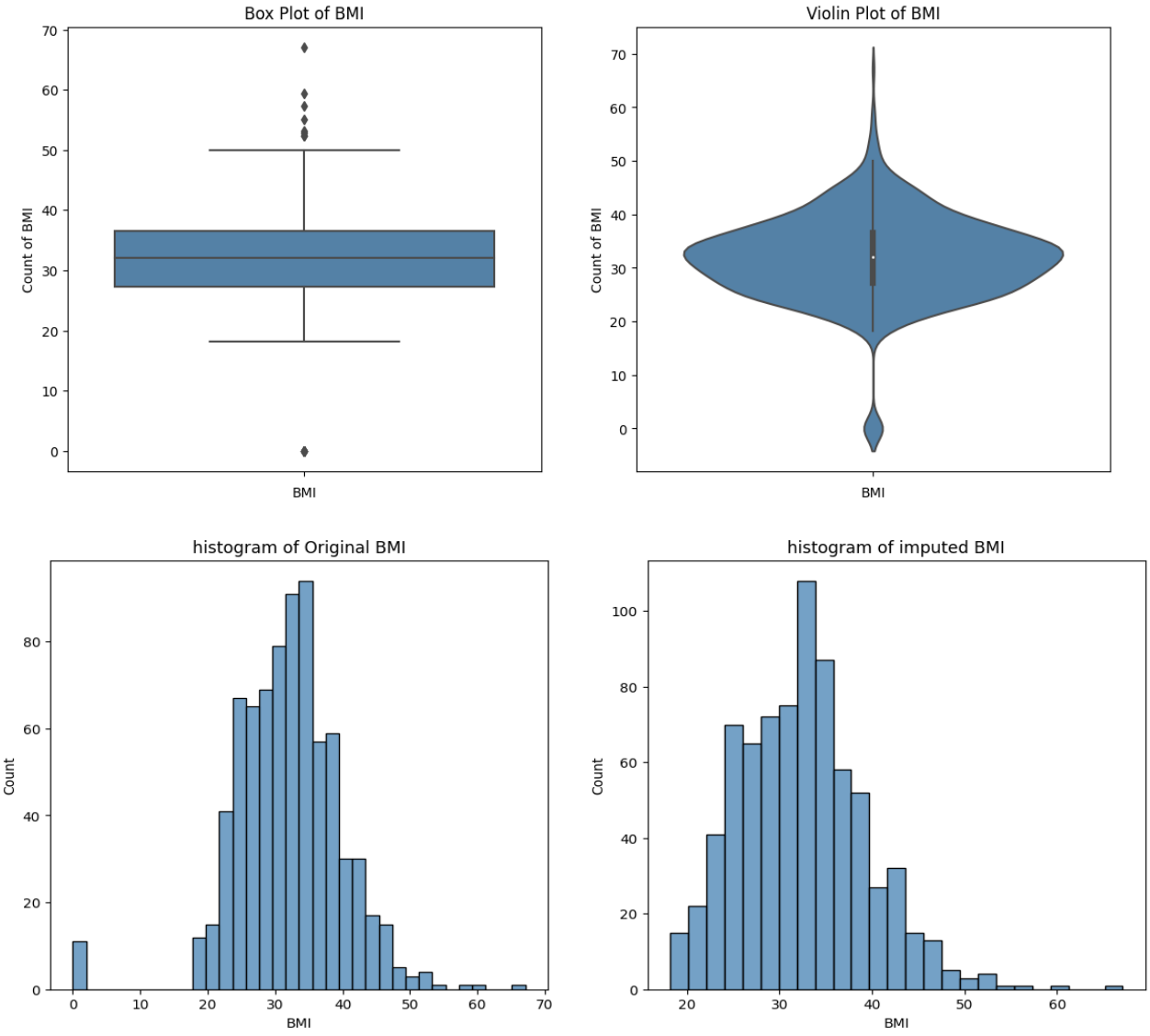


Observations on DPF:

- No missing values were observed in Diabetes Pedigree Function and the distribution is right skewed
- The values beyond 1.2 (upper limit) are considered shown outliers based on the box plot - but since we're unaware of the appropriate range for this feature, we've not treated any outliers
- Since there are no missing or zero values, there is no need to perform any imputation of the column

BMI

Box plot, violin plot, and histogram for BMI are shown below:

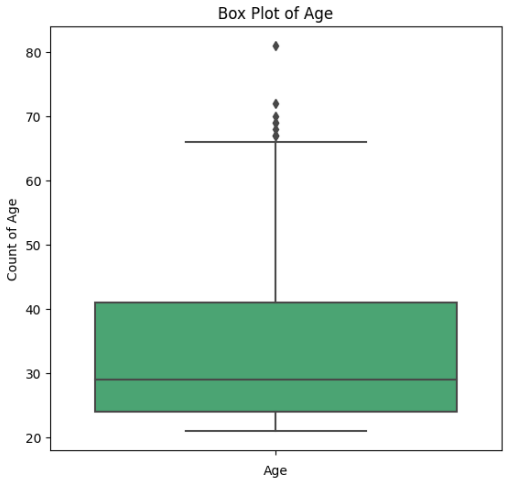


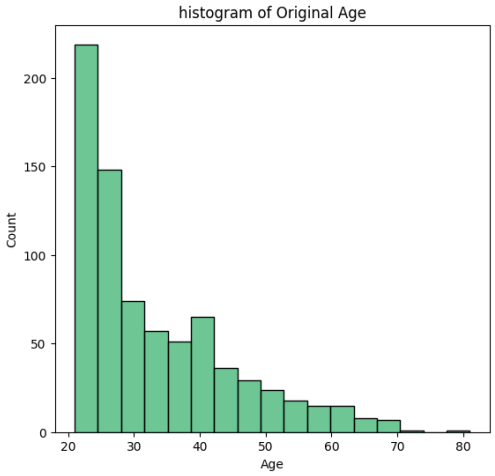
<u>Observations on BMI and treatment</u>:

- There are 11 missing values which are imputed with the median since the data is not normally distributed – this is noticeable in the pre and post-imputation histograms
- There are a few outliers with values exceeding 50 but these values are associated with morbid obesity and hence, we have decided not to treat outliers since this may introduce bias in the feature

Age

Box plot, violin plot, and histogram for Age are shown below:



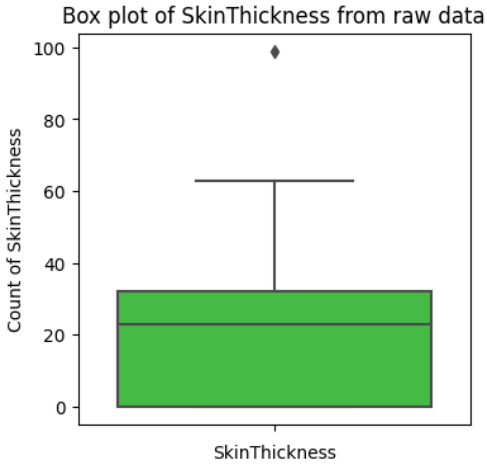


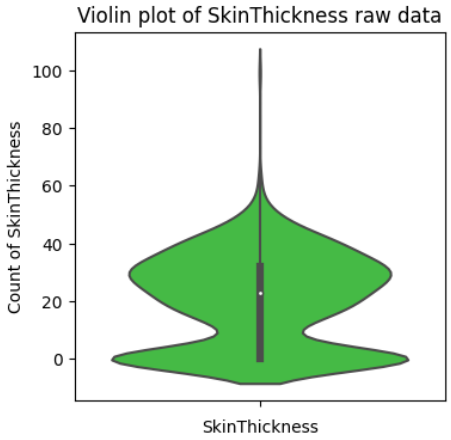
Observations on Age:

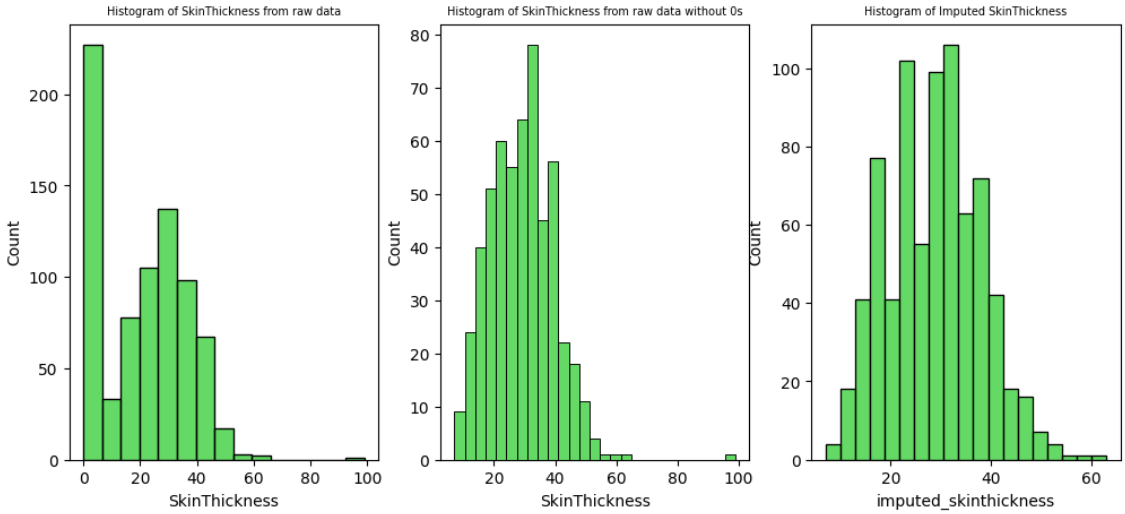
- There is no need to impute the age as no missing value is present
- Although the box plot shows that age beyond 67 (upper whisker) is an outlier, there are Pima women who are older and hence, we have decided to retain the data as is
- The most common age group is between 21 to 40 with 75% of the data being concentrated in this age group

Skin Thickness

Box plot, violin plot, and histogram for Skin Thickness are shown below:







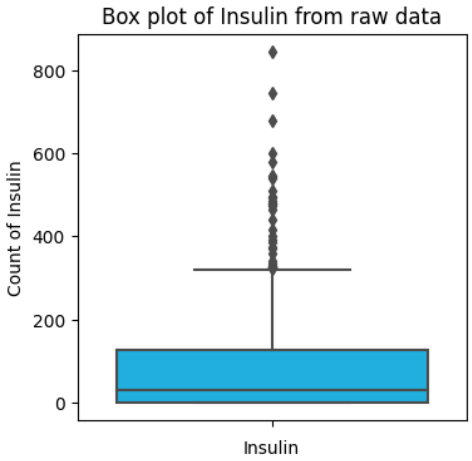
Observations on Skin Thickness and treatment:

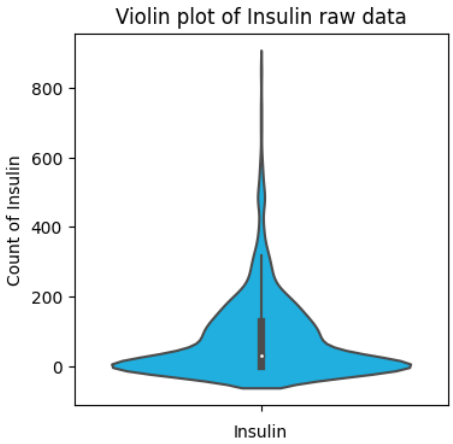
- 0 is not a valid value for this feature and is considered a missing value; however, this makes ~30% of this feature to be missing
- Since a single value imputation like mean or median might remove variability in the data, we decided to adopt a class mean imputation strategy to ensure that imputation was done in a meaningful way
- BMI and Skin thickness show a directional linear trend (as noticeable in the bivariate analysis plot referenced later in the doc) – hence, we can use BMI as the class for imputing skin thickness. Outcome is also used as a class to ensure that there was no bias in the imputation from an independent feature point of view
- Additionally, 1 outlier record with value 99 (noticeable in the box plot) was over the upper whisker (80) and has been winsorized with the 95th percentile value of 44
- Pre and post-imputation distributions are largely comparable - this was our intent as opposed to single value imputation (with mean/median) which would've significantly altered the data distribution with a peak in a single value

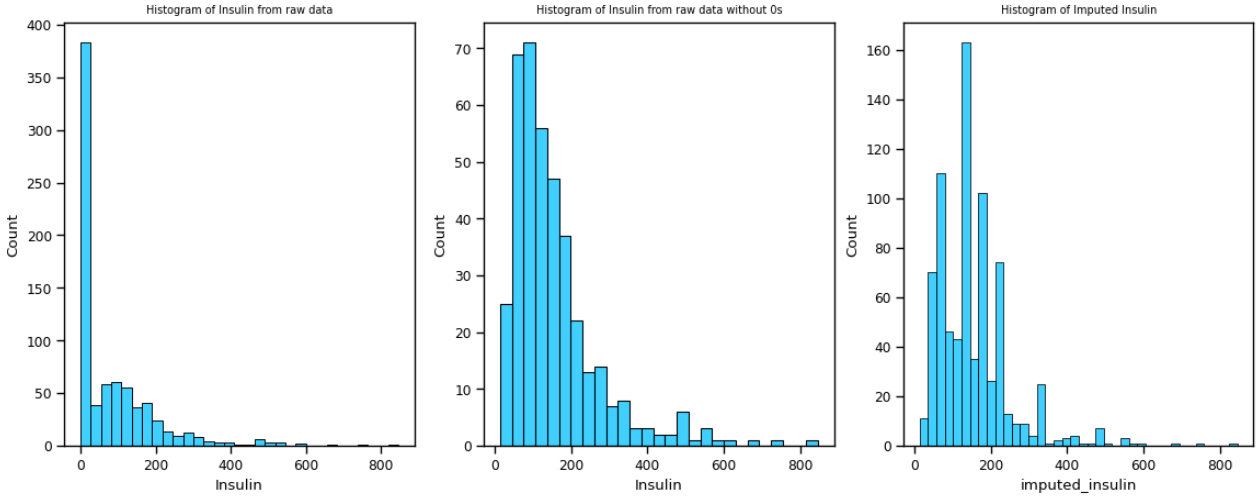
- The transformation of the feature is illustrated in the histogram pair plots

Insulin

Box plot, violin plot, and histogram for Insulin are shown below:







<u>Observations on Insulin and treatment</u>:

- 0 is not a valid value for this feature and is considered a missing value; however, this makes ~49% of this feature to be missing
- Since a single value imputation like mean or median might remove variability in the data, we decided to adopt a class mean imputation strategy to ensure that imputation was done in a meaningful way
- While there is no strong evidence to quantify the relationship between Glucose and

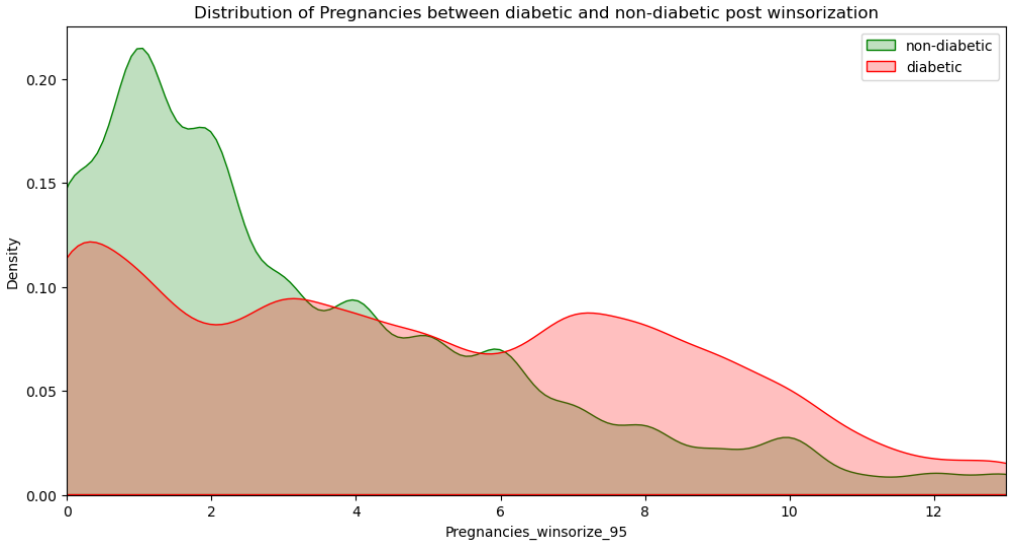
Insulin due to Insulin showing a high % of missing values, it was a judgment call to consider Glucose as a class variable for Insulin. Outcome is also used as a class to ensure that there was no bias in the imputation from an independent feature point of view

- No treatment is done on outliers for Insulin since 49% of data is imputed and we would like to preserve the rest of the data as much as possible
- Pre and post-imputation distributions are somewhat comparable - this was our intent as opposed to single value imputation (with mean/median) which would've significantly altered the data distribution with a peak in a single value
- The transformation of the feature is illustrated in the histogram pair plots

EDA – Bivariate analyses

Dependent-Independent variables' relationships:

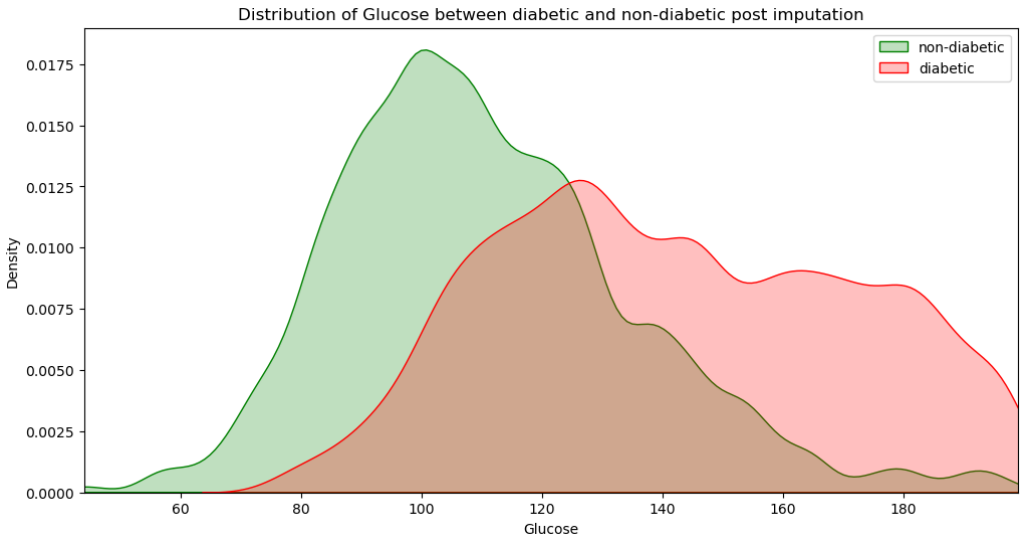
Bivariate analysis for Pregnancies split by Outcome



- The distribution of Pregnancies between Outcome 1 (diabetic) and 0 (non-diabetic) is shown in the above chart
- The distribution of pregnancies between diabetic and non-diabetic is significantly different (independent t-test between these 2 groups has a p-value of 1.5e-09, which is lower than 0.05)
- Based on the shape of distribution seen above, we observed that the density of diabetic women was frequently ~2x more than non-diabetic beyond 7 pregnancies i.e. **Pima women with >=7 pregnancies are ~2x likely to be diagnosed with diabetes vs those with a lower count of pregnancies**
- This conclusion is supported by the independent t-test between 2 groups - < 7

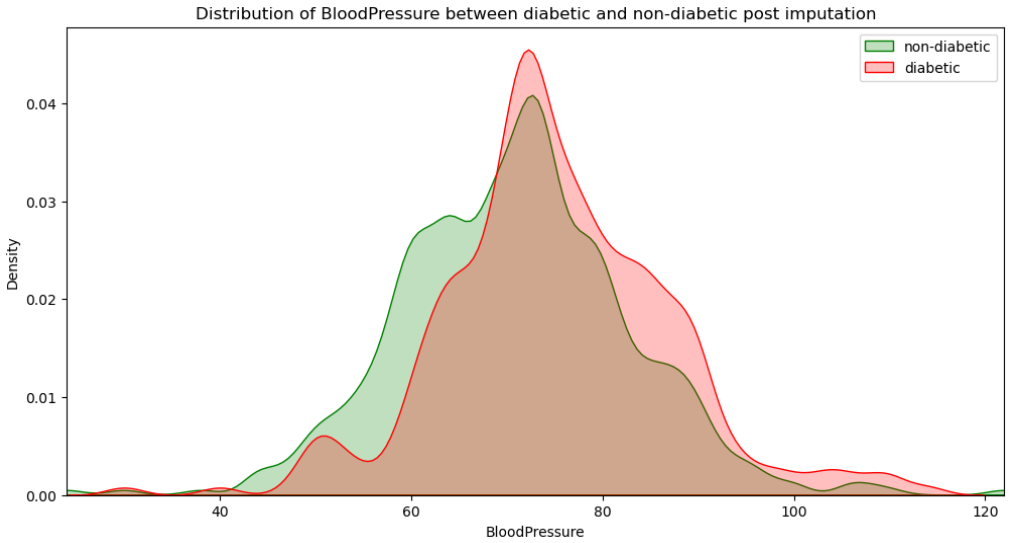
pregnancies and >= 7 pregnancies. The p-value for the same was 4.91e-191, which is less than 0.05

Bivariate analysis for Glucose split by Outcome



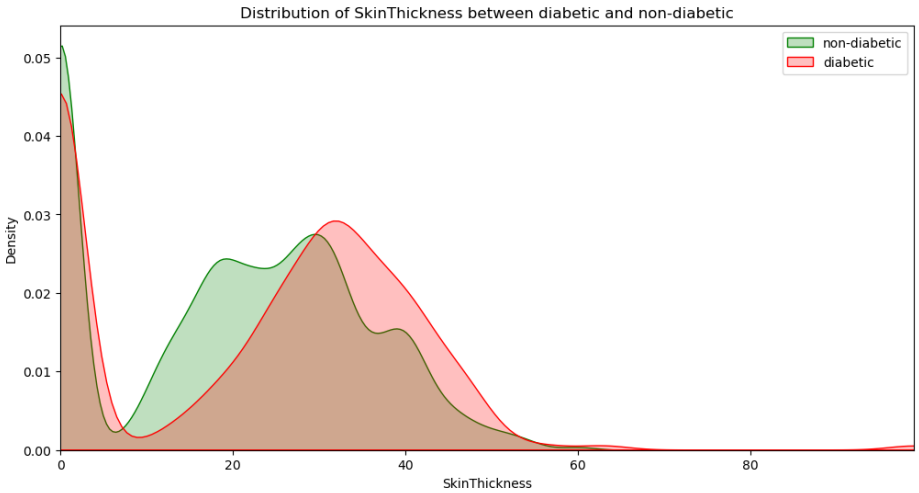
- The distribution of Glucose between Outcome 1 (diabetic) and 0 (non-diabetic) is shown in the above chart
- The distribution of glucose between diabetic and non-diabetic is significantly different (independent t-test between these 2 groups has a p-value of 3.13e-48, which is less than 0.05)
- Based on the shape of distribution seen above, we observed that the density of diabetic women was frequently ~3x more than non-diabetic beyond glucose levels of 120 i.e. **Pima women with >= 120 glucose level were 3.2 times more likely to be diabetic than otherwise**
- This conclusion is supported by the independent t-test between 2 groups - < 120 glucose and >= 120 glucose. The p-value for the same was 6.17e-176, which is less than 0.05.

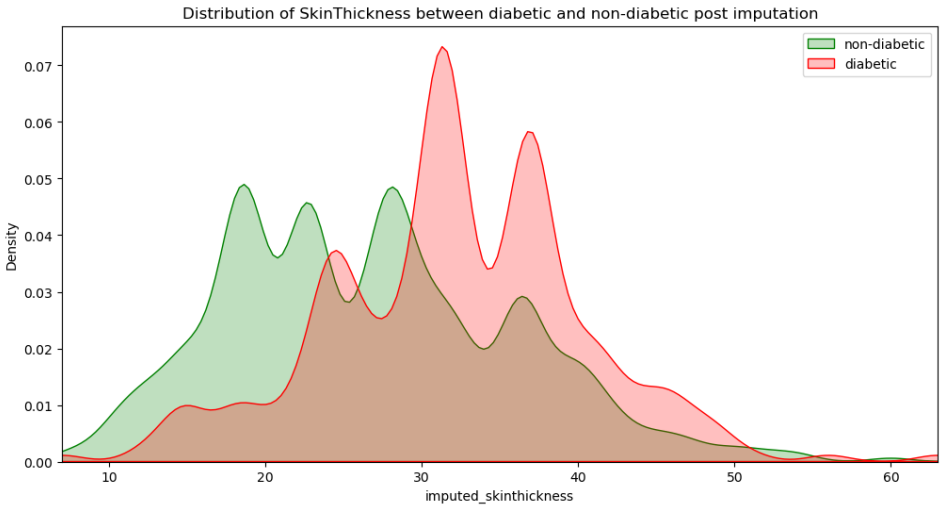
Bivariate analysis for Blood pressure split by Outcome



- The distribution of Blood Pressure between Outcome 1 (diabetic) and 0 (non-diabetic) is shown in the above chart
- The distribution of blood pressure between diabetic and non-diabetic is significantly different (independent t-test between these 2 groups has a p-value of 3.71e-06, which is less than 0.05)
- Based on the shape of distribution seen above, we observed that the density of diabetic women was frequently ~1.5x more than non-diabetic beyond blood pressure levels of 82 i.e. **Pima women with Blood Pressure >= 82 are 1.5 times more likely to be diabetic than otherwise**
- This conclusion is supported by the independent t-test between 2 groups - < 82 blood pressure and >= 82 blood pressure. The p-value for the same was 6.49e-121, which is less than 0.05

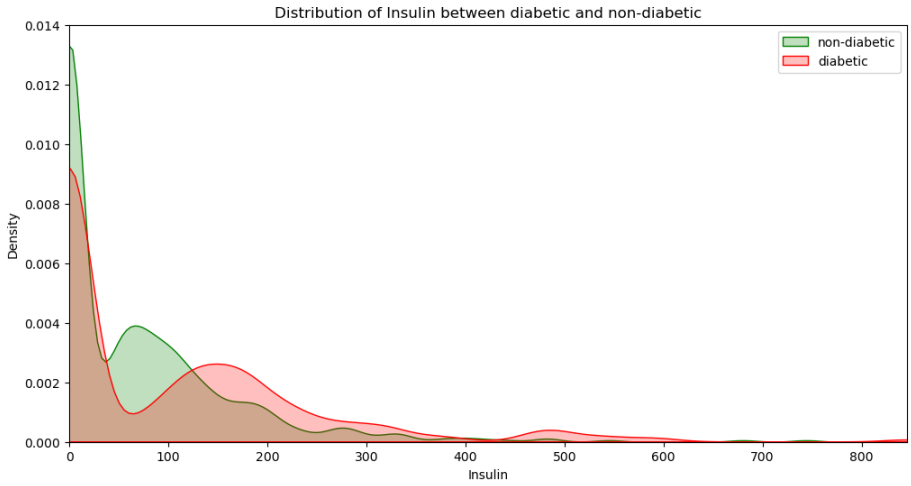
Bivariate analysis for Skin thickness split by Outcome

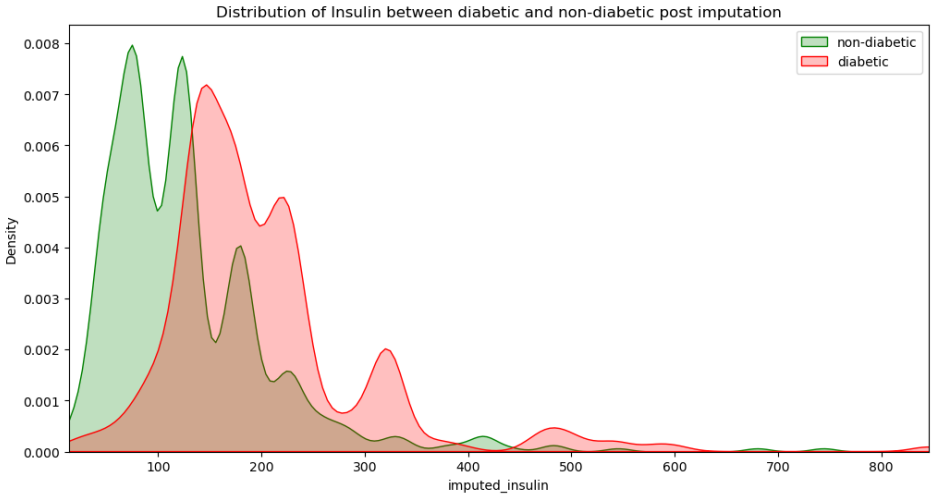




- The distribution of Skin Thickness between Outcome 1 (diabetic) and 0 (non-diabetic) is shown in the above charts for raw and imputed data respectively
- Skin Thickness post class (BMI and Outcome) mean imputation became a multimodal distribution for both diabetic and non-diabetic women. The distribution of Skin Thickness (post imputation) between diabetic and non-diabetic is significantly different (independent t-test between these 2 groups has a p-value of 1.51e-15, which is less than 0.05)
- Based on the shape of distribution seen above, we observed that the density of diabetic women was frequently ~2.5x more than non-diabetic beyond skin thickness levels of 30 i.e. **Pima women with Skin Thickness >= 30 are 2.5 times more likely to be diabetic than otherwise**
- This conclusion is supported by the independent t-test between 2 groups - < 30 skin thickness and >= 30 skin thickness. The p-value for the same was 4.61e-183, which is less than 0.05.

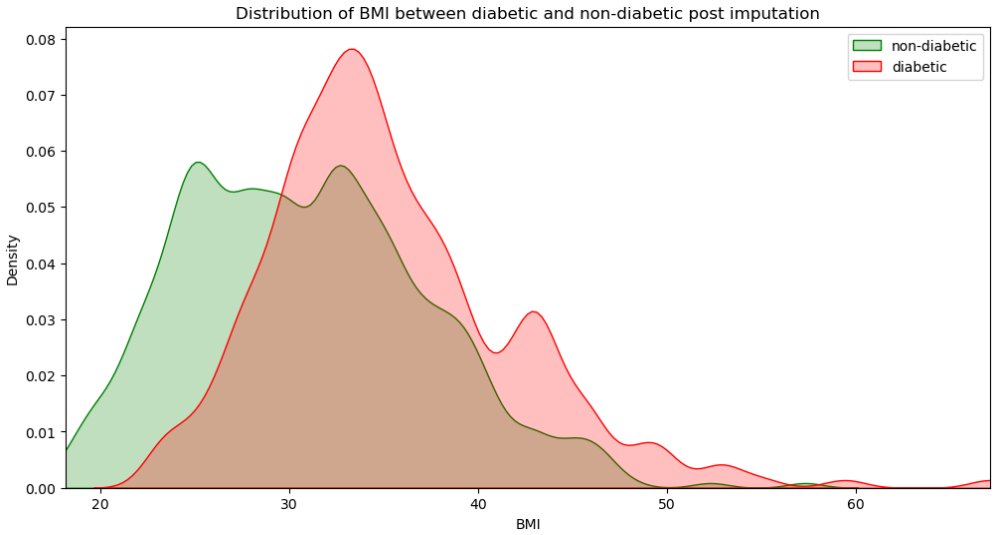
Bivariate analysis for Insulin split by Outcome





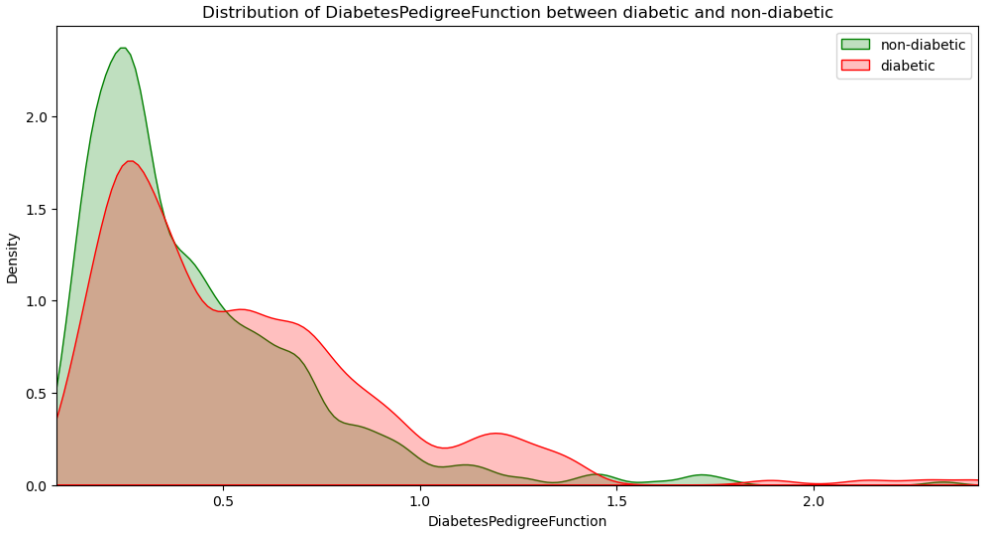
- The distribution of Insulin between Outcome 1 (diabetic) and 0 (non-diabetic) are shown in the above charts for raw and imputed data respectively
- Insulin post-class (Glucose and Outcome) mean imputation became a multimodal distribution for both diabetic and non-diabetic women. The distribution of Insulin (post imputation) between diabetic and non-diabetic is significantly different (the independent t-test between these 2 groups has a p-value of 2.73e-24, which is less than 0.05).
- Based on the shape of distribution seen above, we observed that the density of diabetic women was frequently ~2.5x more than non-diabetic beyond insulin levels of 150 i.e. **Pima women with Insulin >= 150 are 2.7 times more likely to be diabetic than otherwise**
- This conclusion is supported by the independent t-test between 2 groups - < 150 insulin and >= 150 insulin. The p-value for the same was 1.7e-122, which is less than 0.05.

Bivariate analysis for BMI split by Outcome



- The distribution of BMI between Outcome 1 (diabetic) and 0 (non-diabetic) is shown in the above chart
- The distribution of BMI between diabetic and non-diabetic is significantly different (independent t-test between these 2 groups has a p-value of 8.33e-19, which is less than 0.05)
- Based on the shape of distribution seen above, we observed that the density of diabetic women was frequently ~2.5x more than non-diabetic beyond a BMI of 30 i.e. **Pima women with BMI >= 30 are 2.7 times more likely to be diabetic than otherwise**
- This conclusion is supported by the independent t-test between 2 groups - < 30 BMI and >= 30 BMI. The p-value for the same was 3.33e-144, which is less than 0.05.

Bivariate analysis for Diabetes pedigree function split by Outcome

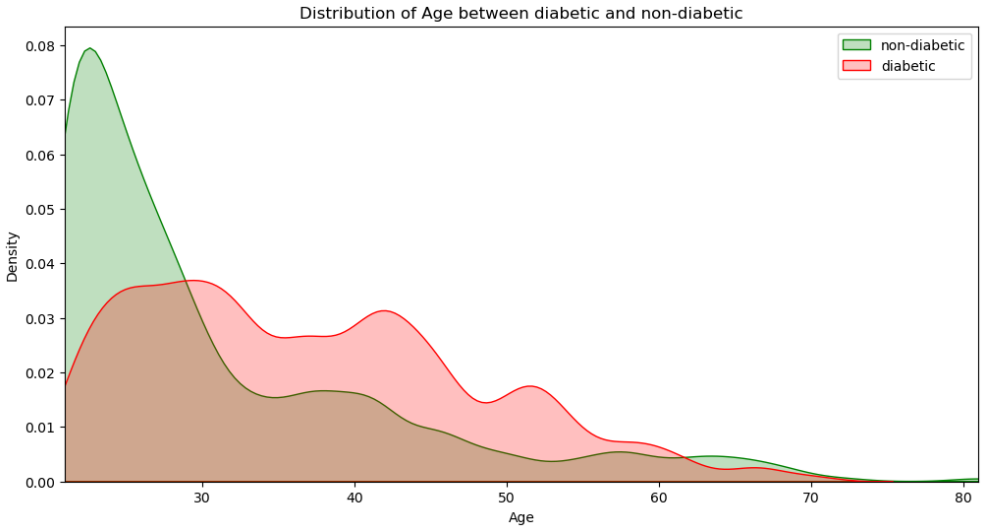


- The distribution of Diabetes Pedigree Function between Outcome 1 (diabetic) and 0 (non-

diabetic) is shown in the above chart

- The distribution of Diabetes pedigree function between diabetic and non-diabetic is significantly different (independent t-test between these 2 groups has a p-value of 1.25e-06, which is less than 0.05)
- Based on the shape of distribution seen above, we observed that the density of diabetic women was frequently ~1.5x more than non-diabetic beyond DPF of 0.5 i.e. **Pima women with Diabetes Pedigree Function >= 0.5 are 1.6 times more likely to be diabetic than otherwise**
- This conclusion is supported by the independent t-test between 2 groups - < 82 Diabetes pedigree function and >= 82 Diabetes pedigree function. The p-value for the same was 7.37e-155, which is less than 0.05.

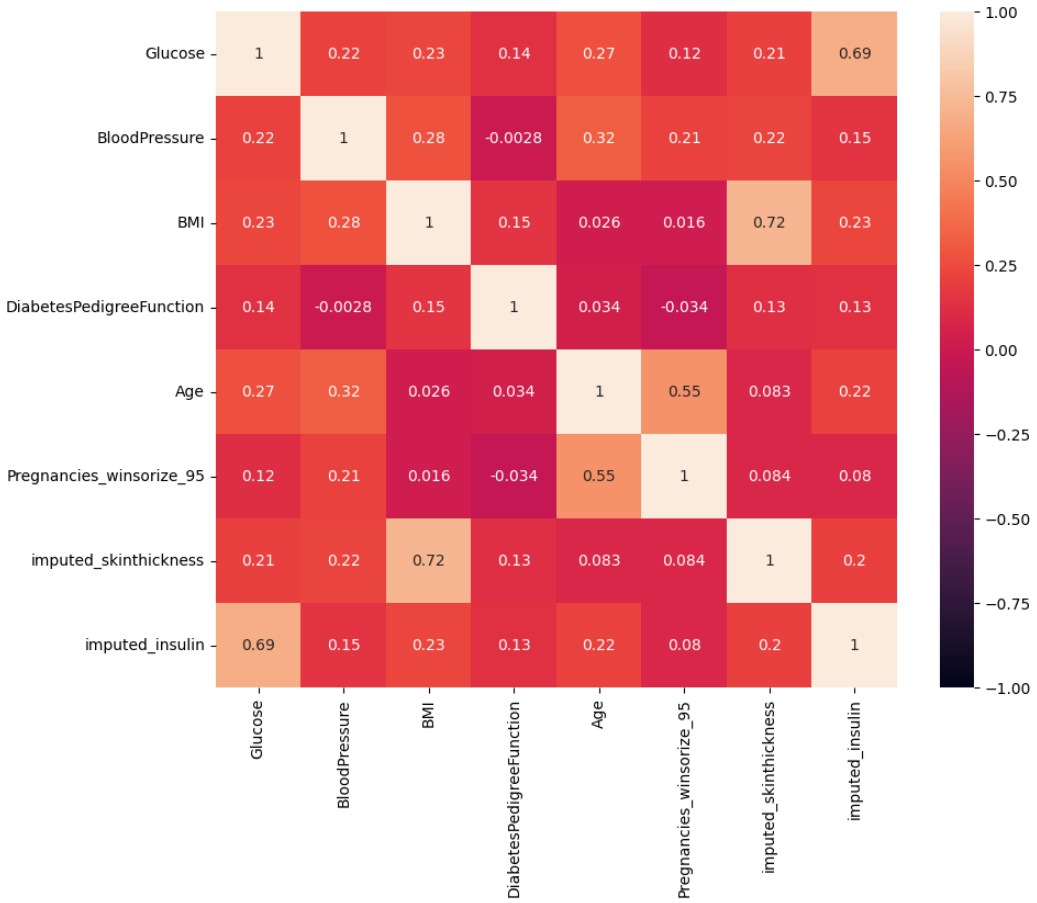
Bivariate analysis for Age split by Outcome



- The distribution of Age between Outcome 1 (diabetic) and 0 (non-diabetic) is shown in the above chart
- The distribution of Age between diabetic and non-diabetic is significantly different (independent t-test between these 2 groups has a p-value of 2.21e-11, which is less than 0.05)
- Based on the shape of distribution seen above, we observed that the density of diabetic women was frequently ~2x more than non-diabetic beyond age 30 i.e. **Pima women who are >= 30 years old are 2.3 times more likely to be diabetic than those who are younger**
- This conclusion is supported by the independent t-test between 2 groups - < 30 age and >= 30 age. The p-value for the same was 1.27e-160, which is less than 0.05.

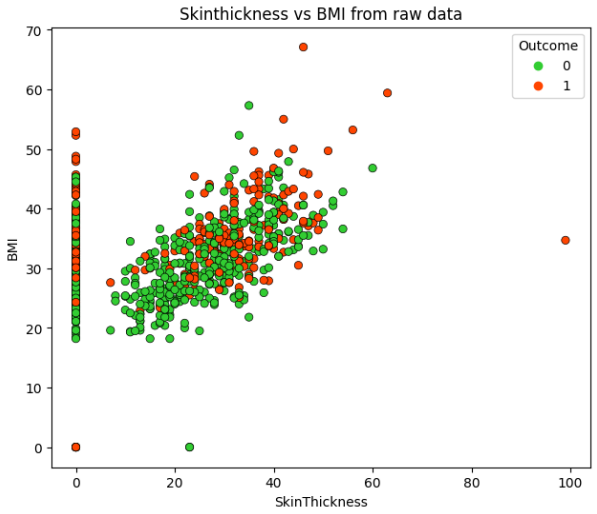
Relationship between all Independent variables – correlation heatmap:

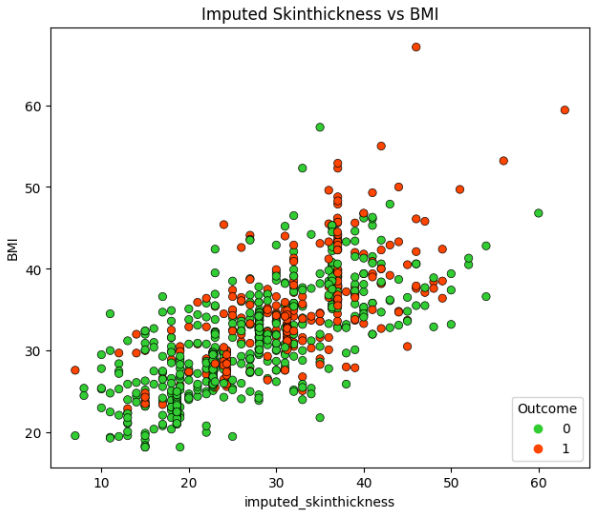
No major correlation was observed between any of the features (cut-off for correlation = 0.85)



- The correlation heatmap post data treatment (winsorization & imputation) is shown in the above chart
- No major correlation was observed between any of the independent variables
- A somewhat “medium” correlation of 0.72 was observed between the imputed SkinThickness and BMI, this is due to the noticeable linear relationship shown between these variables and is explained later in the doc
- A correlation of 0.69 was observed between the imputed Insulin and Glucose, this is likely due to the magnitude of imputations in Insulin which was based on Glucose

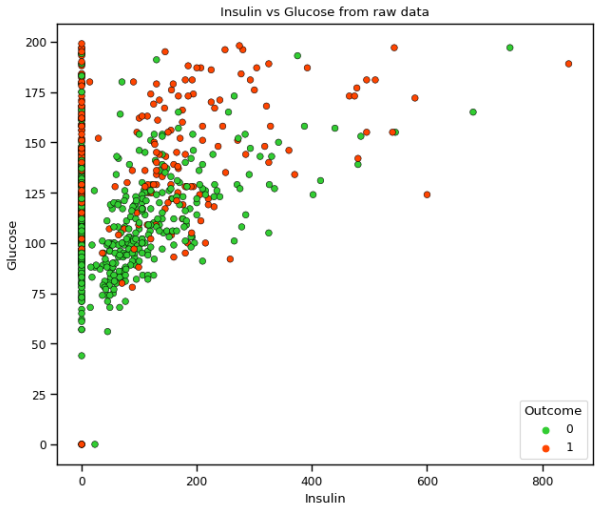
Bivariate analysis for specific independent variables: BMI vs Skin thickness

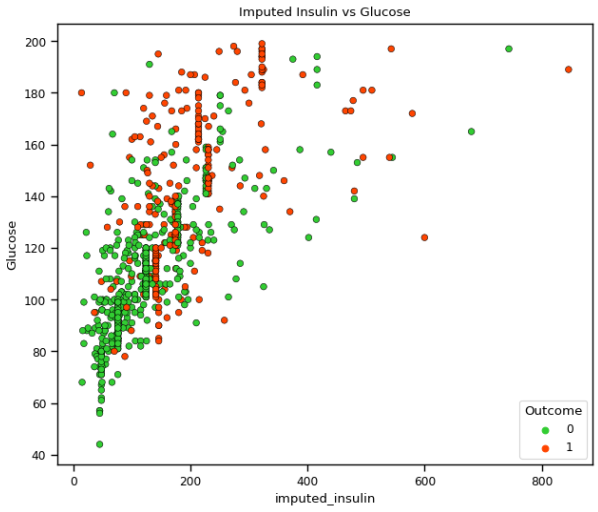




- From the raw data, a positive linear relationship between BMI and SkinThickness is visually noticeable in the raw data chart if missing values are disregarded
- This was the reason why BMI was considered as the class for class mean imputation for SkinThickness
- Post imputation, we can see a more meaningful relationship between the imputed features which shows a correlation of 0.72
    - Specifically, the concentration of 0s is noticeable in the first chart but these imputed values are spread out in the second chart and blend in with the remaining data points.

Bivariate analysis for specific independent variables: Glucose vs Insulin





- As seen in the raw data scatter plot, a significant volume of data (49%) is concentrated in 0 for Insulin
- Post imputation, we're able to redistribute the data through class mean imputation (with Glucose and Outcome combination being the class) while largely maintaining the overall trend of remaining data points
- Hence, by adopting class mean imputation, we're able to retain the directional trend of the data and can salvage this column without any major distribution changes

Summary of Exploratory Data Analysis

- The overall takeaway based on our exploratory analysis is that there is evidence to suggest that older Pima women with higher Insulin and glucose levels and considerably overweight are at greater risk of being diabetic
- Pima women who are 30 years or older are 2.3x more likely to be diabetic than otherwise
- Insulin post imputation became a multimodal distribution for both diabetic and non-diabetic Pima women, with the difference in Insulin >= 150 being 2.7x more likely being diabetic than lower readings
- Pima women with >= 120 glucose levels were 3x more likely to be diabetic than otherwise
- Women with Skin Thickness >= 30 are 2.5x more likely to be diabetic than otherwise
- Pima women with BMI >= 30 are 2.7x more likely to be diabetic than otherwise
- Pima women with >= 7 pregnancies are 2x more likely to be diabetic than those with lower
- Pima women with Blood Pressure >= 82 are 1.5x more likely to be diabetic than those otherwise
- Pima women with Diabetes Pedigree Function >= 0.5 are 1.6 times more likely to be diabetic than otherwise

Predictive Modeling

Approach

Our goal is to predict if a Pima woman is diabetic or not based on the provided features. Hence, we experimented with 4 classification algorithms to get a range of models that we can choose from based on the criteria of model performance like accuracy, precision, recall, explainability, etc. We chose to experiment with 4 tree-based models since these are intuitive to understand, perform well with different types of data and certain algorithms are less prone to data issues like missing values, outliers, etc. For our project, we used the following 4 algorithms:

- Decision trees
- Random Forest
- Gradient Boosted Machine
- Extreme Gradient Boosting or XGBoost

Modeling steps and evaluation metrics

The steps followed in the modeling are largely the same across all 4 models:

- 70% of the treated dataset is randomly split for training while 30% is kept aside for testing
- Models are trained on the training dataset and predicted on the test set, after which the actual vs predicted on the testing dataset is done to measure the accuracy of prediction
- Hyperparameter tuning is performed to optimize the model performance
- Another iteration of the modeling is done with the optimal parameters using random search (RandomizedSearchCV from sklearn) – this is the final model
- The following metrics are calculated to measure the model performance:
    - Confusion matrix: A confusion matrix represents the prediction summary in matrix form. It shows how many predictions are correct and incorrect per class. It helps in understanding the classes that are being confused by the model as other classes.
    - Accuracy: ratio of the number of correct predictions to the total number of input samples

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{True\ Positives\ (TP) + False\ Positives\ (FP) + True\ Negatives\ (TN) + False\ Negatives\ (FN)}$$

    - Precision: Precision is the ratio between the True Positives and all the Positives, the formula being:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

    - Recall: The recall is the measure of our model correctly identifying True Positives, the formula being:

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

    - F1 score: The F1 score is calculated as the harmonic mean of the precision and recall scores. A large F1 score of 1 indicates excellent precision and recall, while a low score indicates poor model performance.
    - ROC curve:
        - A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds
        - This curve plots two parameters: True Positive Rate (TPR) and False Positive Rate (FPR)
        - The ROC curve essentially separates the 'signal' from the 'noise' and shows the performance of a classification model at all classification thresholds
        - The Area Under the Curve (AUC) is the measure of the ability of a binary classifier to distinguish between classes and is used as a summary of the ROC curve
        - The higher the AUC, the better the model's performance at distinguishing

between the 1 and 0

- Feature importance:
  - One of the advantages of tree-based models is the ability to understand the relative importance of each independent feature used in the model
  - Variable importance is a way of measuring how each variable contributes to the overall performance of the model. For single decision trees, the variable “higher up” in the tree has the greater influence on predicting the target variable

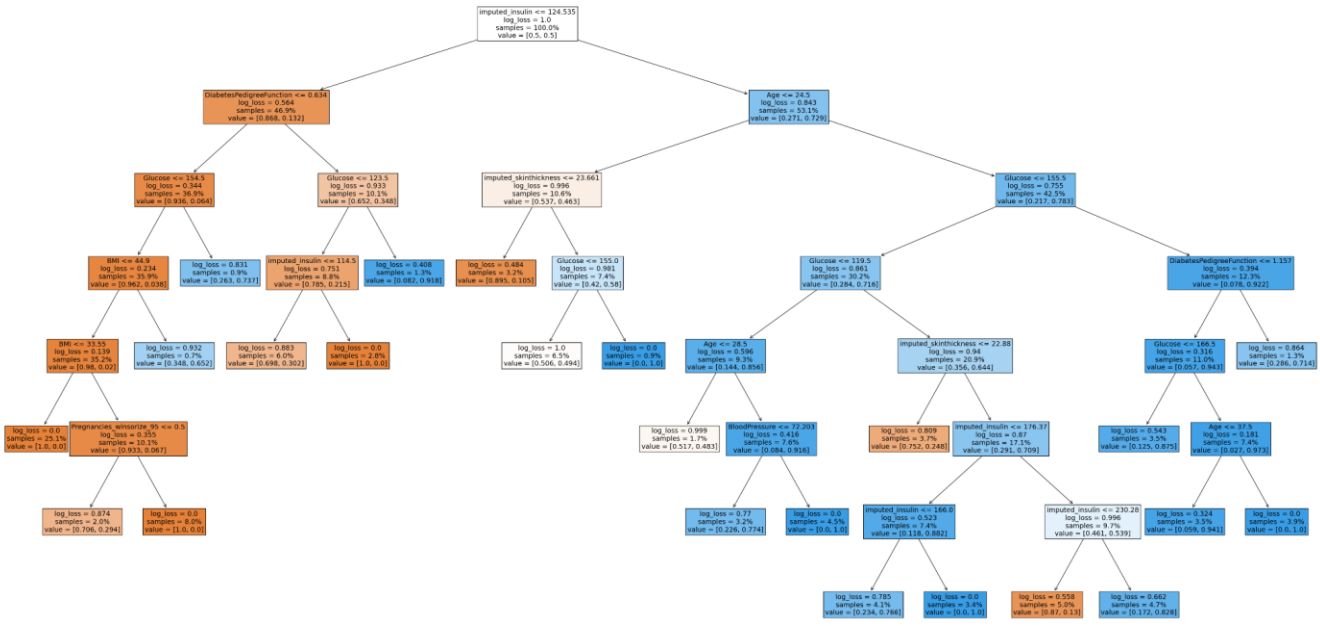
Decision trees

Decision tree classifiers use a tree-like model to make decisions based on features, dividing data into subsets at each node to reach a final classification.

Each node represents a decision made on one of the independent features. Depending on the decision made at the node, the dataset is divided into two or more subsets.

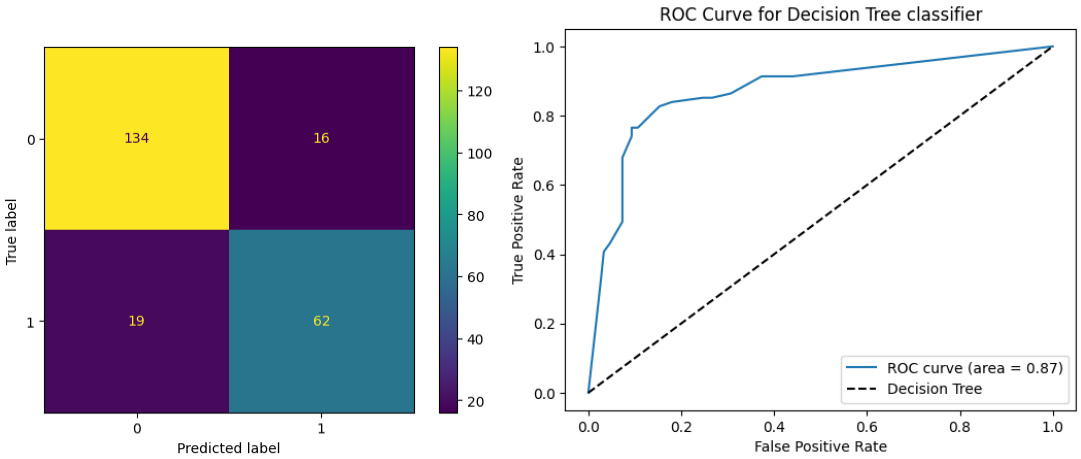
The splitting decision is guided by the heuristic that the resulting splits will increase the amount of information that can be obtained from the data and reduce the uncertainty in predicting the outcome.

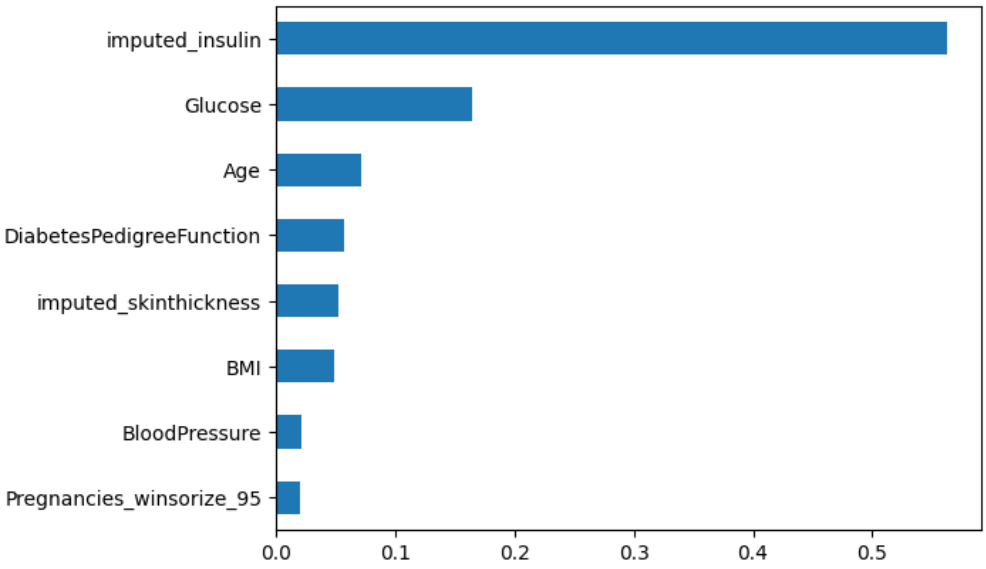
We keep splitting the data in each branch until we reach a pure/ leaf node which is a node where we cannot split the data any further as there is no information gain and no uncertainty in the prediction of outcome.



The above figure shows the complete decision tree that was obtained after hyperparameter tuning of the DecisionTreeClassifier.

Statistical descriptors or our model:

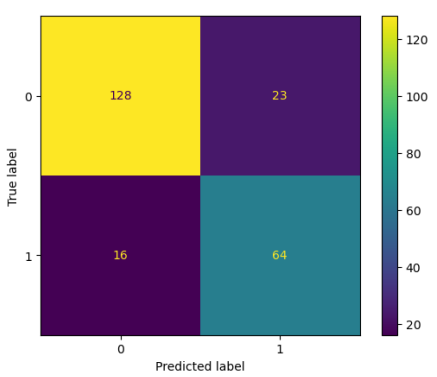


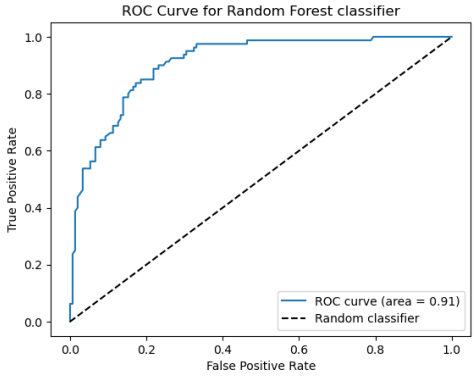


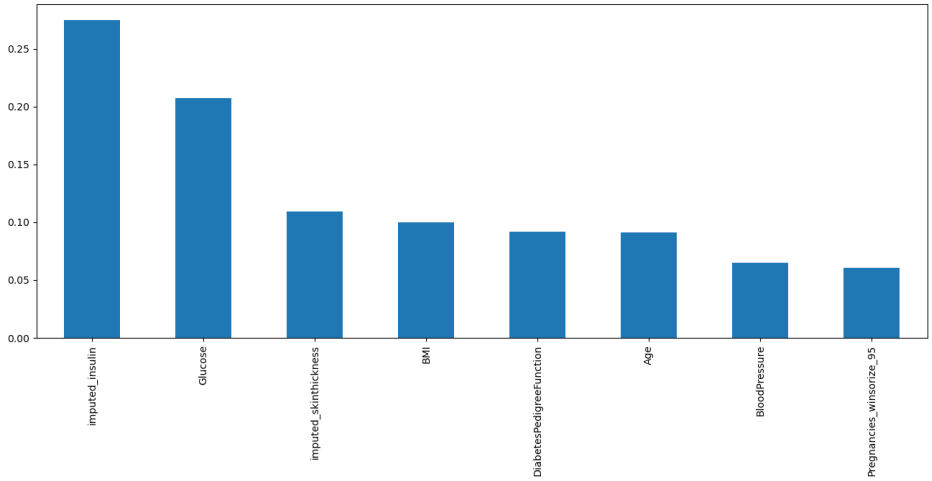
Random Forest

- Random Forest is an ensemble of decision trees, which uses the bagging method. Bagging is a combination of bootstrapping and aggregation algorithms
- Random forest models improve the generalization of the the model by aggregating the results across a collection of decision trees, thereby reducing the possibility of over-fitting
- Hyperparameter tuning was performed to optimize the model performance leading to an increase in accuracy by ~1 percentage point.
    - First iteration accuracy = 82.7%
    - Accuracy after tuning = 83.1%
- Area under curve = 0.91 which is quite good

- Insulin (imputed feature) was the most important variable of the features used for building the model.

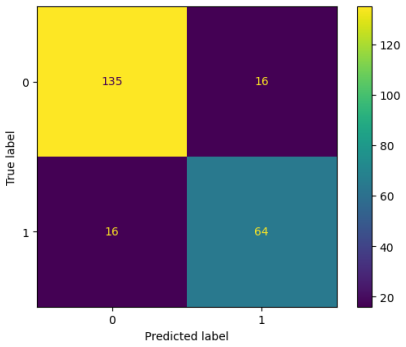


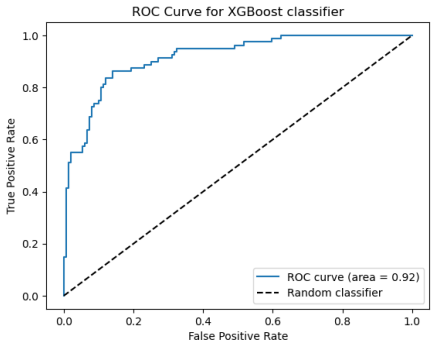


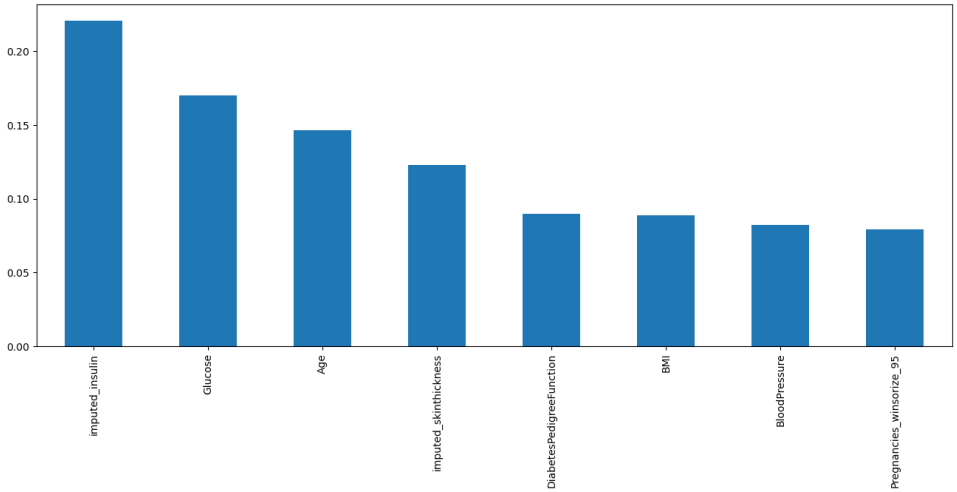


XGBoost

- XGBoost (full form: Extreme Gradient Boosting) is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning algorithm
- It provides parallel tree boosting and is the leading machine learning method for regression, classification, and ranking problems
- XGBoost is different from the Random Forest since it uses Boosting as a way to improve the predictive power of weak learner trees in series but the trade-off is that it is slightly more computationally expensive and complex
- Hyperparameter tuning was performed to optimize the model performance leading to an increase in accuracy by ~4 percentage points.
    - First iteration accuracy = 82.7%
    - Accuracy after tuning = 86.1%
- The area under the curve = 0.92 which is quite good
- Insulin (imputed feature) was the most important variable of the features used for building the model







Gradient Boosted Machine

- The Gradient Boosting Machine utilizes decision trees as its machine learning model
- This technique involves adding models to the ensemble one after another, with each subsequent model improving upon the performance of the previous ones
    - Boosting algorithms work by initially constructing a model based on the training dataset, followed by building a second model to correct any errors made by the first one
- In the gradient boosting process, first, we create a base model that predicts the observations in the training dataset. Then we calculate the pseudo residuals, which are obtained by subtracting the predicted values from the observed values
- Later, a model is built using these pseudo residuals to make predictions
- Then, the output values for each leaf of the decision tree are determined
- A leaf can have more than one residual, so the final output for all the leaves needs to be determined. Finally, the predictions of the previous model are updated.
- Hyperparameter tuning was performed to optimize the model performance leading to an increase in accuracy by ~1 percentage point
    - First iteration accuracy = 87.0%
    - Accuracy after tuning = 87.88%

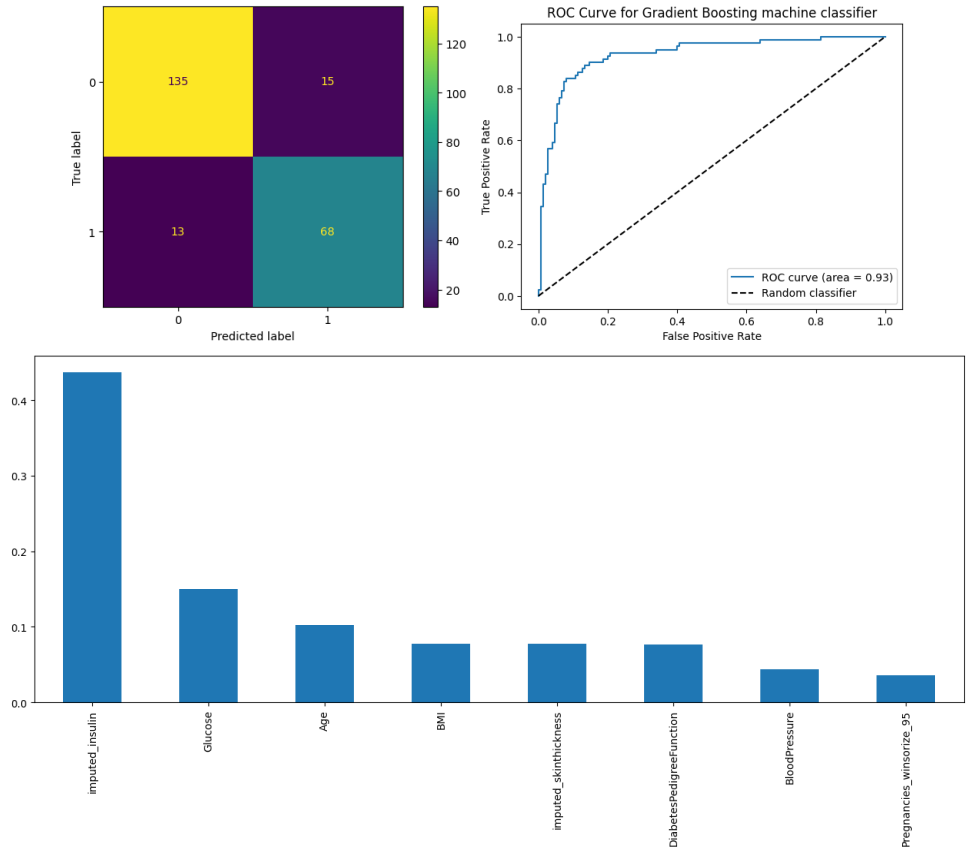
- The area under the curve = 0.93 which is quite good
- Insulin (imputed feature) was the most important variable of the features used for building the model, followed by Glucose and Age as highlighted in the variable importance plot

Consolidated summary of modeling approaches

Gradient Boosted Machine (GBM) gave us the best overall accuracy, precision, and recall among the 4 approaches attempted due to the boosting approach of improving the predictive power of each tree in the ensemble model. XGBoost was a close second.
With further tuning of the hyperparameters, it is possible to improve the model performance but there may be a risk of overfitting; in our approach, we have prioritized model explainability over performance.

Modeling summary and key takeaways

- **We were able to achieve a peak accuracy of 88% on the GBM model post hyperparameter tuning with Insulin, Glucose, and Age being the most important variables**
- **Insulin was found to be the most important variable in predicting the Outcome** i.e. diabetic or not.
    - Insulin was followed by Glucose and Age in the variable importance plot
- Insulin being the most important variable makes intuitive sense since a higher Insulin reading is most commonly associated with being diabetic, Glucose reading the 2nd most important variable is also intuitive
- By using the variable treatment approach and the model as illustrated in our approach, we can predict with a fairly high degree of accuracy whether any new Pima Indian woman is at risk of diabetes or not
- Shapley values are useful in determining the model explainability - they provide the importance of the variables along with the direction of influence.
    - This might aid medical professionals in early medical intervention to manage type 2 diabetes.
    - In particular, we can recommend that **older Pima women with high Insulin and Glucose readings are at higher risk of type 2 diabetes and hence, should be prioritized for preventive care**



Approach without Outcome being considered in class mean imputations

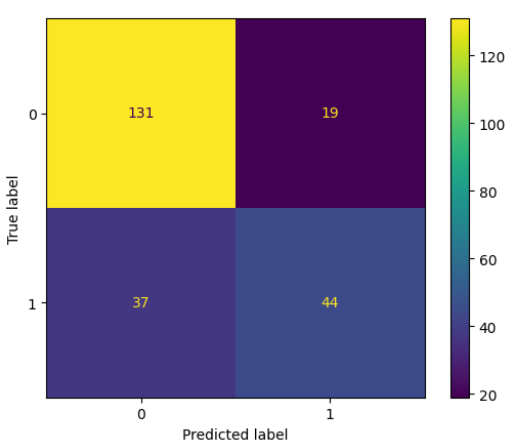
Gradient Boosted Machine

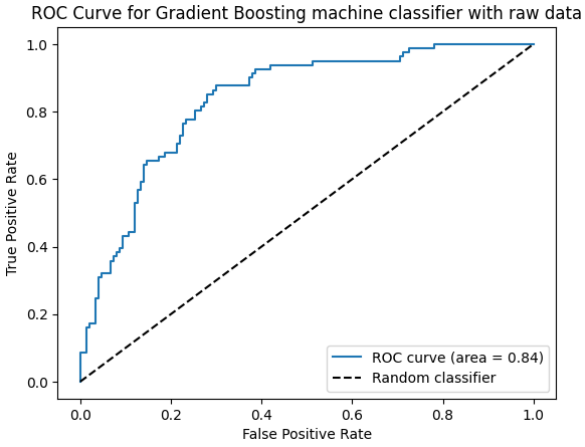
Based on the guidance provided during the group presentation on 5th December, 2023, we have run 2 iterations of Gradient Boosted Machine (considered since it was our champion model).

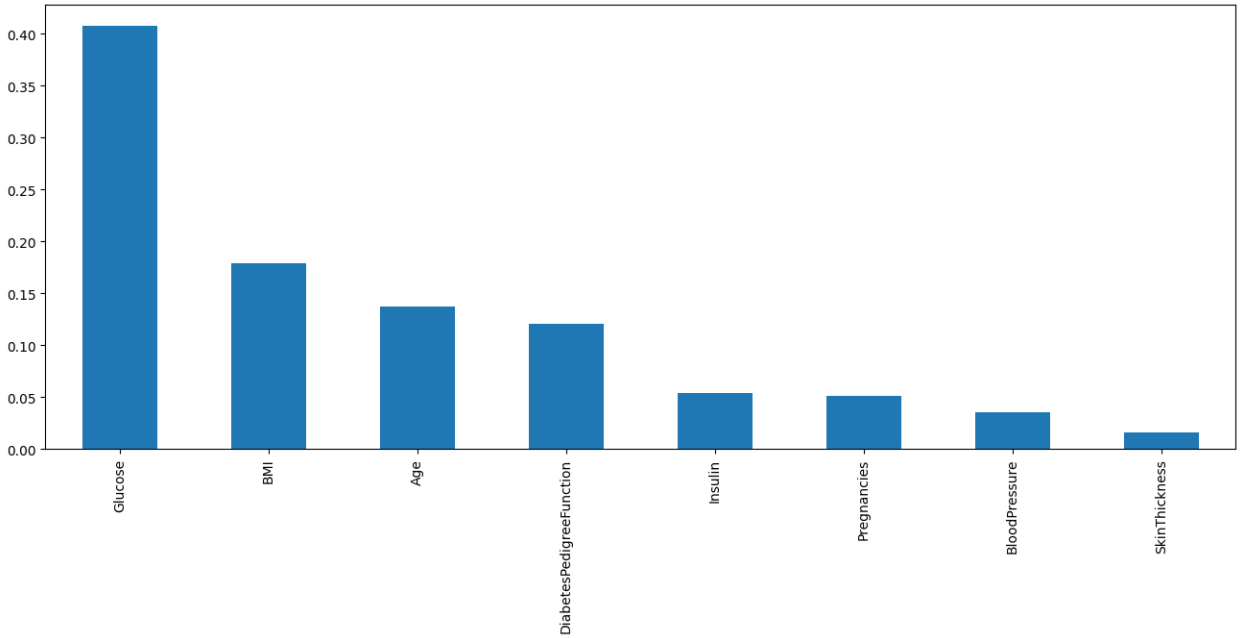
- Baseline model with features from raw data
- Model without taking Outcome for class mean imputations for Skin Thickness and Insulin features

Following are the results of the modeling done with each approach:

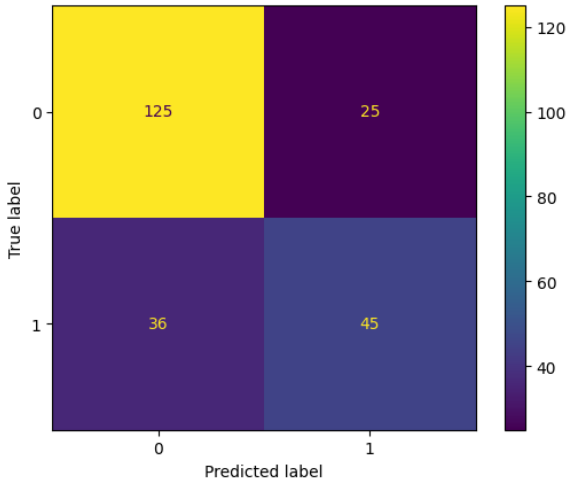
Baseline model

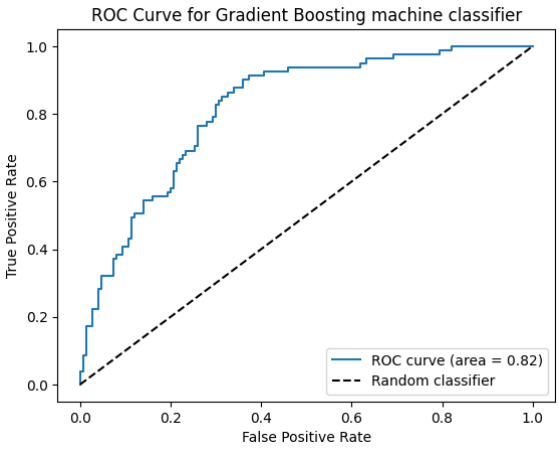


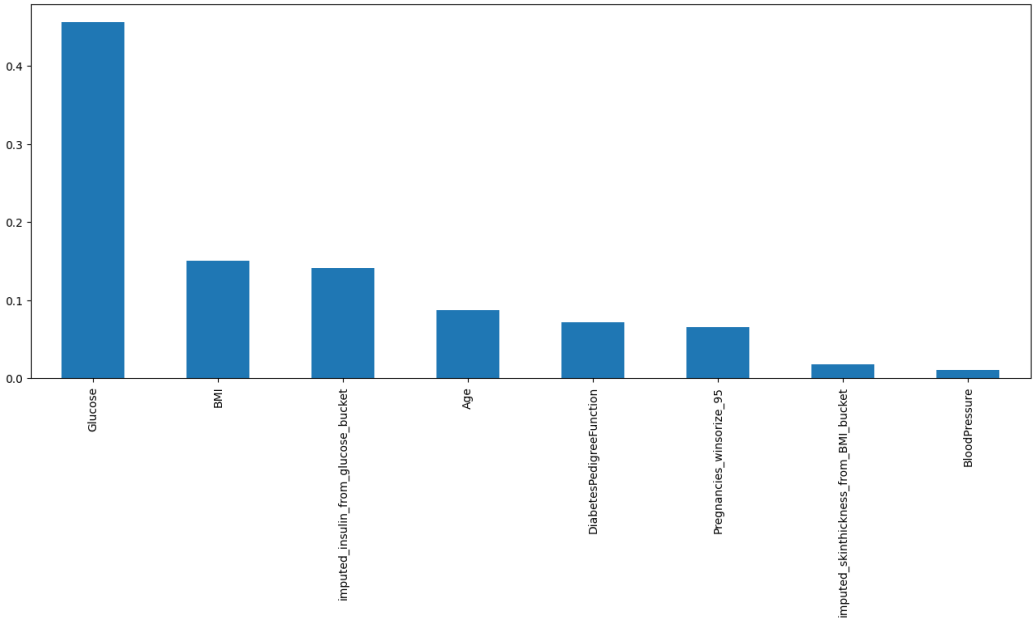




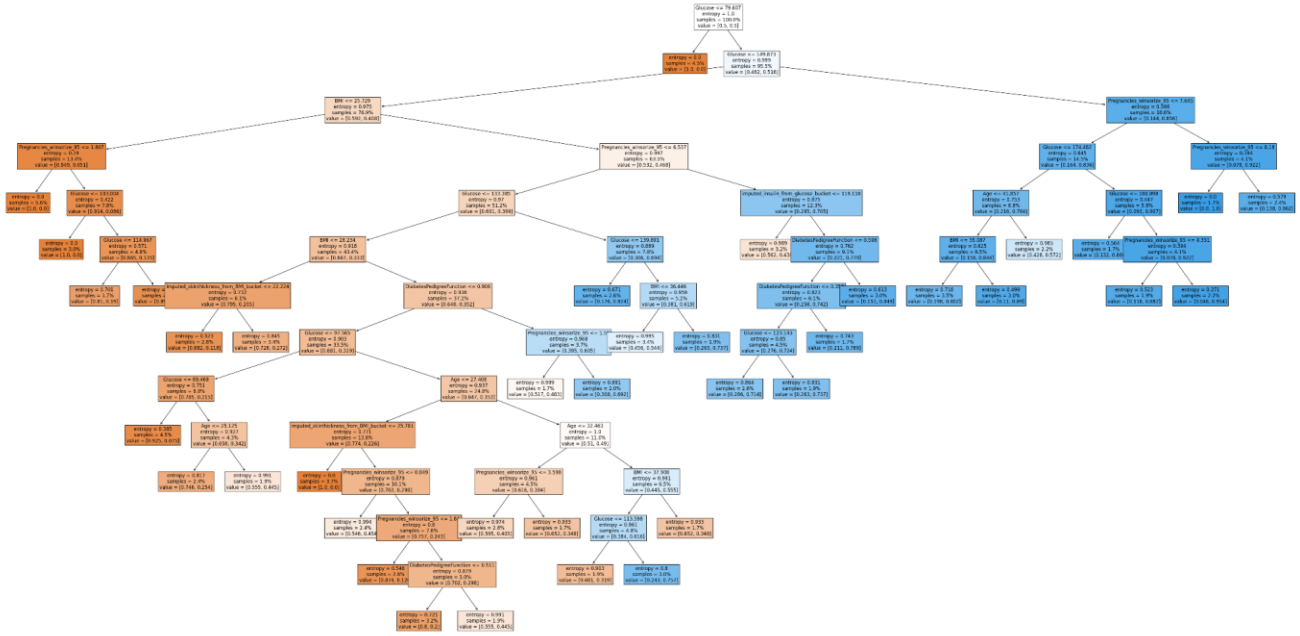
Model with imputed data that doesn't take Outcome as class

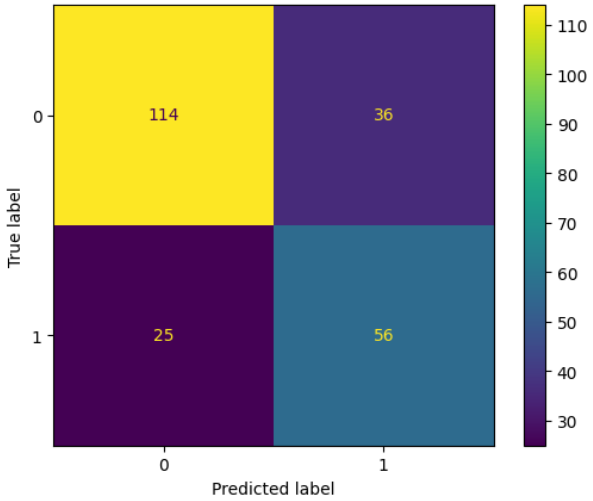


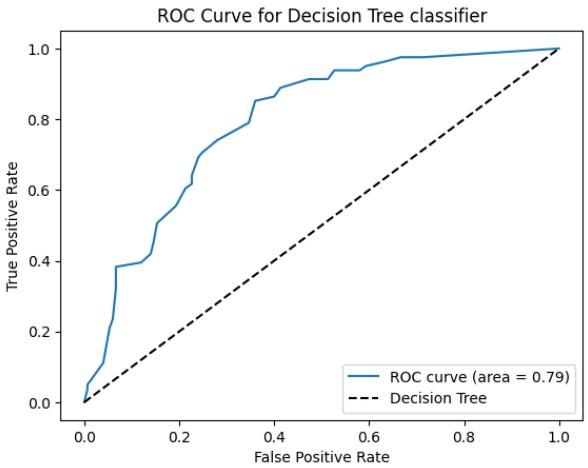


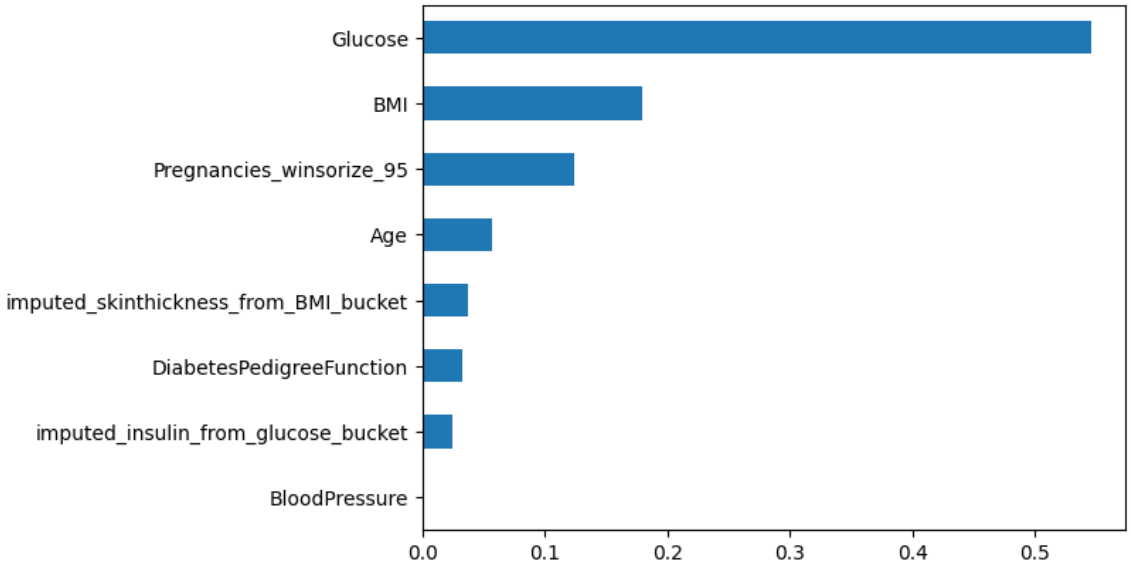


Decision Tree:

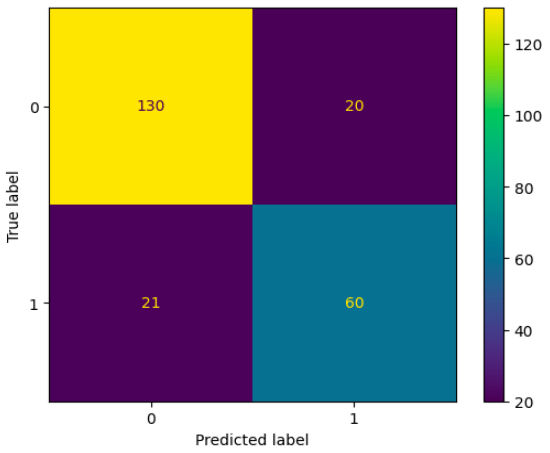


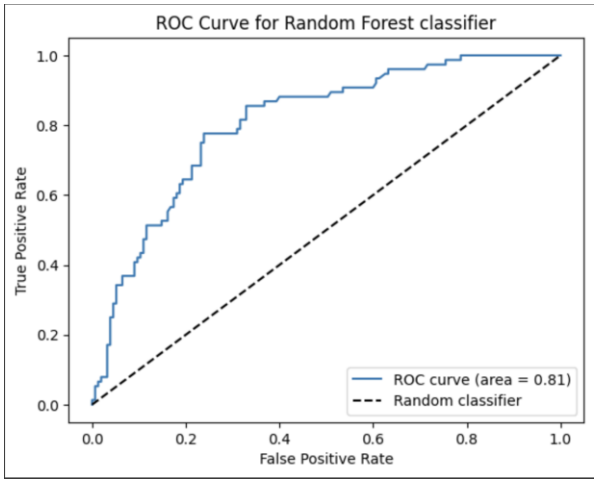


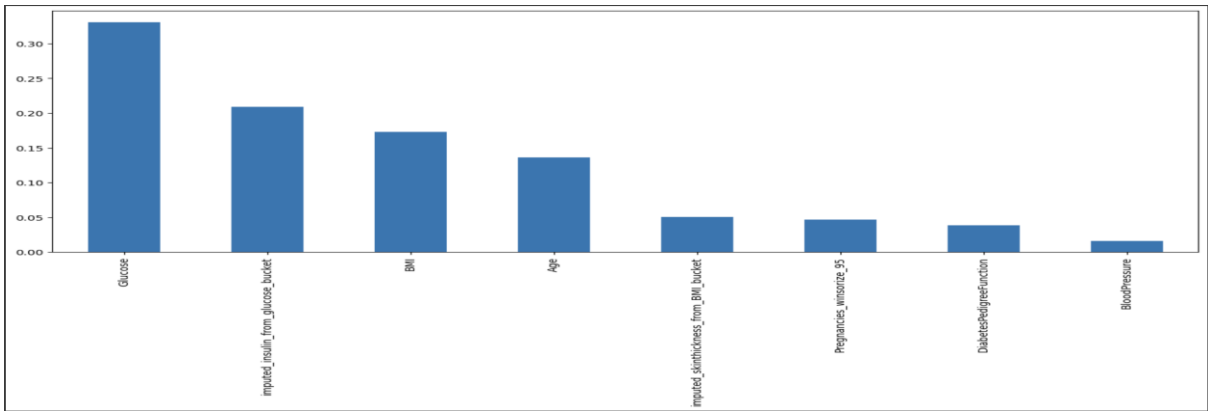




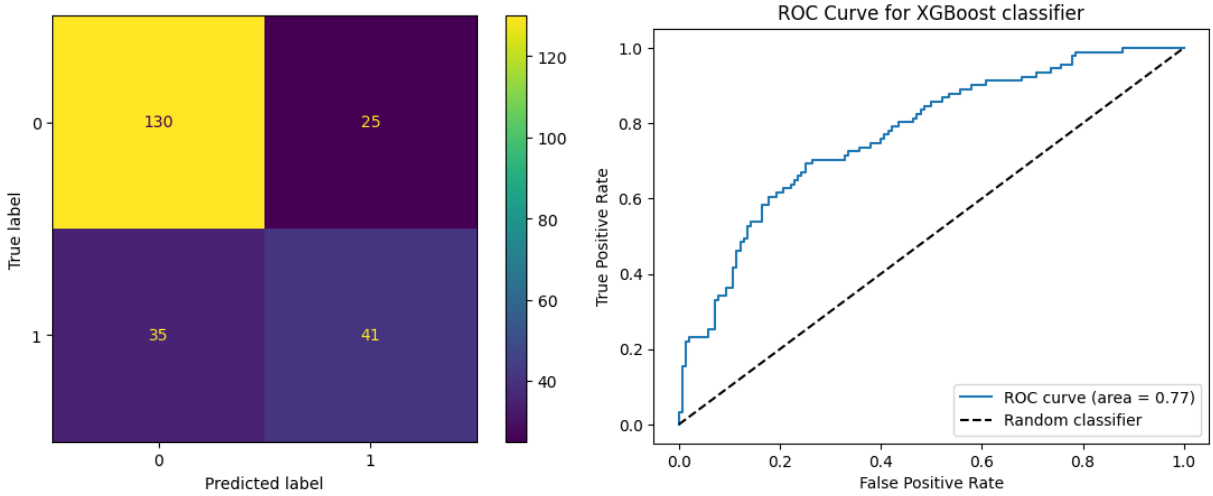
Random Forest:

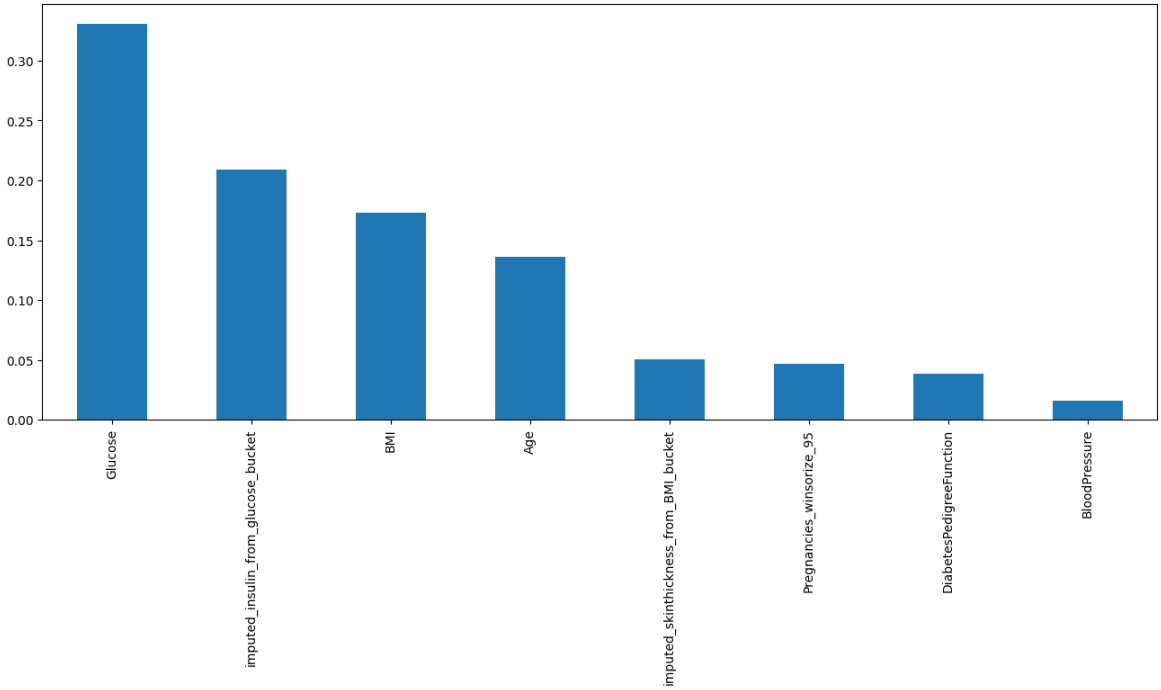






XGBOOST:


130
25
35
41
0
1
True label
Predicted label
120
100
80
60
40
ROC Curve for XGBoost classifier
True Positive Rate
False Positive Rate
ROC curve (area = 0.77)
Random classifier
0.0
0.2
0.4
0.6
0.8
1.0


0.30
0.25
0.20
0.15
0.10
0.05
0.00
Glucose
imputed_insulin_from_glucose_bucket
BMI
Age
imputed_skinthickness_from_BMI_bucket
Pregnancies_winsorize_95
DiabetesPedigreeFunction
BloodPressure

Metric	Decision tree	Random Forest	XGBoost	Gradient Boosted Machine
Accuracy	73.59%	74.03%	73.16%	73.59%
Precision	60.86%	62.12%	71.64%	64.29%
Recall	69.13%	53.95%	52.75%	55.56%
AUC	0.79	0.80	0.77	0.82
F1 score	0.64	0.58	0.61	0.6

Summary and takeaways

Metric	GBM without Outcome considered in imputation	GBM with raw data (baseline model)
Accuracy	73.59%	75.76%
Precision	64.29%	69.84%
Recall	55.56%	54.32%
AUC	0.82	0.84
F1 score	0.6	0.61

- We see that the accuracy of the baseline model is slightly higher and this is likely due to the influence of the 0s in Insulin and Skin Thickness
- Not much change is observed in the Recall and AUC though
- A considerable drop in accuracy is observed as compared to our original approach
- In both cases, Glucose and BMI are the top 2 important variables

Tableau dashboard link

Tableau public server dashboard link is:

https://public.tableau.com/app/profile/sakshi.jain1349/viz/Pime-daibetes/Dashboard1?publish=yes

References

- High-Risk Populations: The Pimas of Arizona and Mexico: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4418458/#:~:text=By%201970%2C%20the%20prevalence%20of,similarly%20aged%20Caucasians%20%5B5%5D
- Reducing Diabetes in Indian Country: Lessons from the Three Domains Influencing Pima Diabetes: https://www.smu.edu/~/media/Site/Dedman/Departments/Anthropology/pdf/Smith-Morris/HO%20Virchow.ashx
- PIMA Indian Diabetes dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- Effect of pregnancies on diabetes: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124395/
- Relationship between Insulin and Glucose: https://pubmed.ncbi.nlm.nih.gov/8422798/
- Relationship between Insulin and Glucose: https://pubmed.ncbi.nlm.nih.gov/8898771/
- Women have a higher baseline TSF and hence upper limit of 30 can be considered [study: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9127233/#:~:text=In%20the%20study%20population%2C%20the,vs%2014.3%20%C2%B1%206.8%20mm).]
- Insulin and age study for Pima women: https://www.foodandnutritionjournal.org/volume7number2/significance-of-health-related-predictors-of-diabetes-in-pima-indians-women/
- BMI vs diabetes: https://pubmed.ncbi.nlm.nih.gov/2031485/