Data Science Capstone

Health Care Project

- Project Task: Week 1

a) Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make sense and thus indicates missing value:

- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI

```
print(f'unique featues: {len(counts[0])}')
print(f'binary featues: {len(counts[1])}')
print(f'categorical/Numerical featues: {len(counts[2])}')
```

```
unique featues: 0
binary featues: 0
categorical/Numerical featues: 8
```

```
print("{0:34} {1:20} {2:20}".format('Column Name','Null Value count','Zero Value count'))
for col in x.columns:
    null = x[col].isna().sum()
    zero = sum(x[col].values ==0)
    print("{0:24} {1:20} {2:20}".format(col,null,zero))
```

```
Column Name                      Null Value count     Zero Value count
Pregnancies                                 0                  111
Glucose                                     0                    5
BloodPressure                               0                   35
SkinThickness                               0                  227
Insulin                                     0                  374
BMI                                         0                   11
DiabetesPedigreeFunction                    0                    0
Age                                         0                    0
```

- There is no null values in any columns.
- Only zeros values are found.
- A value of zero does not make sense and thus indicates missing value: Glucose, BloodPressure, SkinThickness, Insulin, BMI.

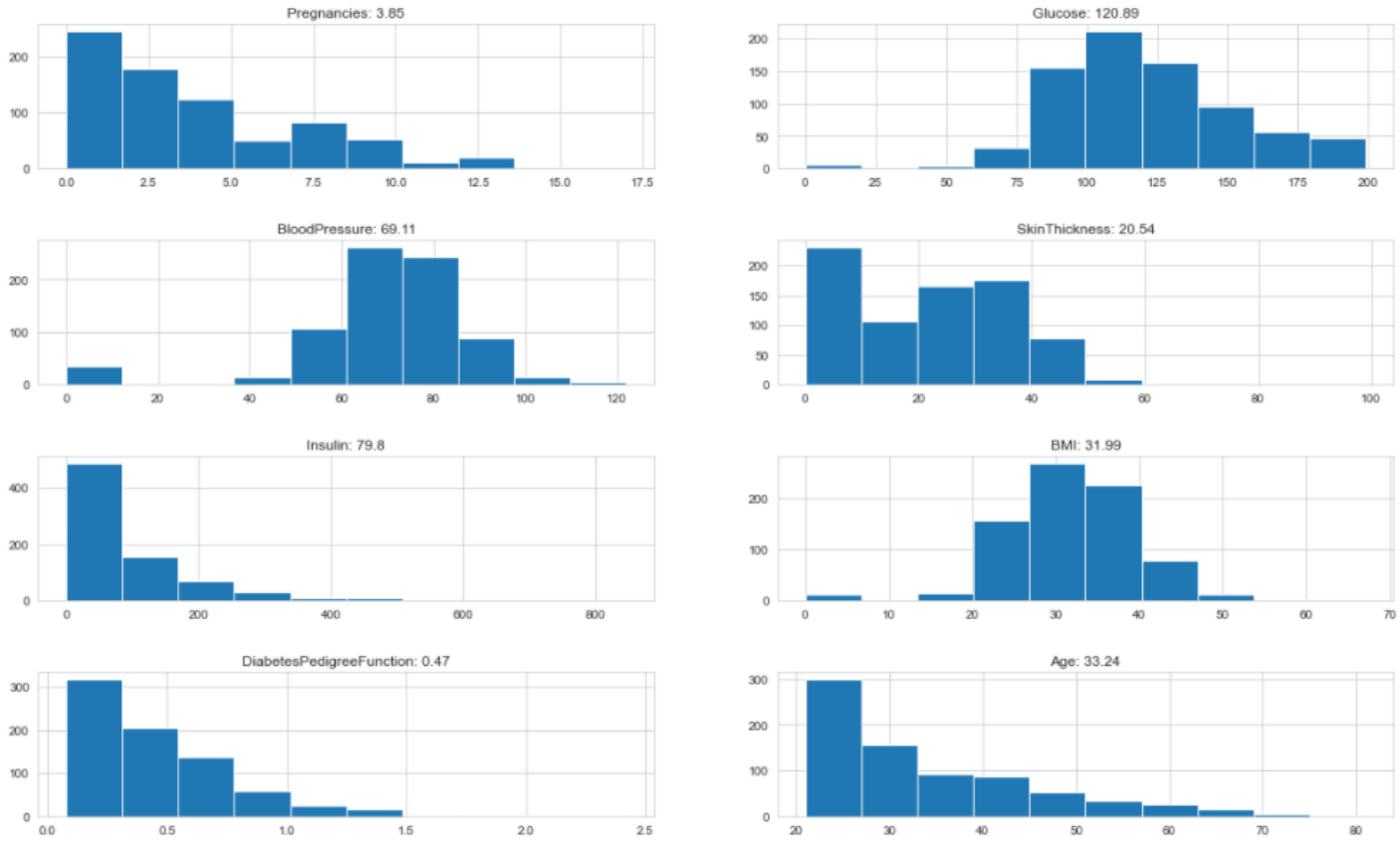
Zero values treatment for the below columns:

- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI

Fill column mean at the place of Zero

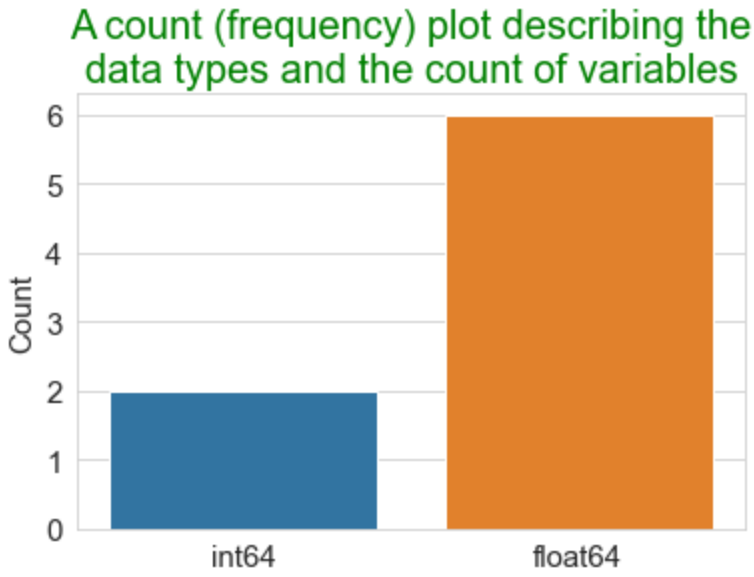
```
columns = ['Glucose','BloodPressure','SkinThickness','Insulin','BMI']
# missing value (zero value) treatment
for col in columns:
    for i in range(len(x[col])):
        if x.loc[i,col] == 0:
            x.loc[i,col]=round(x[col].mean(),2)
```

b) Visually explore these variables using histograms. Treat the missing values accordingly.



There is no missing values.

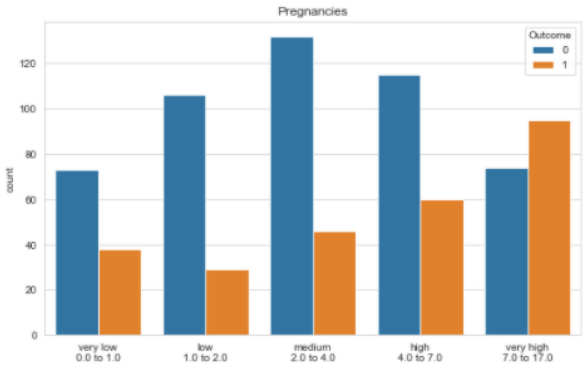
c) There are integer and float data type variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

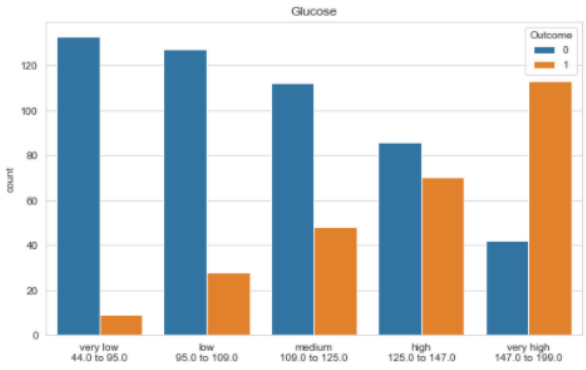


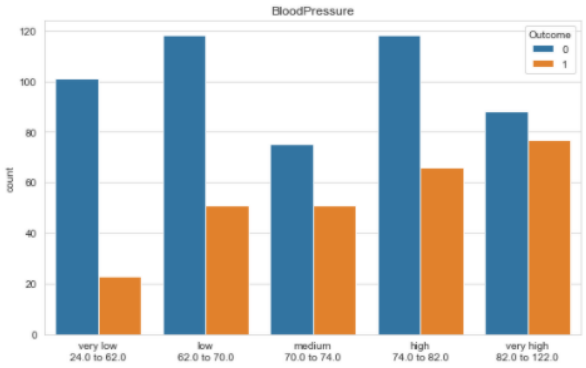
There are six "float64" type columns and 2 "int64" type.

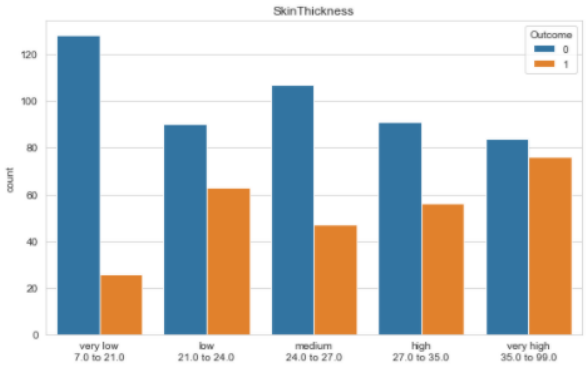
- Project Task: Week 2

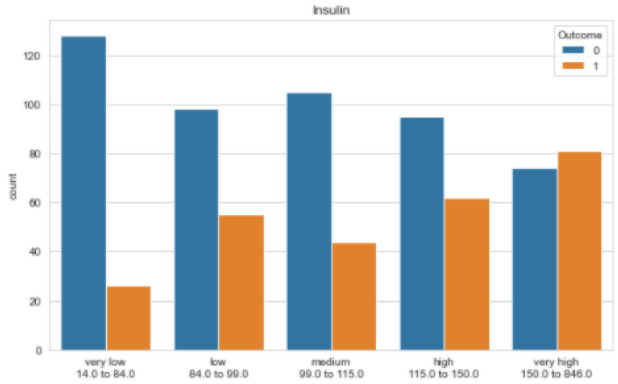
a) Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.

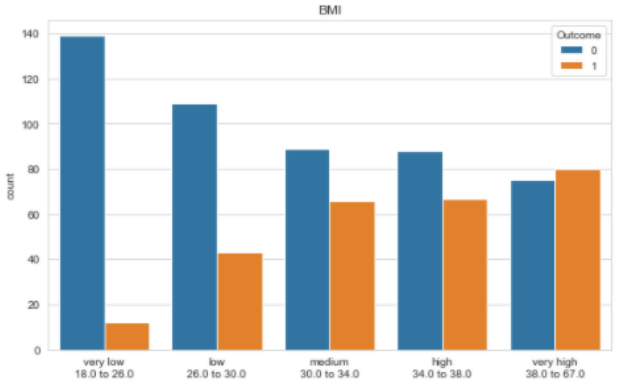


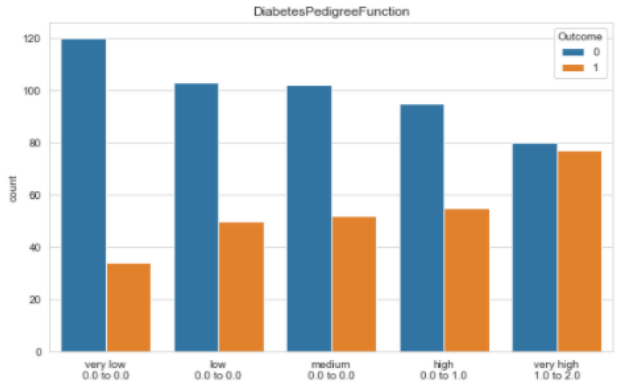


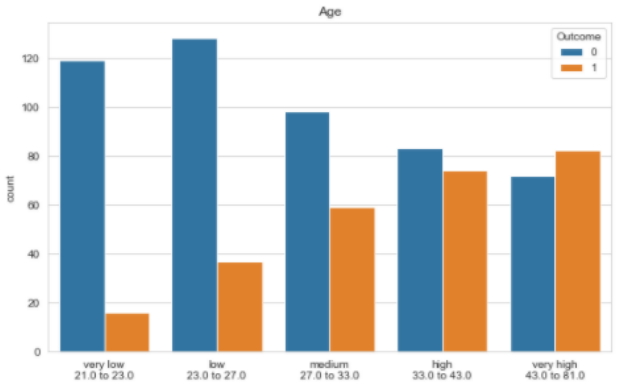






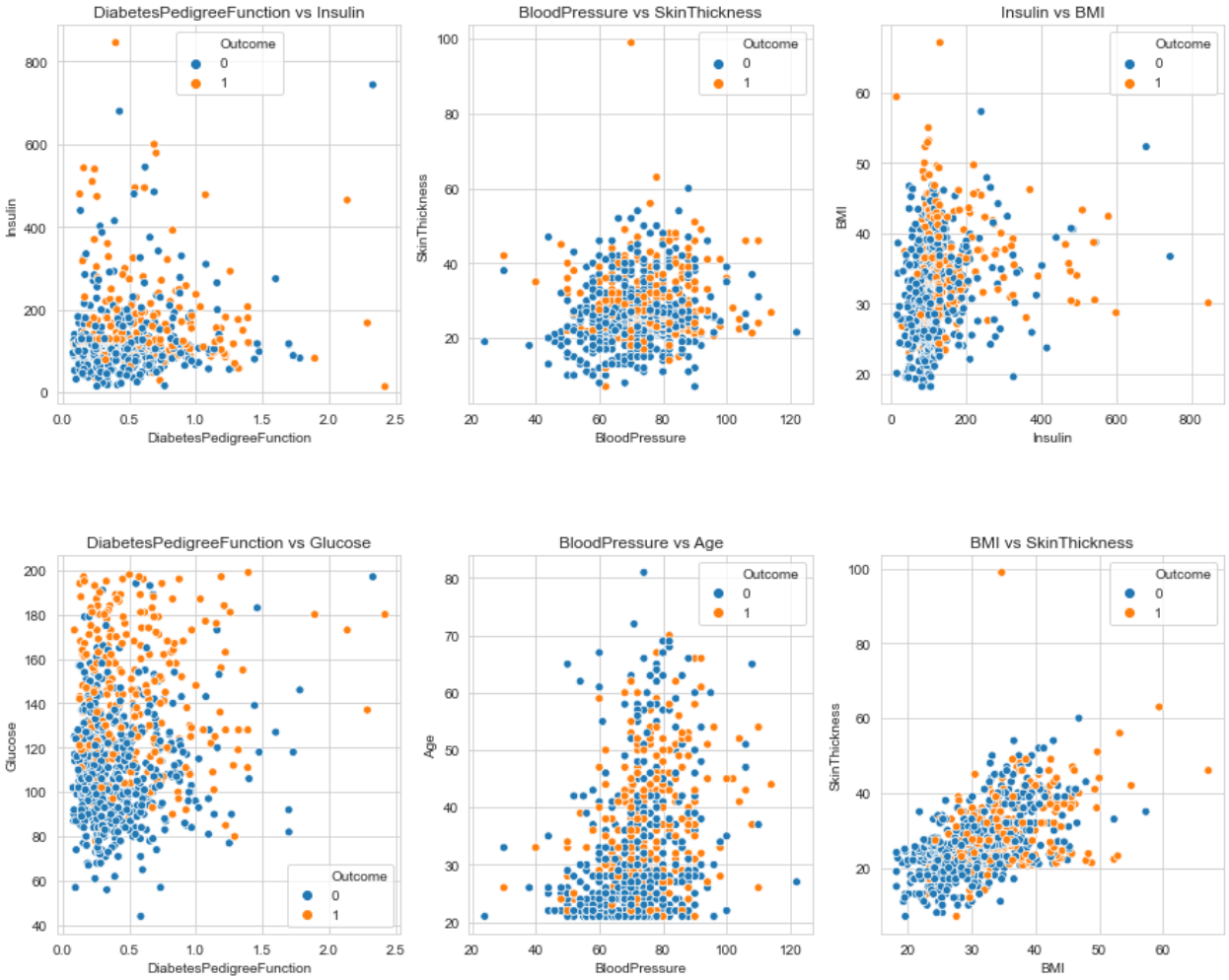






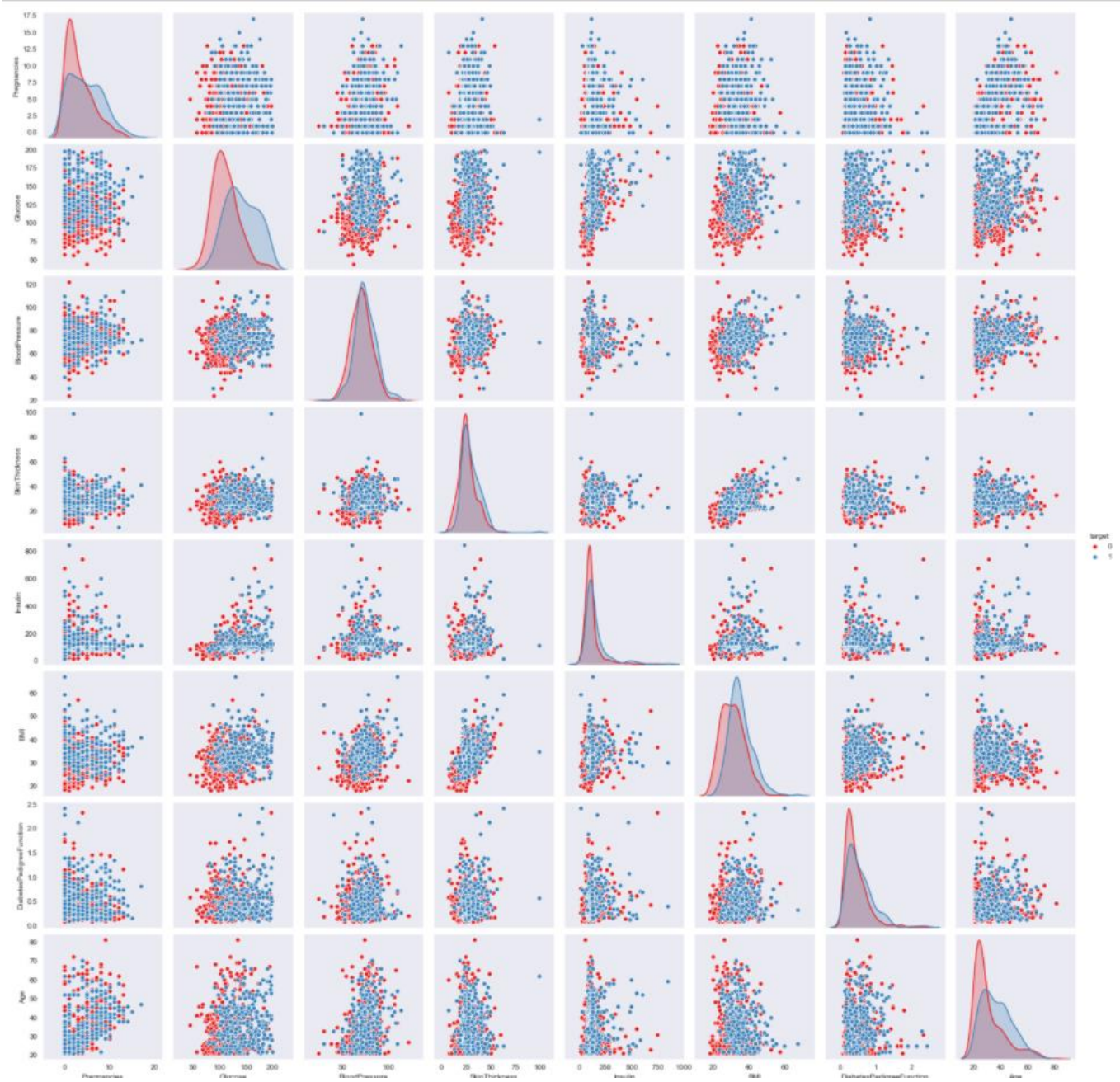
- From the above set of graph it is clear that as there is a solid positive correlation of Age, BMI, Glucose with the target variable i.e. Diabetes positive cases.
- Higher the blood pressure and glucose level higher will be the chances of diabetes.
- All the columns shows positive correleation with the output variable.

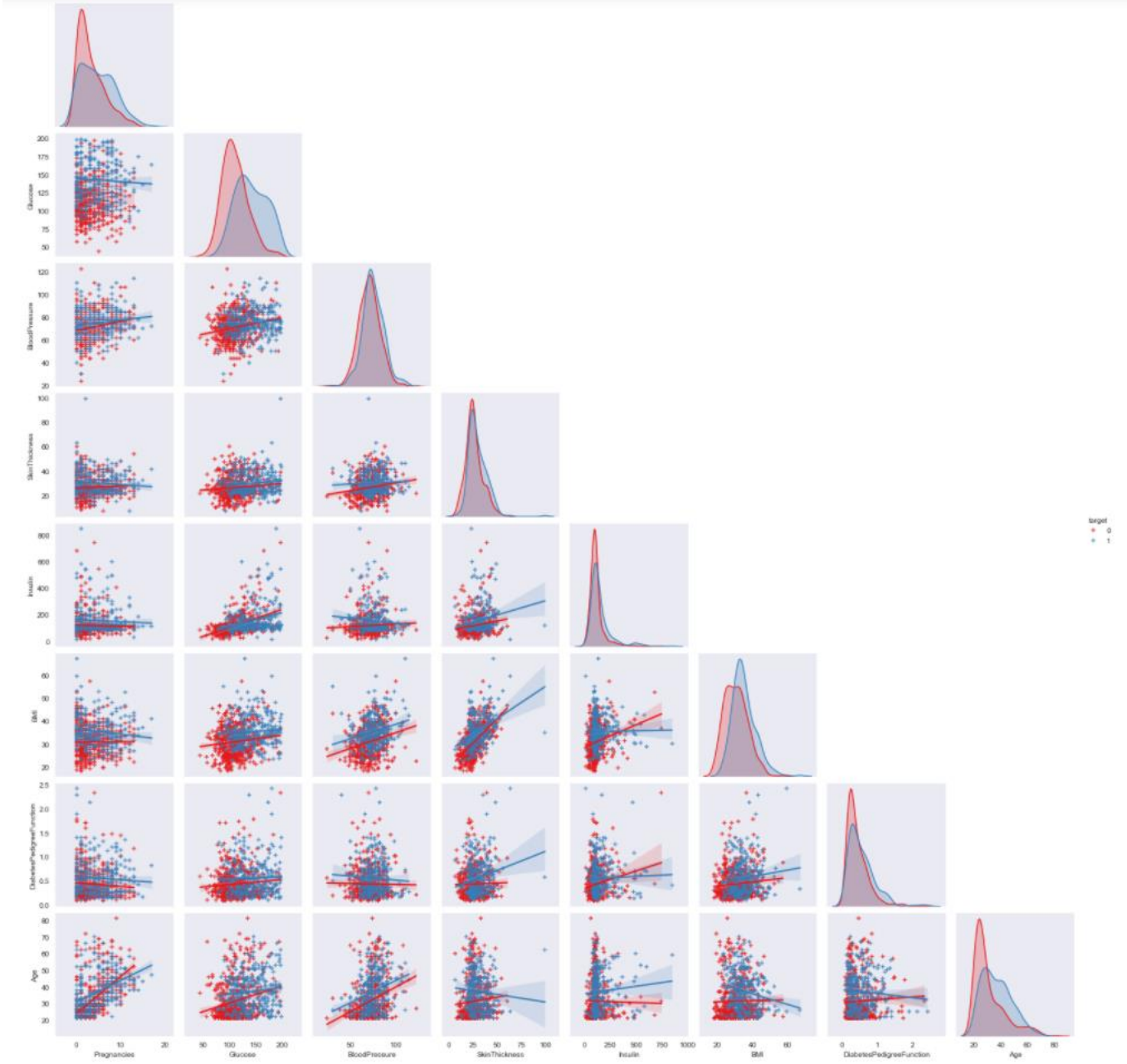
b) Create scatter charts between the pair of variables to understand the relationships. Describe your findings.



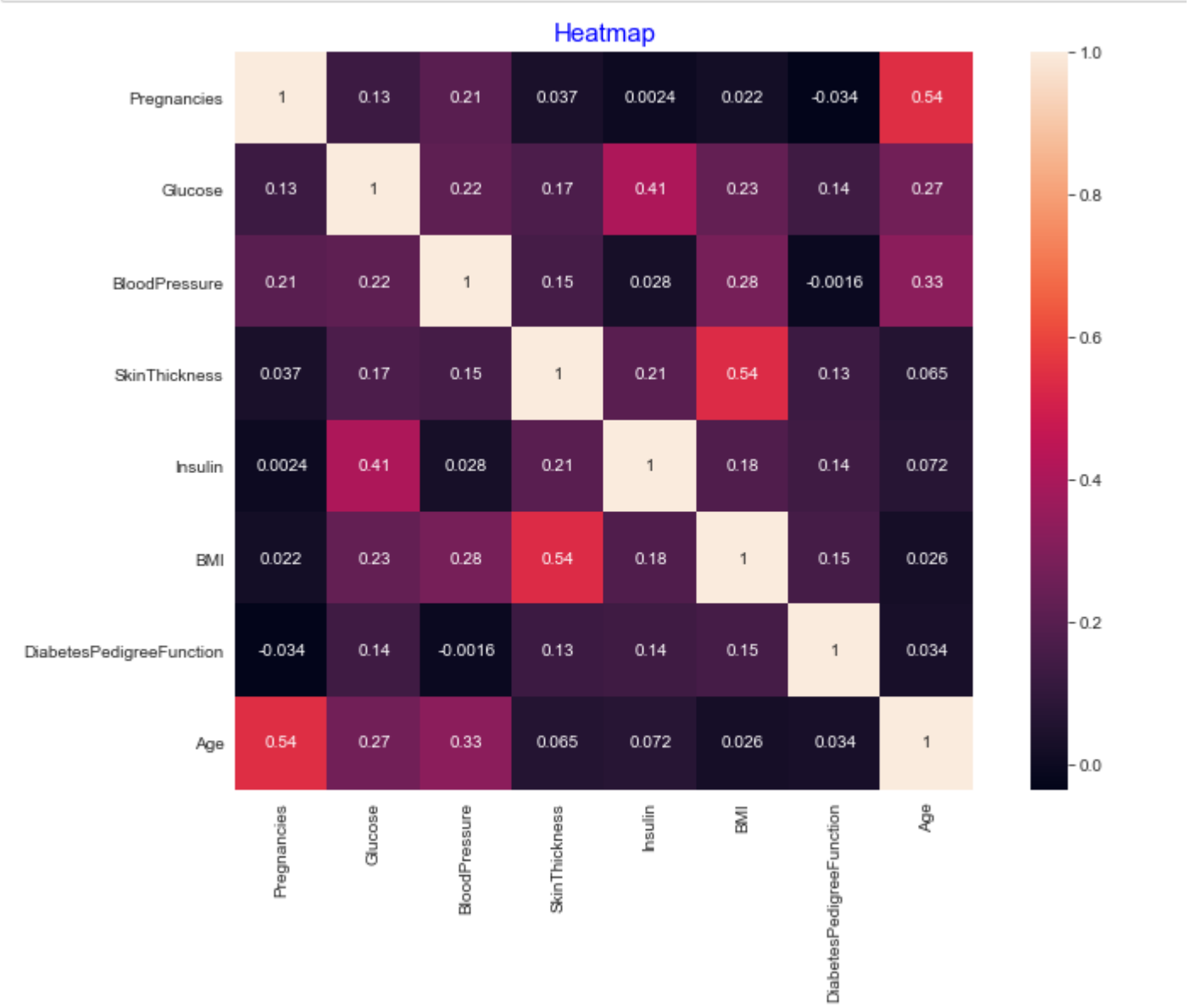
- The above scatter plot shows that we will able to separate target variable upto some extinct.
- For visualizing all relations, now we will plot pairplot.

Pair plot


Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
target
0
1


target
0
1
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age

c) Perform correlation analysis. Visually explore it using a heat map.



- Project Task: Week 3

a) Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.

- This problem is a classification problem in which we have to predict whether a person is suffering from diabetes or not.
- I will use four models:

  a) logistic Regression

  b) Random Forest

  c) SVM

  d) KNN

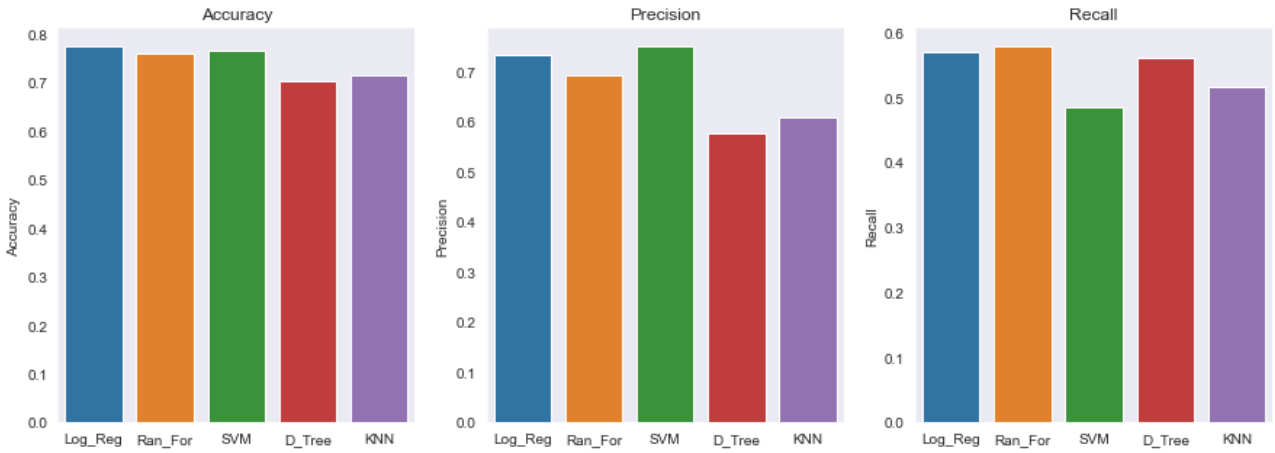
- K-Fold validation technique will be use in all models.

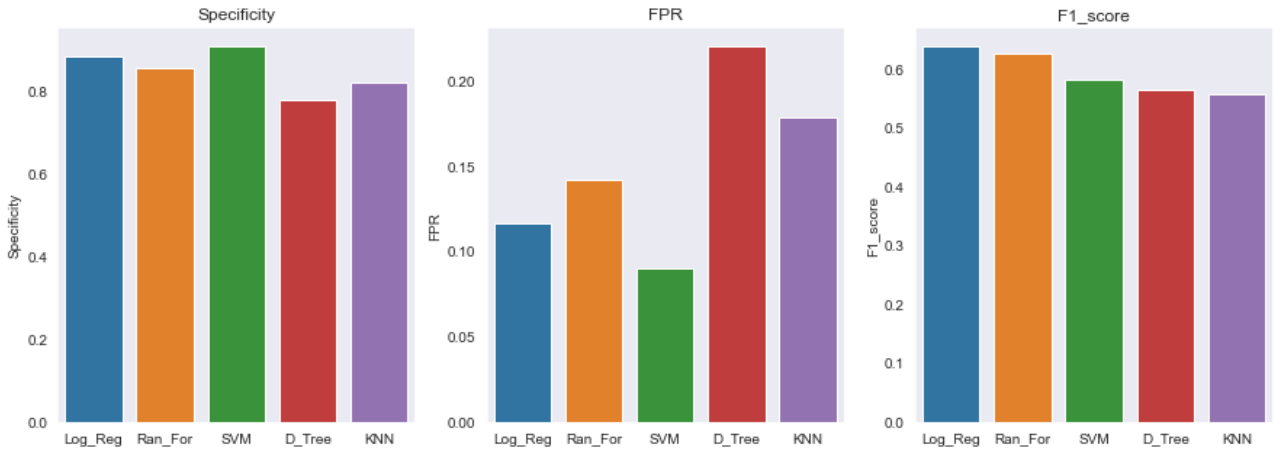
b) Apply an appropriate classification algorithm to build a model. Compare various models with the results from KNN algorithm.

c) Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc. Please be descriptive to explain what values of these parameter you have used.

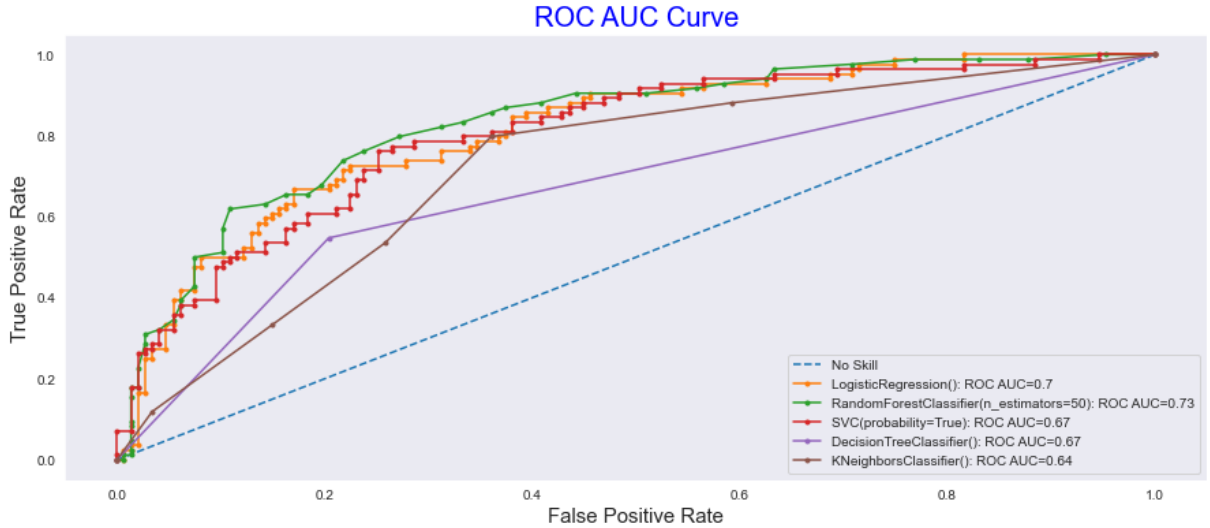
```
k_fold_df
```

	Log_Reg	Ran_For	SVM	D_Tree	KNN
Accuracy	0.773	0.760	0.764	0.704	0.714
Precision	0.735	0.693	0.750	0.578	0.609
Recall	0.569	0.578	0.486	0.562	0.516
Specificity	0.884	0.858	0.910	0.780	0.822
FPR	0.116	0.142	0.090	0.220	0.178
F1_score	0.638	0.626	0.582	0.565	0.557





- Accuracy of Logistic Regression, Random Forest and SVM are almost same.
- Precision of SVM and Logistic Regression is better than the other three models.
- Recall value of Decision Tree, Logistic Regression and Random Forest are better.



- According to above ROC AUC curve it is clear that Random forest performs best.
- Whereas three models --> Logistic regression, SVM and Random Forest gives almost similar result.

- Random Forest Result:

```
confusion_matrix(y_test,y_pred_rfc)
```

```
array([[134,  13],
       [ 37,  47]], dtype=int64)
```

```
rfc_acc = accuracy_score(y_test,y_pred_rfc)
rfc_acc
```

```
0.7835497835497836
```

- KNN (K-Nearest Neighbor):

```
confusion_matrix(y_test,y_pred_knn)
```

```
array([[124,  23],
       [ 43,  41]], dtype=int64)
```

```
knn_acc = accuracy_score(y_test,y_pred_knn)
knn_acc
```

```
0.7142857142857143
```

- Project Task: Week 4