18CSC301J-INFORMATION VISUALIZATION

SEMESTER VI

YEAR: 21-22

Diabetes predictions and its contributors

Report Submitted by

S.AJAY [RA1911003010102]
TARUUN.P [RA1911003010103]
SAIPAVAN TIPPANU [RA1911003010121]
S.KARTHIKEYAN [RA1911003010120]

Faculty in-charge

Debajyoti Modal
Kavitha

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

S.R.M. Nagar, Kattankulathur – 603203, Kancheepuram District

Agenda:

To take a dataset of people with diabetes and see the factors that contribute to diabetes and other ailments.

Abstract:

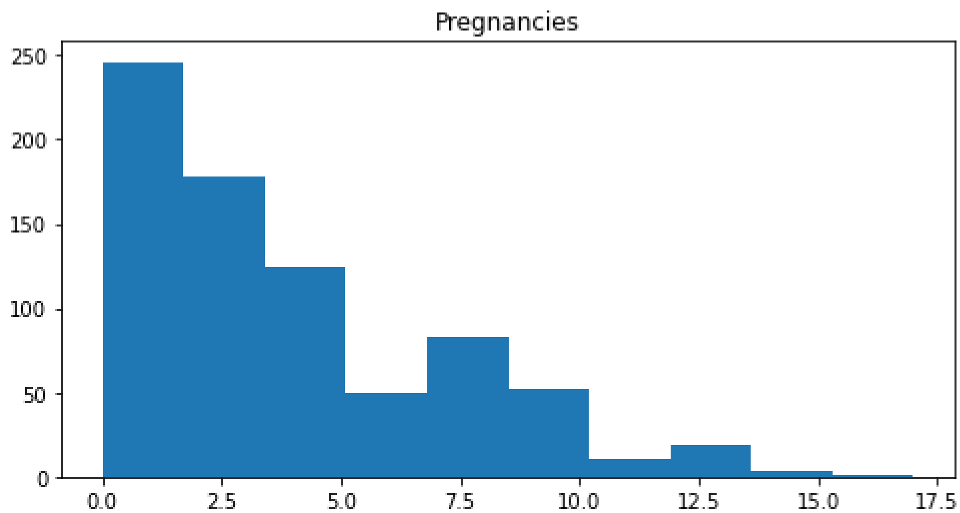
Diabetes is a prolonged infection or set of metabolic ailments where somebody suffers from a prolonged level of Blood Glucose(BG) in the body, due to deficiency of, or due to the non-reaction of the cells and do not react appropriately to insulin. Nowadays, this disease is leading to long-term impediments and stern health problems. Healthcare industry contains huge volumes of highly sensitive information that has to be dealt with appropriately.Diabetes Mellitus(DM)is considered to be a deadly sickness on the planet. Clinical experts need a reliable framework of expectations for the analysis of diabetes.

Introduction:

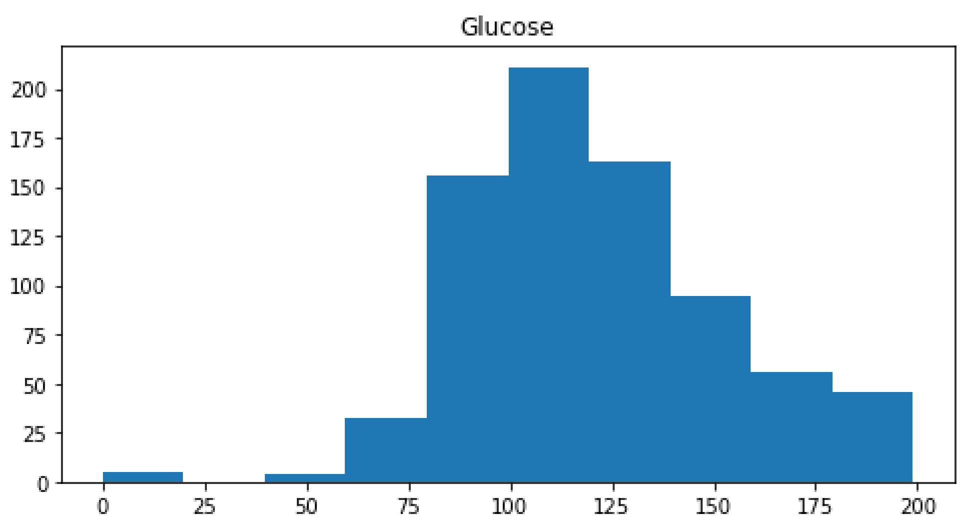
In India, diabetes is a major issue. Between 1971 and 2000, the incidence of diabetes rose ten times, from 1.2% to 12.1%. 61.3 million people 20–79 years of age in India are estimated living with diabetes (Expectations of 2011). It is expected that by 2030 this number will rise to 101,2 million.There are a few AI strategies that are utilized to perform prescient analysis of big data in a variety of domains. Prognostic health examination is a daunting work, but it can ultimately aid experts to make timely and well-versed judgments about the health and handling of patients. The core objective of the study is to assist physicians and experts in the primary diagnosis of diabetes by means of ML methods and visualizations.

OBSERVATION:

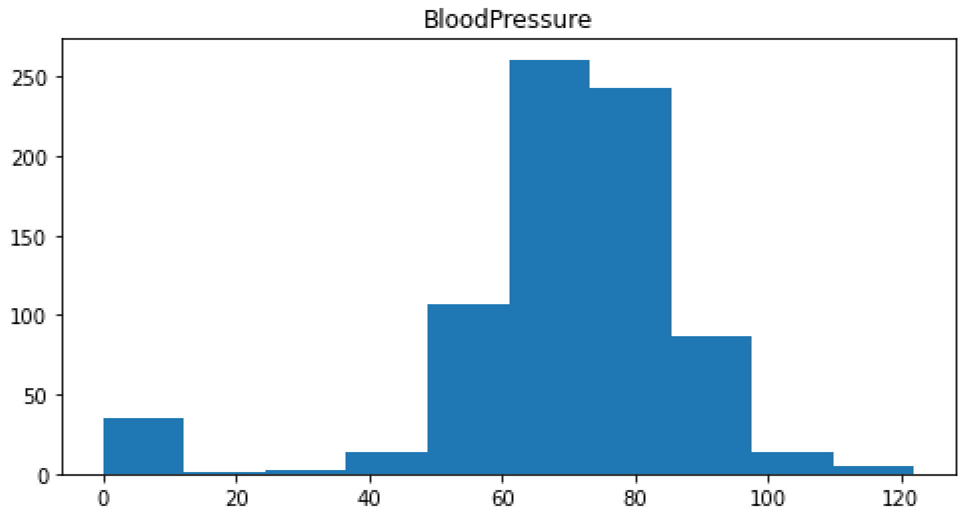
- **UNIVARIATE ANALYSIS**



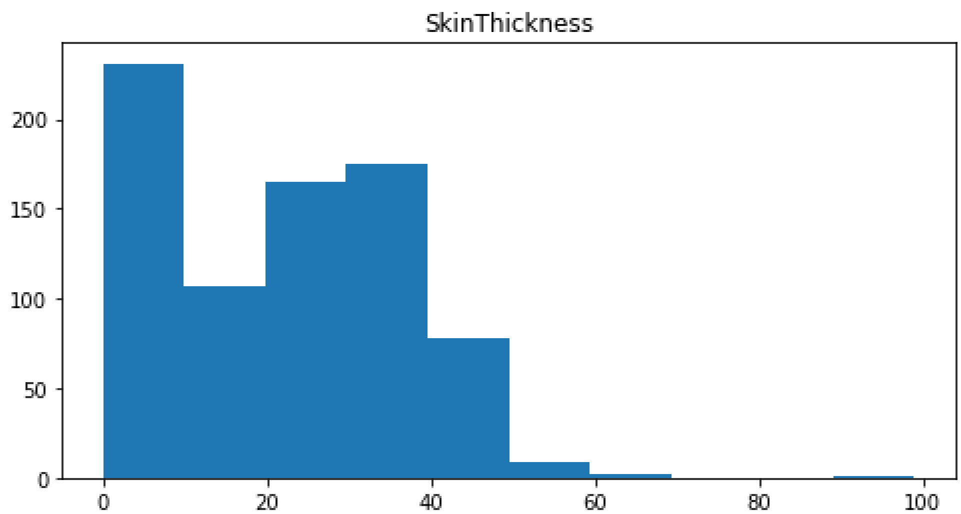
This depicts the number of pregnancies that are there amongst the dataset.Here, we can clearly see the number of pregnancies rates. i.e How many people have a certain number of pregnancy periods.



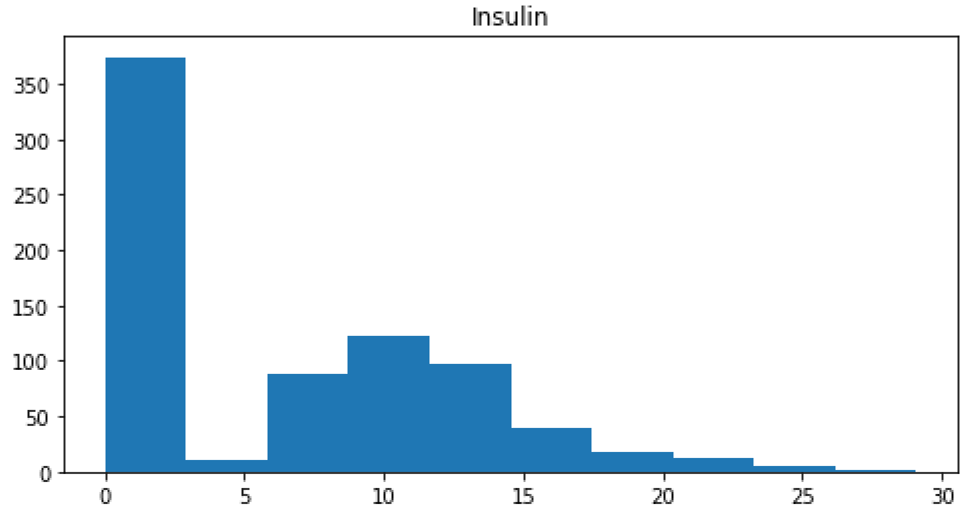
This graph depicts the glucose levels of the people pertaining to the dataset. There are many people with high glucose levels in the middle sector of the graph.



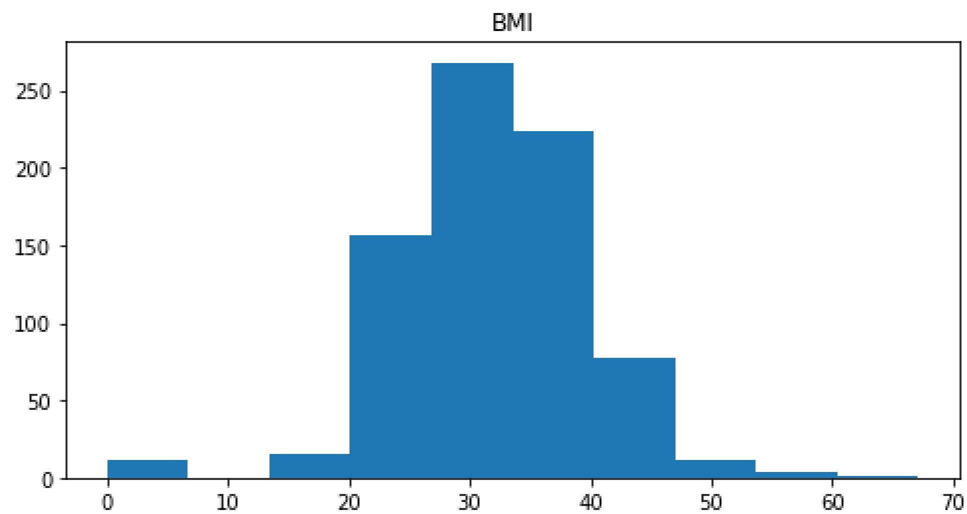
This graph depicts the blood pressure levels of the people pertaining to the dataset. Here, we can see that many people have BP within the range of 60-85 and a smaller number of people have less than 40.



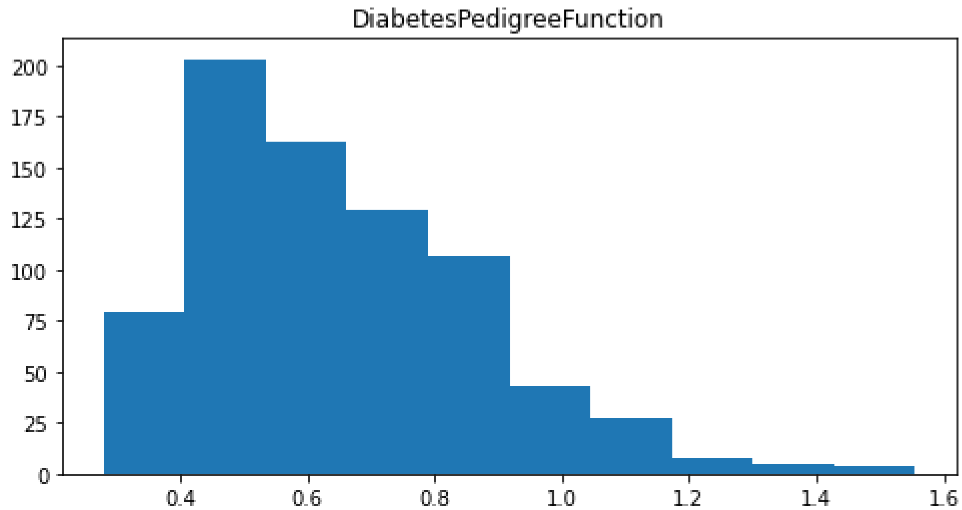
This graph depicts the skin thickness of the people pertaining to the dataset Here we can see that many people have skin thickness ranging from 0-40 and very minimal people have it in the range of 50+



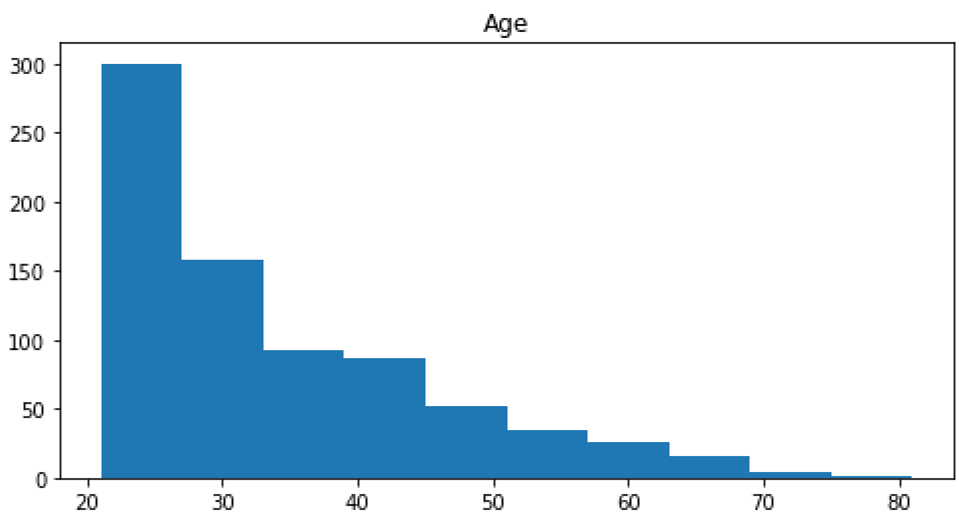
This graph depicts the insulin levels of the people pertaining to to the dataset
Here we can see that people have high insulin ranging from 0-3 and starts to gradually decrease from its second peak at 10.



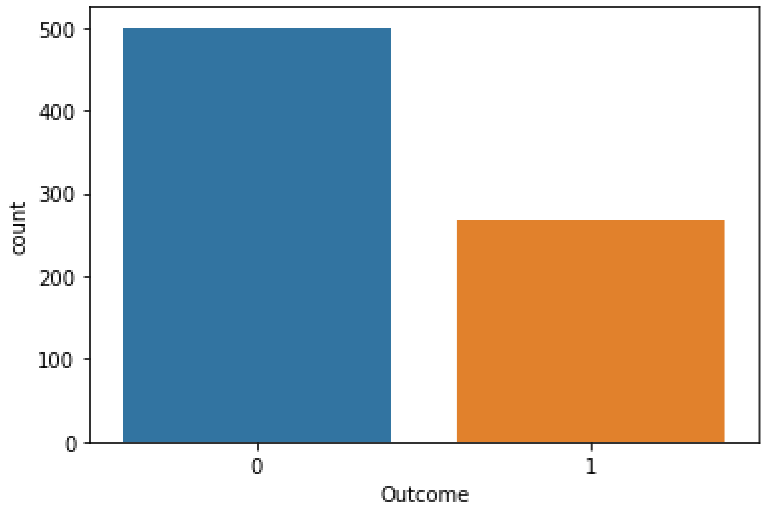
This graph depicts the BMI of the people pertaining to the dataset.
Here we can see that bmi peaks out at 30 and is high in the range of 20-45



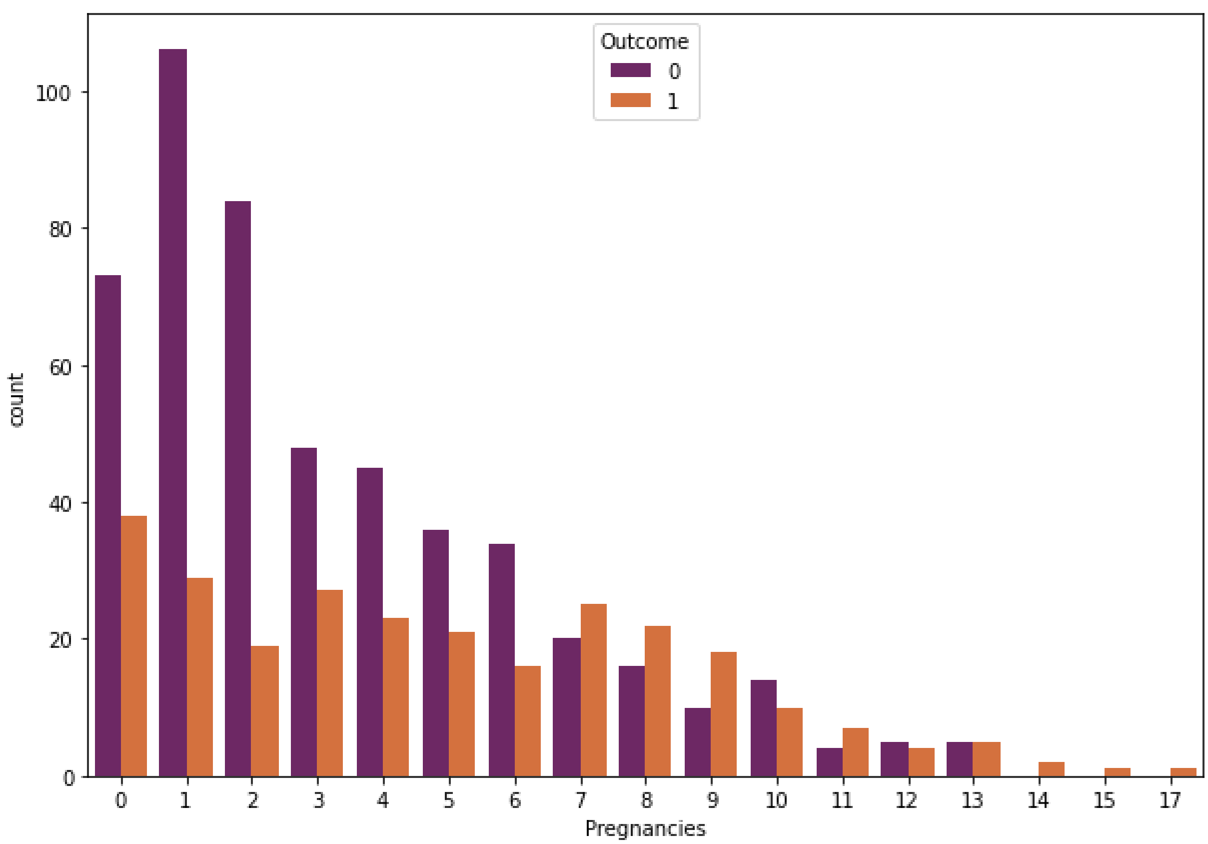
This graph depicts the Diabetes Pedigree function of the people pertaining to the dataset. Here we can see the graph peaking at 0.5 and gradually decreases upon further examination.



This graph depicts the Age of the people pertaining to the dataset
Here we can see the graph peaking at 0.5 and gradually decreases upon further examination

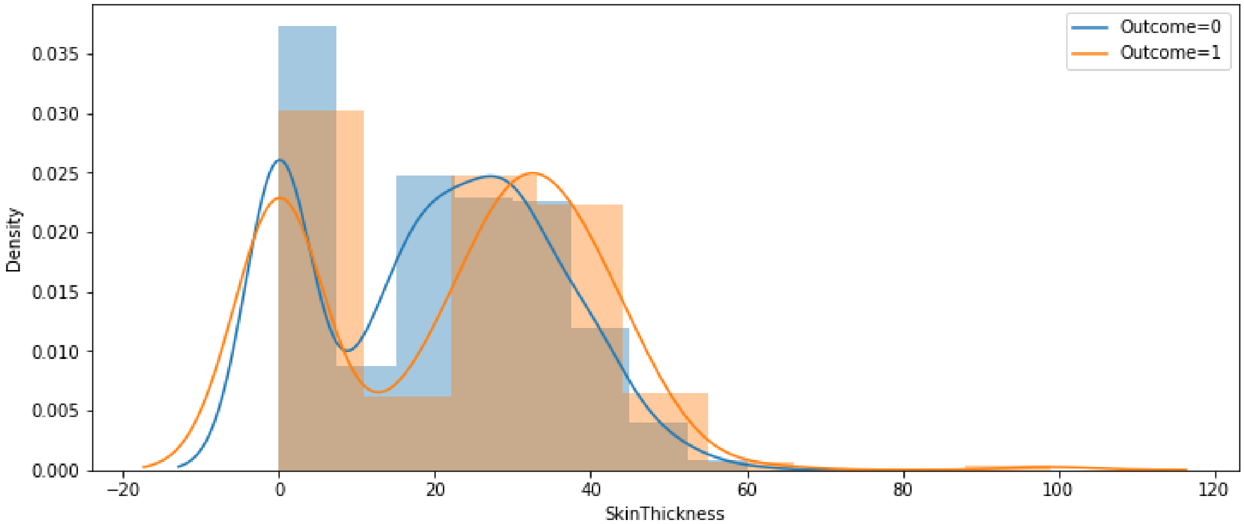


In this graph we can observe the outcome of the univariate analysis based on the previous graph's data as input. This graph can be deciphered by taking the orange as has diabetes and blue as does not have diabetes

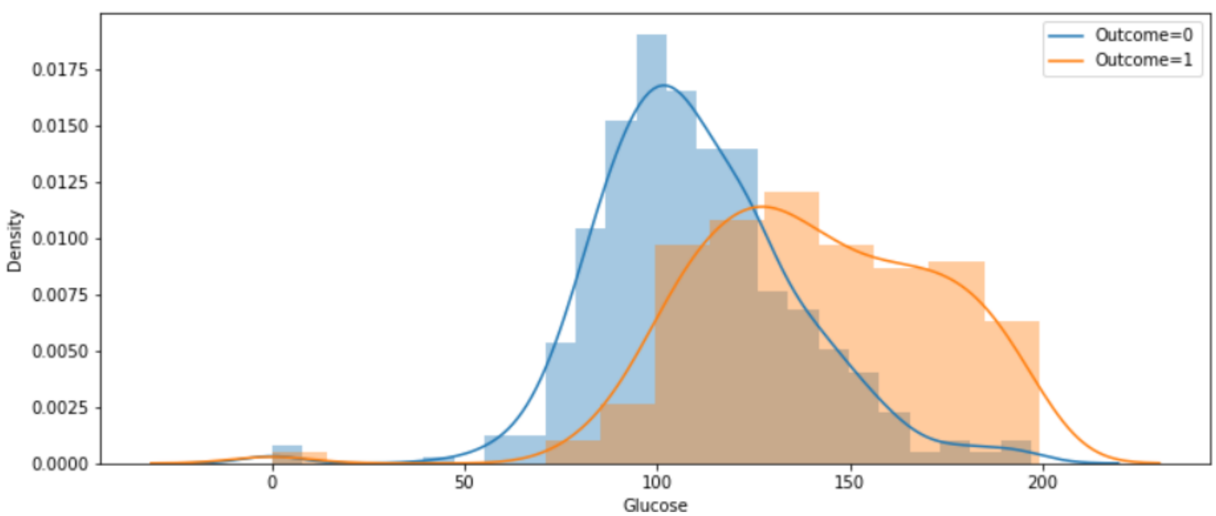


Via this viz, we can see that pregnancies might not always contribute to diabetes as the outcome of most of the pregnant women are not diabetic. This shows us that this might not only be the main contributing factor if the patient is diabetic.

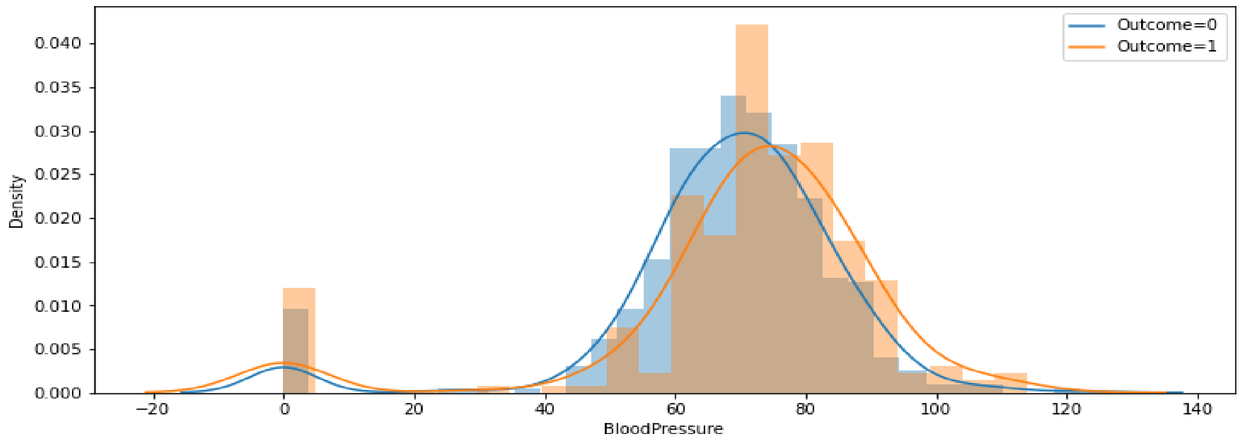
- **BIVARIATE ANALYSIS**



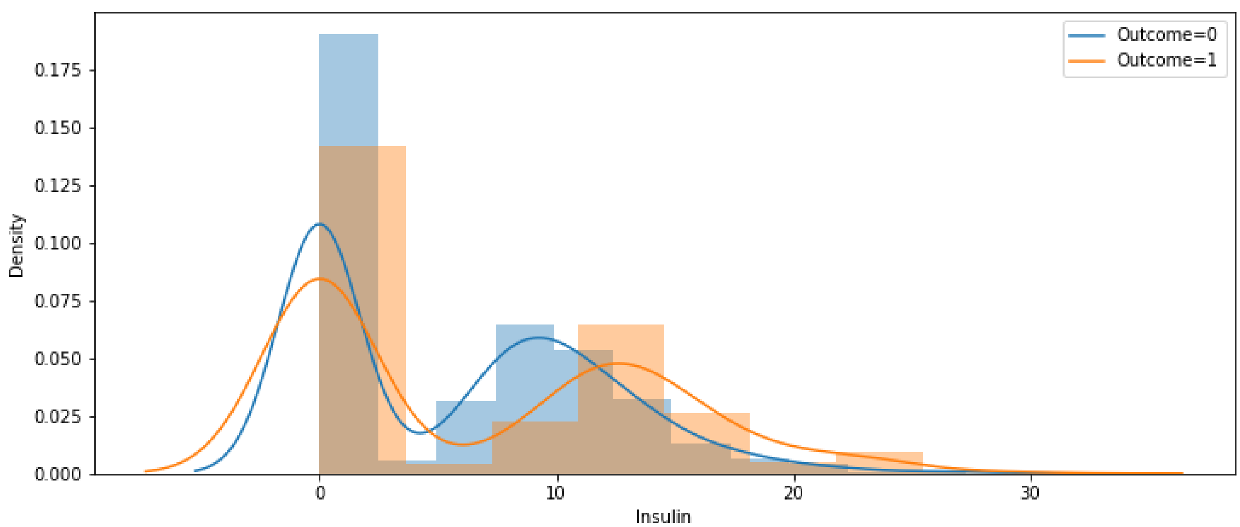
In this Graph we are able to grasp the correlation between thickness of skin and having diabetes.The graph also shows that, the outcome of diabetes is nearly the same for all the people of different skin thickness. Only in the vicinity of 0.025 density of 0 skin thickness and around 10-20, there is a clear correlation where we can see that the outcome is 0. And from around 30-60, the outcome is 1 where, we can conclude that this might be a contributing factor for diabetes.



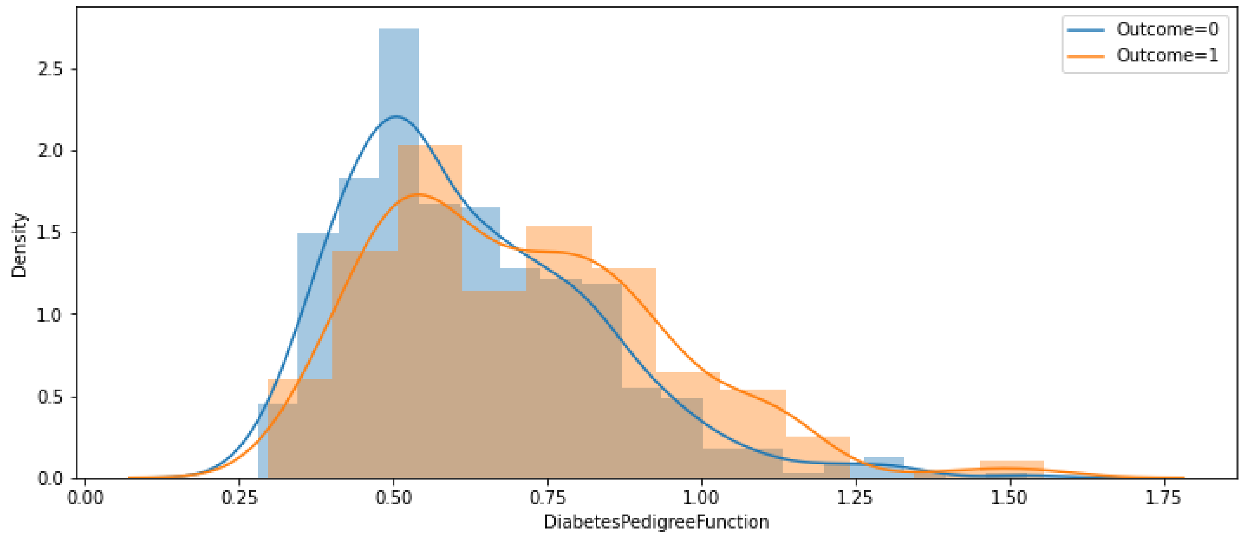
In this Graph we are able to grasp the correlation between Glucose and having Diabetes.The graph depicts the glucose level is directly proportional to the chances of diabetes. We can clearly see the graphs outcome around 100-250, the outcome 1.where, we can conclude that this might be a contributing factor for diabetes.



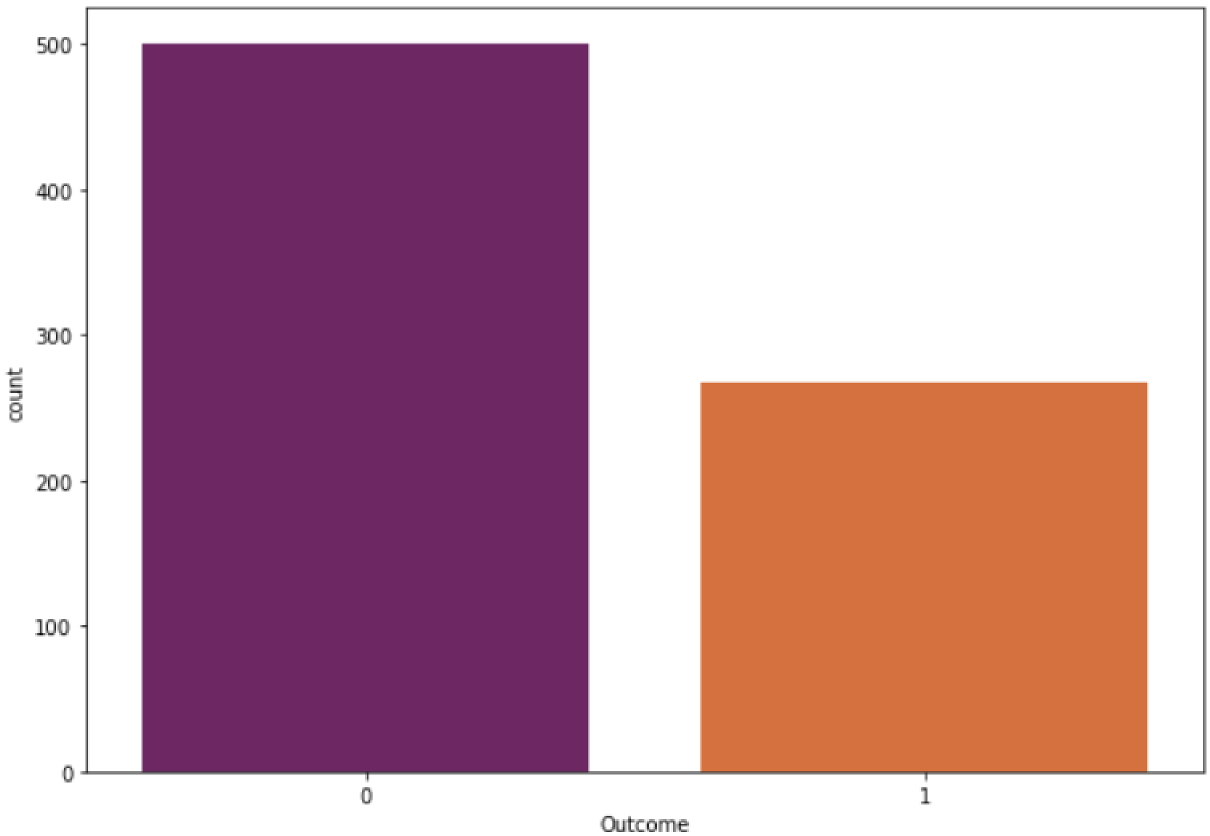
In this Graph we are able to see the correlation between blood pressure and having diabetes. The graph hasn't shown much variance till 40 but from 40-120 there is a slight variance. The dataset shows that people who have a bp higher than 80 have a higher chance of diabetes.



In this Graph we are able to grasp the correlation between Insulin and having Diabetes. The variance in insulin levels play a crucial role in telling a person if he has diabetes or not. As we can see from the graph the region around 5-12 has the outcome of 0 but from 15 and higher the outcome is 1. So higher or lower the insulin level, the more likely a person has diabetes. Ie. having less or excess amount of insulin might lead to diabetes.

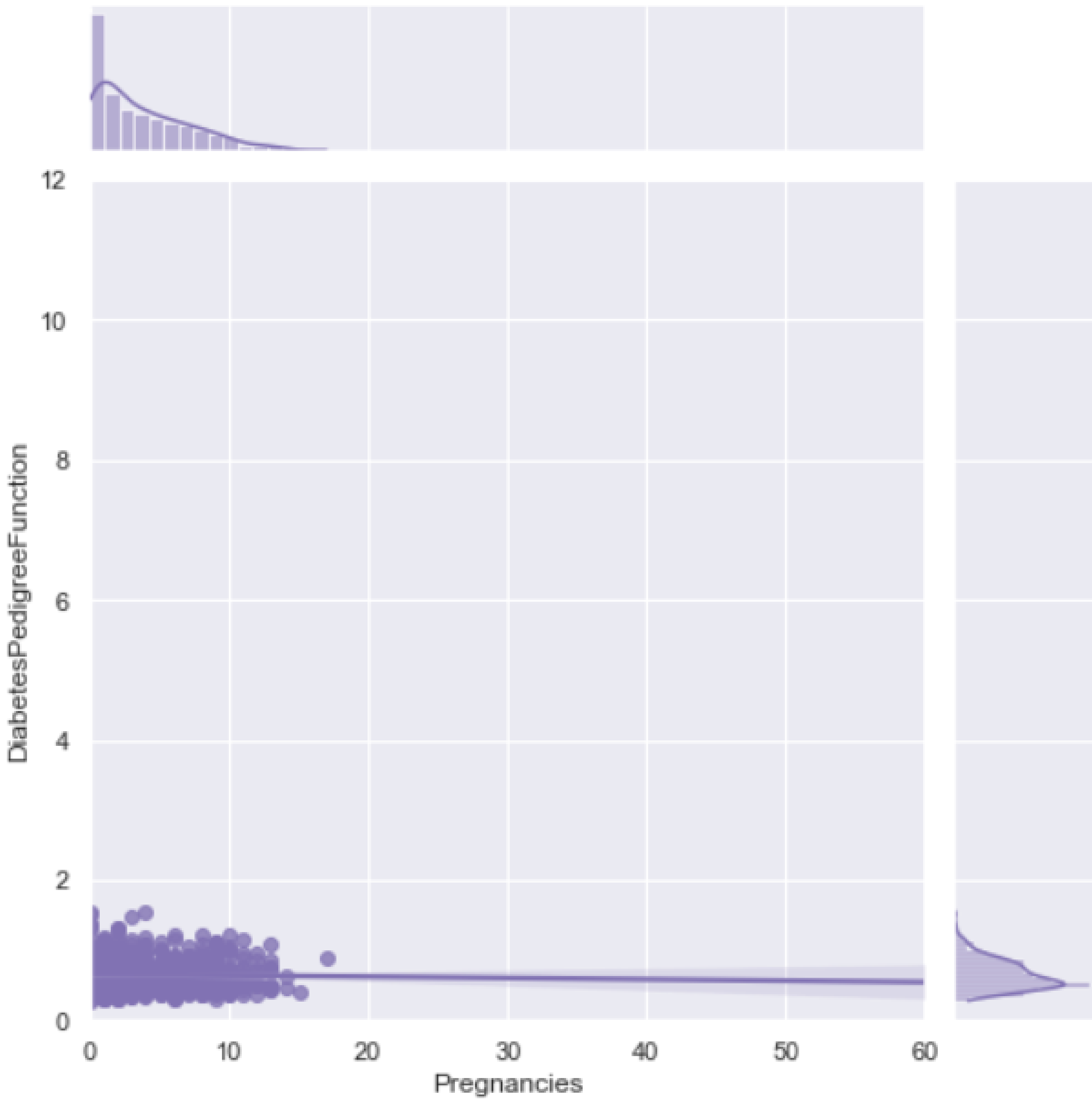


In this Graph we are able to grasp the correlation between Diabetes and having diabetes. From the given graph we can see from the graph around the region 0.25-0.75 the outcome is 0 but as the numbers get higher ie. from 0.75 - 1,25 the outcome is 1. We can conclude that with the increase in diabetes pedigree function, the outcome of diabetes increases.

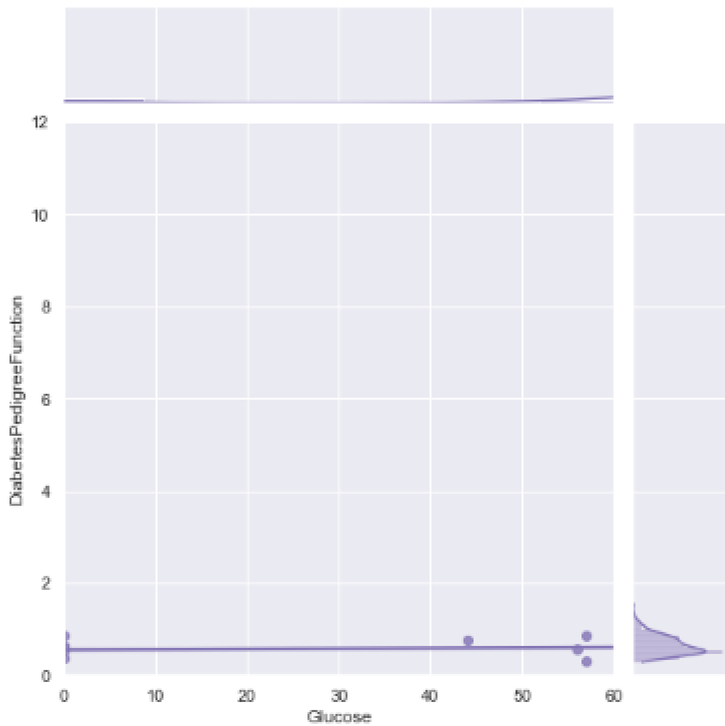


In this graph we can observe the outcome of the Bivariate analysis based on the previous graph's data as input. This graph can be deciphered by taking the orange as has diabetes and purple as does not have diabetes.

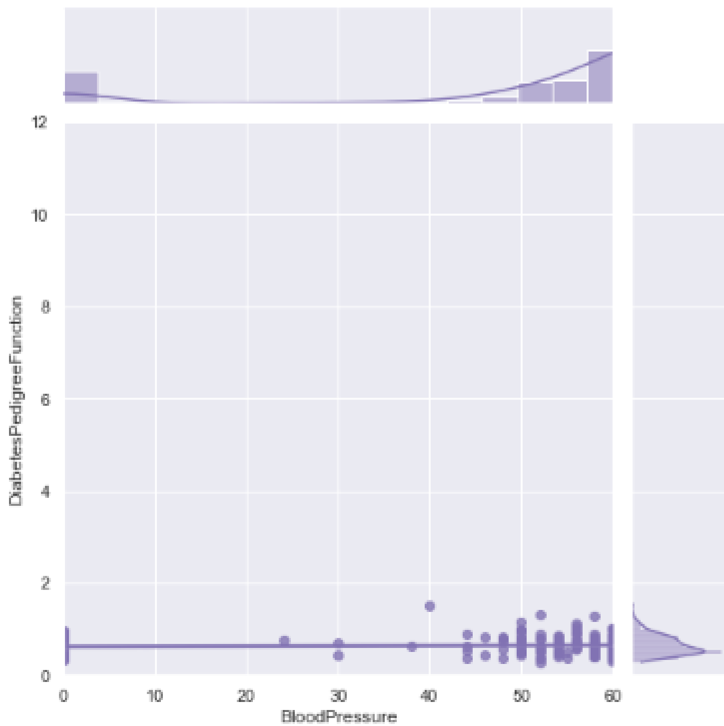
- **Linear Regression With Marginal Distribution**



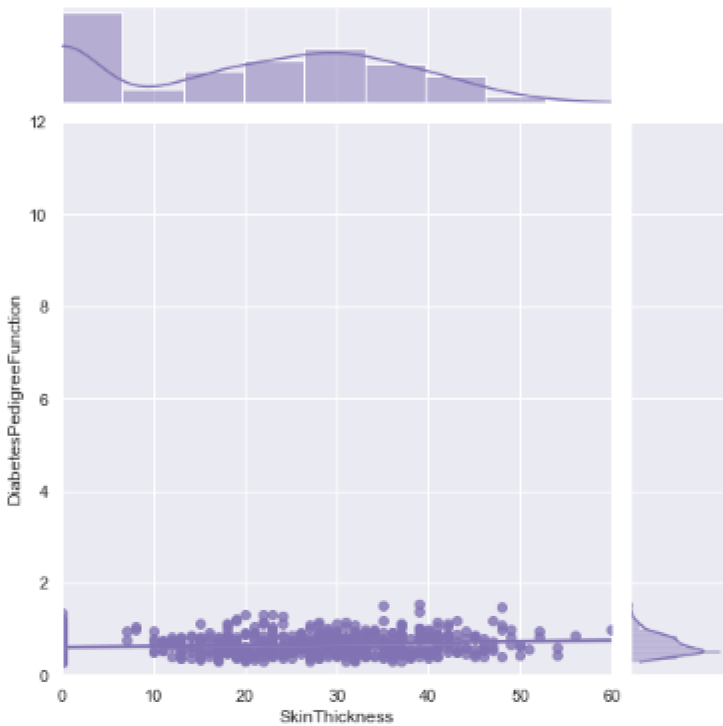
In this graph we are able to perceive the correlation between Diabetes Pedigree Function and Pregnancies. The number of pregnancies doesn't matter in this case as the diabetes pedigree function stays constant throughout. On the side we have the marginal side plots, wherein we can see the marginal distribution pertaining to diabetes pedigree function and pregnancies. It increases at the start and slopes down as the density decreases.



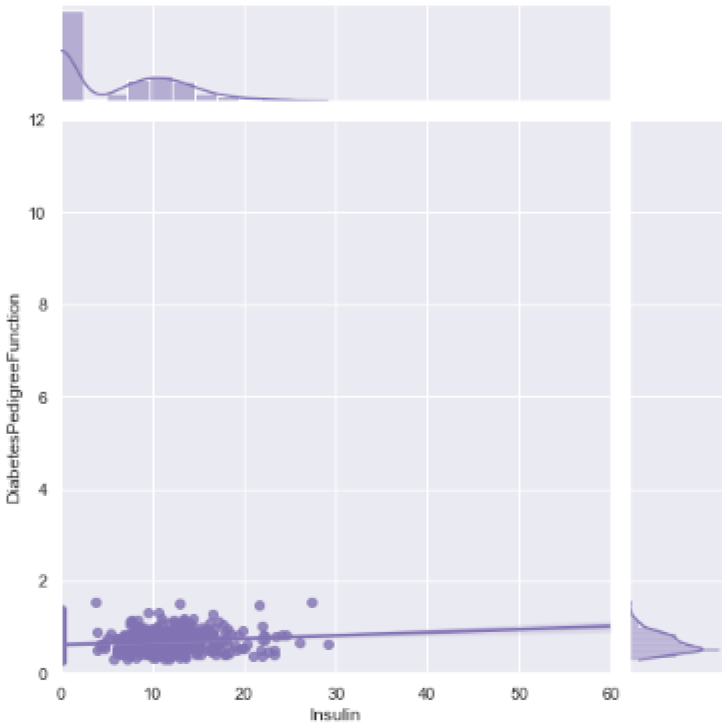
In this graph we are able to perceive the correlation between Diabetes Pedigree Function and Glucose.The factor of glucose doesn't contribute much in this case as the diabetes pedigree function stays constant throughout. On the side we have the marginal side plots, wherein we can see the marginal distribution pertaining to diabetes pedigree function and glucose. In this only at the end of the graph does the density and the graph have some minimal density and value because of which we can say that this is not that big of a factor of diabetes.



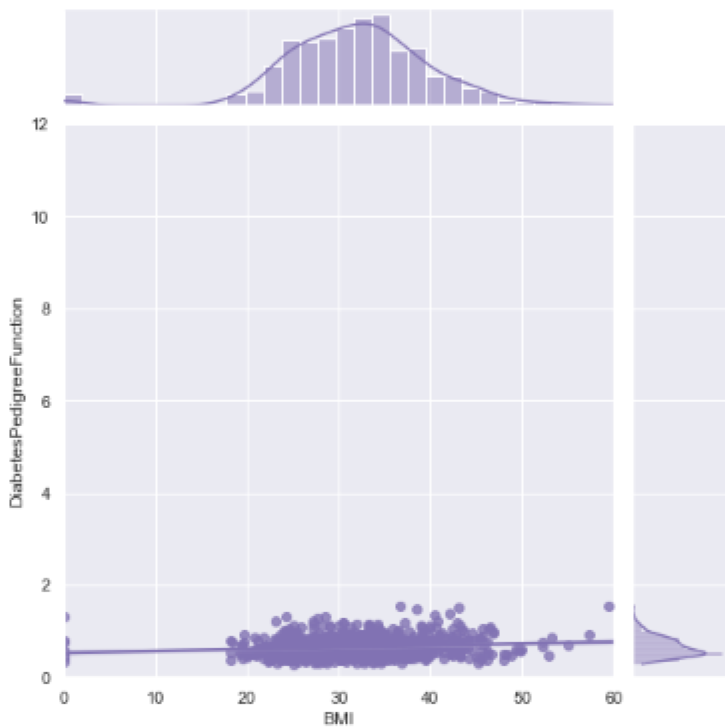
In this graph we are able to perceive the correlation between Diabetes Pedigree Function and Blood Pressure.The factor of Blood pressure doesn't contribute much as in this case as the diabetes pedigree function stays constant throughout at the beginning. On the side we have the marginal side plots, wherein we can see the marginal distribution pertaining to diabetes pedigree function and Blood Pressure. It starts to increase in the middle and spike up as the density increases. So, we can conclude that only above a certain margin, diabetes pedigree function might matter.



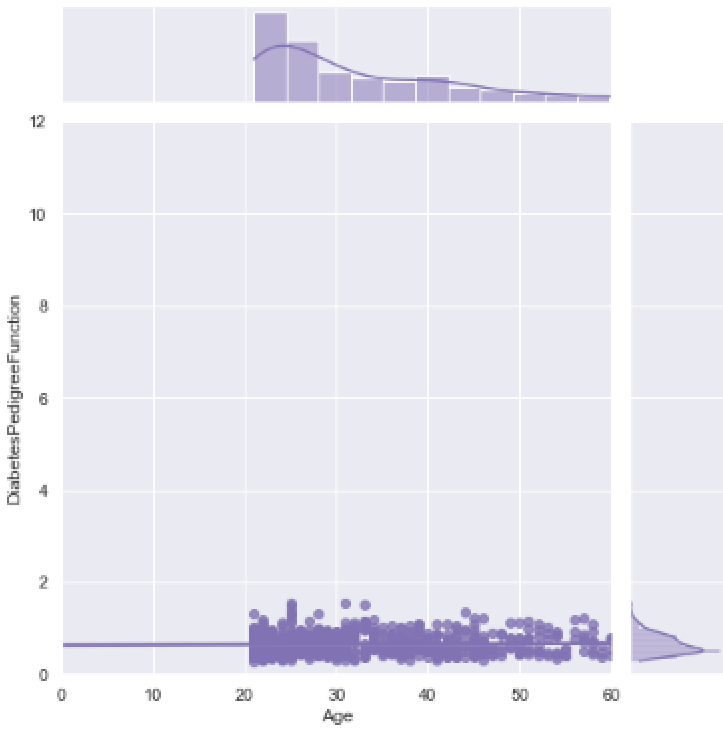
In this graph we are able to perceive the correlation between Diabetes Pedigree Function and Skin Thickness. The factor of Skin thickness contributes a lot as in this case as the diabetes pedigree function stays in the middle throughout and has a lot of density. On the side we have the marginal side plots, wherein we can see the marginal distribution pertaining to diabetes pedigree function and Skin Thickness. It increases at the start and slopes down a little then shoots up again as the density increases and gradually slopes down as the density decreases.



In this graph we are able to perceive the correlation between Diabetes Pedigree Function and Insulin. The factor of Insulin contributes initially as in this case as the diabetes pedigree function varies throughout at start. On the side we have the marginal side plots, wherein we can see the marginal distribution pertaining to diabetes pedigree function and Insulin.It increases at the start and slopes down a little then shoots up again as the density increases and gradually slopes down as the density decreases and flatlines. We can conclude this by saying that, a lot of people take in insulin when they have diabetes, wherein Diabetes pedigree function might be a factor.

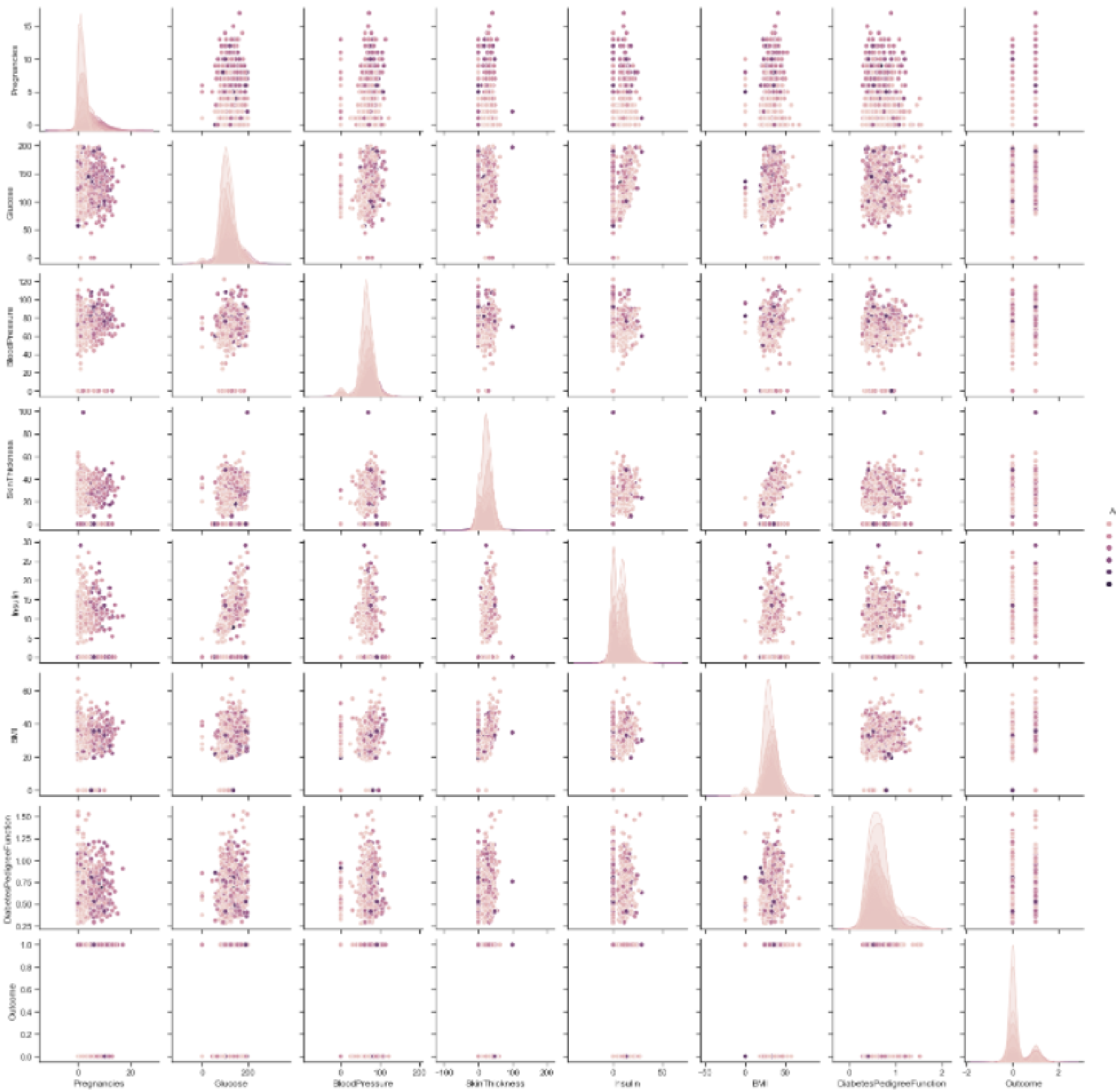


In this graph we are able to perceive the correlation between Diabetes Pedigree Function and BMI. The factor of BMI contributes a lot in this case as the diabetes pedigree function stays high throughout the middle sector. On the side we have the marginal side plots, wherein we can see the marginal distribution pertaining to diabetes pedigree function and BMI. In this we see a small rise in the beginning that flatlines immediately till we reach the middle where the graph peaks when it is having the maximum density and gradually slopes down as the density decreases. So, the BMI of diabetic people does vary between 20-50 heavily where the diabetes pedigree function might have a role in it as most of the people have at most 1 as the function value.



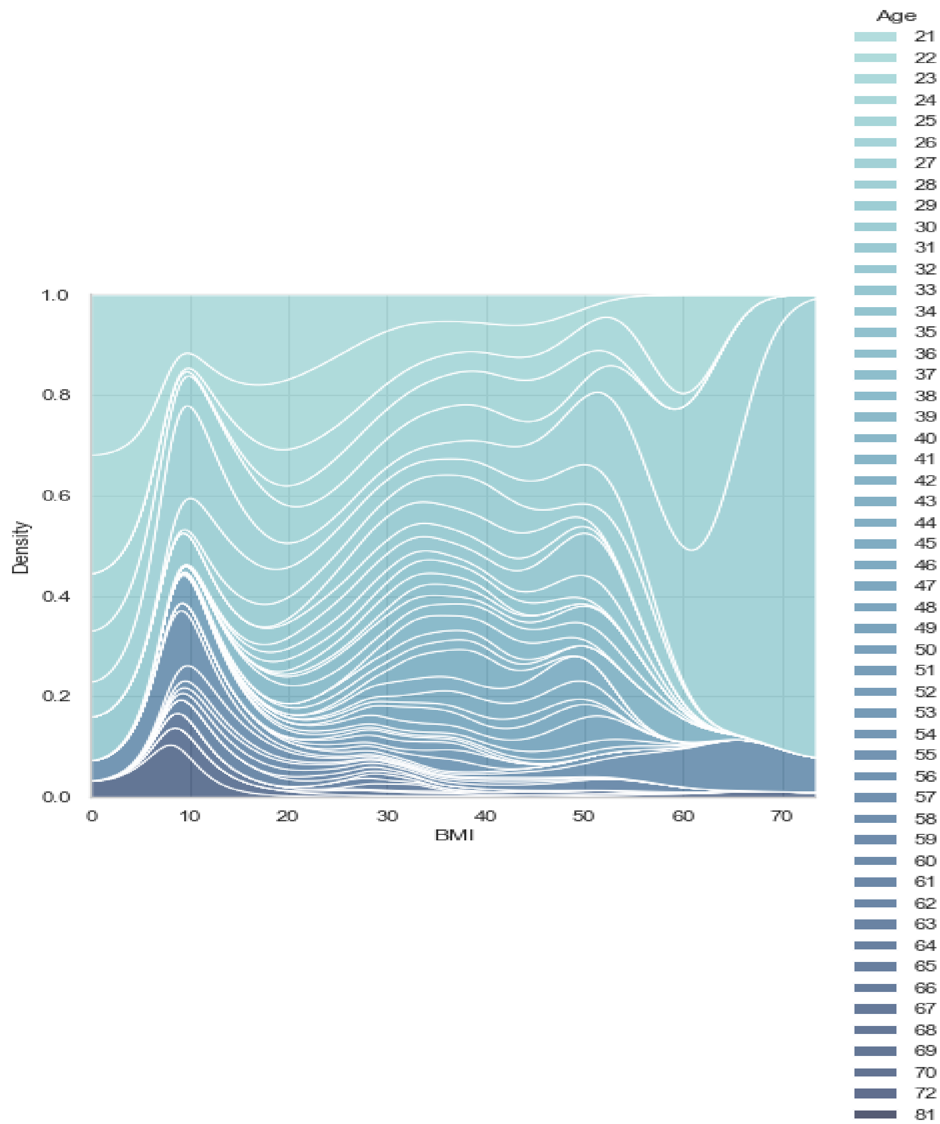
In this graph we are able to perceive the correlation between Diabetes Pedigree Function and Age. The factor of AGE contributes much in this case as the diabetes pedigree function stays increases and then decreases throughout. On the side we have the marginal side plots, wherein we can see the marginal distribution pertaining to diabetes pedigree function and AGE. in this we are able to see some data on the graph from the middle around the age of 20 as that is where the density begins to increase and maxes out within no time and gradually slopes down as the density decreases. The above marginal graph gives a great idea on how the density increases then decreases.

● Scatterplot Matrix

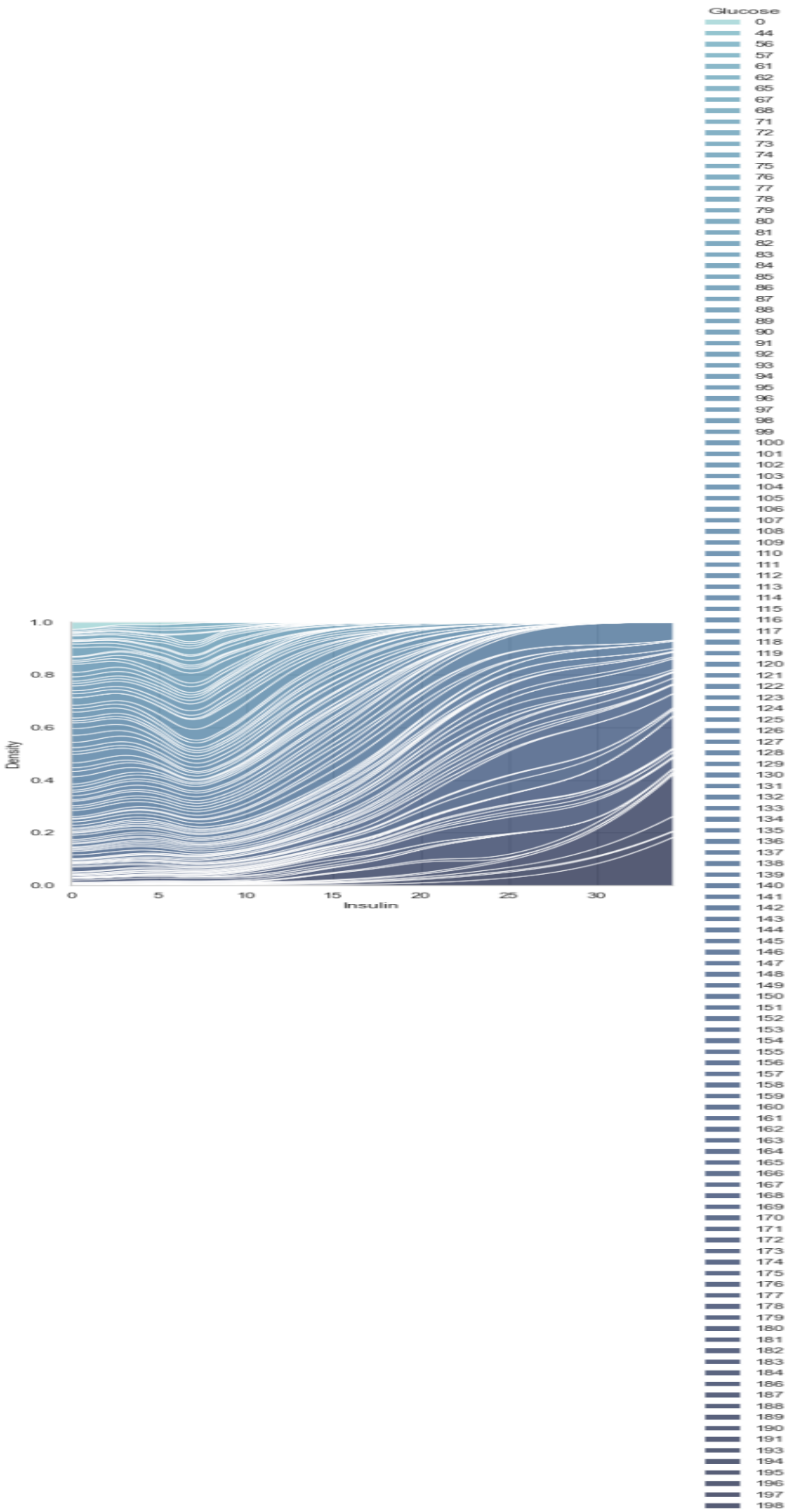


In this we are comparing all the attributes that are provided in the dataset and plotting a scatter plot matrix. Here, we have a scatterplot matrix wherein we use age as the base comparison feature. This plot gives a clear idea on how age contributes to each and every feature or body problems. Age is labeled in the form as hue where in when the age increases, the color darkness increases.

• Conditional kernel density estimate

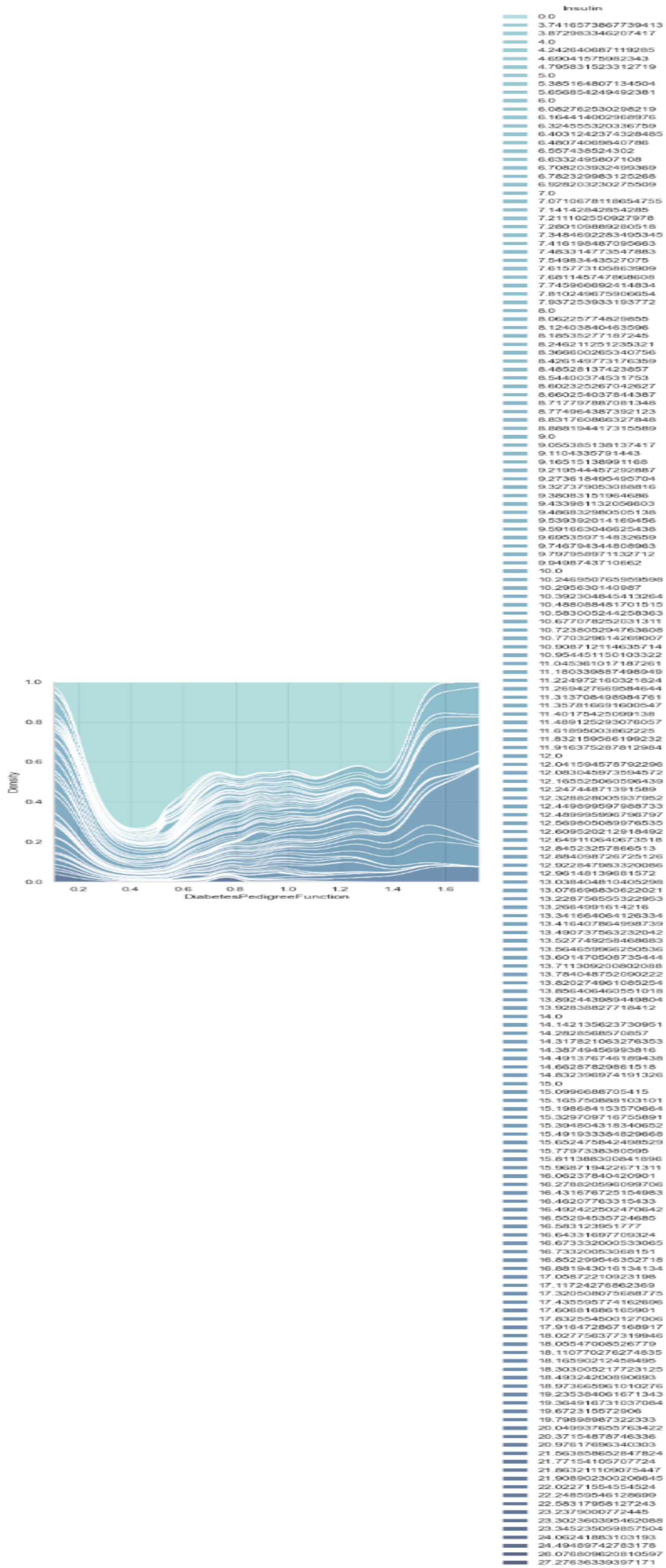


In this graph we are able to visualize the correlation between age and bmi as a factor of having diabetes. The bottom quarter of the relation is dominated by the old people, since their BMI is usually quite low as they tend to lose weight.ie. in the area around 10 BMI. The middle part consists of the age group of around 40-55. The upper portion of the viz. Consists of lower aged people since they tend to be of different weight and height proportions.



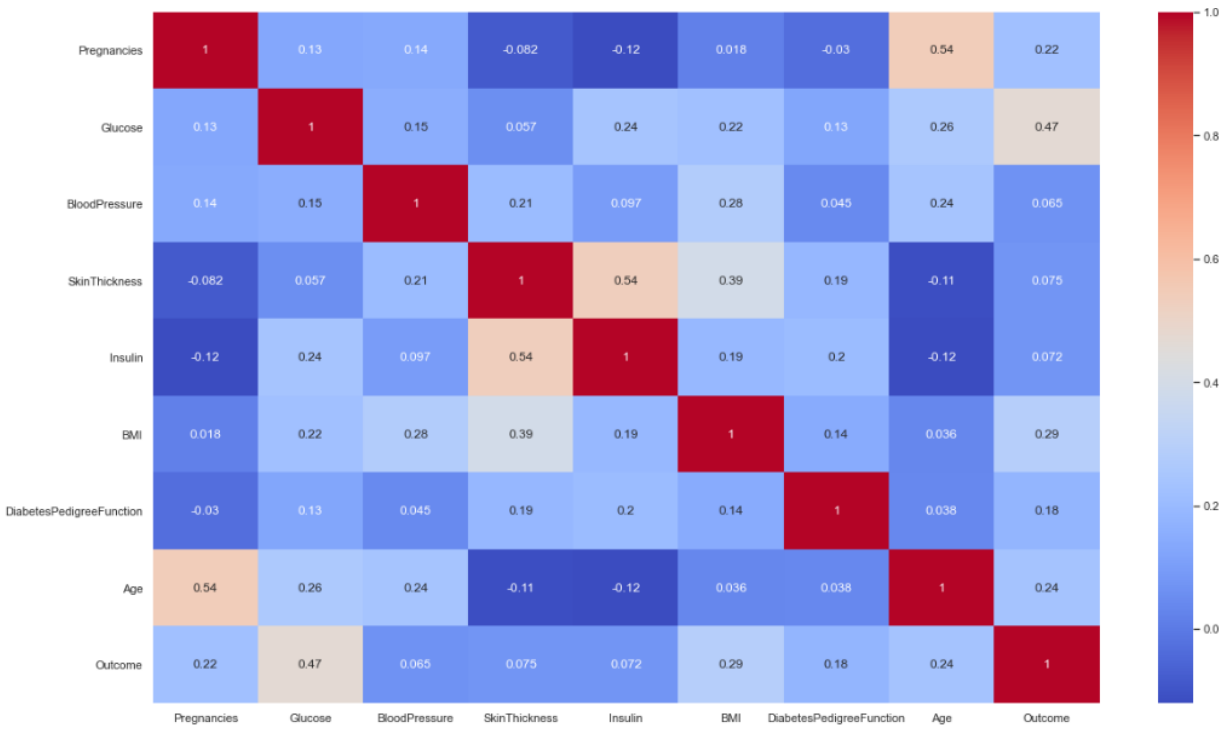
Glucose
0
44
56
57
61
62
65
67
68
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
186
187
188
189
190
191
193
194
195
196
197
198
199
1.0
0.8
0.6
0.4
0.2
0.0
Density
0
5
10
15
20
25
30
Insulin

In this graph we are able to visualize the correlation between the insulin level and glucose as a factor of having diabetes. Consider the levels from 5-15 the glucose levels are in the middle level but as the number goes up around 20 to 30 the graph gets darker. We can conclude from the graph that with the increasing levels of insulin, the glucose also goes higher.


1.0
0.8
0.6
0.4
0.2
0.0
Density
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
DiabetesPedigreeFunction
Insulin
0.0
3.7416573867739413
3.872983346207417
4.0
4.242640687119285
4.69041575982343
4.795831523312719
5.0
5.385164807134504
5.656854249492381
6.0
6.082762530298219
6.164414002968976
6.324555320336759
6.4031242374328485
6.48074069840786
6.557438524302
6.6332495807108
6.708203932499369
6.782329983125268
6.928203230275509
7.0
7.0710678118654755
7.14142842854285
7.211102550927978
7.280109889280518
7.3484692283495345
7.416198487095663
7.483314773547883
7.54983443527075
7.615773105863909
7.681145747868608
7.745966692414834
7.810249675906654
7.937253933193772
8.0
8.06225774829855
8.12403840463596
8.18535277187245
8.246211251235321
8.366600265340756
8.426149773176359
8.48528137423857
8.54400374531753
8.602325267042627
8.660254037844387
8.717797887081348
8.774964387392123
8.831760866327848
8.888194417315589
9.0
9.055385138137417
9.1104335791443
9.16515138991168
9.219544457292887
9.273618495495704
9.327379053088816
9.38083151964686
9.433981132056603
9.486832980505138
9.539392014169456
9.591663046625438
9.695359714832659
9.746794344808963
9.797958971132712
9.9498743710662
10.0
10.246950765959598
10.295630140987
10.392304845413264
10.488088481701515
10.583005244258363
10.677078252031311
10.723805294763608
10.770329614269007
10.908712114635714
10.954451150103322
11.045361017187261
11.180339887498949
11.224972160321824
11.269427669584644
11.313708498984761
11.357816691600547
11.40175425099138
11.489125293076057
11.61895003862225
11.832159566199232
11.916375287812984
12.0
12.041594578792296
12.083045973594572
12.165525060596439
12.24744871391589
12.328828005937952
12.449899597988733
12.489995996796797
12.569805089976535
12.609520212918492
12.649110640673518
12.84523257866513
12.884098726725126
12.922847983320086
12.96148139681572
13.038404810405298
13.076696830622021
13.228756555322953
13.2664991614216
13.341664064126334
13.416407864998739
13.490737563232042
13.527749258468683
13.564659966250536
13.601470508735444
13.711309200802088
13.784048752090222
13.820274961085254
13.856406460551018
13.892443989449804
13.92838827718412
14.0
14.142135623730951
14.2828568570857
14.317821063276353
14.38749456993816
14.491376746189438
14.66287829861518
14.832396974191326
15.0
15.0996688705415
15.165750888103101
15.198684153570664
15.329709716755891
15.394804318340652
15.491933384829668
15.652475842498529
15.7797338380595
15.811388300841896
15.968719422671311
16.06237840420901
16.278820596099706
16.431676725154983
16.46207763315433
16.492422502470642
16.55294535724685
16.583123951777
16.64331697709324
16.673332000533065
16.73320053068151
16.852299546352718
16.881943016134134
17.05872210923198
17.11724276862369
17.320508075688775
17.435595774162696
17.60681686165901
17.832554500127006
17.916472867168917
18.027756377319946
18.05547008526779
18.110770276274835
18.16590212458495
18.303005217723125
18.49324200890693
18.973665961010276
19.235384061671343
19.364916731037084
19.672315572906
19.79898987322333
20.049937655763422
20.37154878746336
20.97617696340303
21.563858652847824
21.77154105707724
21.863211109075447
21.908902300206645
22.02271554554524
22.24859546128699
22.58317958127243
23.2379000772445
23.302360395462088
23.345235059857504
24.06241883103193
24.49489742783178
26.076809620810597
27.27636339397171
29.086079144497972

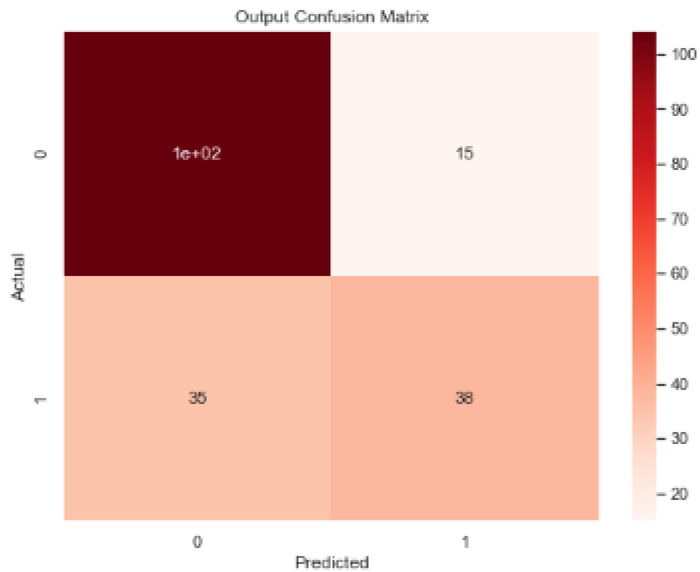
In this graph we are able to visualize the correlation between the insulin level and diabetes pedigree function as a factor of having diabetes. Diabetes pedigree function (**a function which scores likelihood of diabetes based on family history**). We can see from the graph that the low levels around 0 - 0.4 and 0.5 to higher levels can be found easily. All the other middle area looks pretty normal

• Correlation Matrix



This matrix gives us a clear picture of how each and every feature is correlated to the other. The warmer the label is, the more the contributing factor it is in that feature. This is also a diagonal matrix with value 1 where it has its own feature repeated. The other values correspond to the relationship between the feature and the other feature it is contributing to.

• Implementing Confusion Matrix



This matrix is derived after producing the supervised training methods. This confusion matrix gives a clear picture of how each and every factor has contributed. Eg. Accuracy,F1 score.

Conclusion:

With the help of algorithms like Univariate, Bivariate, Scatterplot Matrix, Linear Regression, Conditional Kernel density estimation,correlation matrix and confusion matrix we were able to visualize the relationship between the people can posses that could be a trait of having and possibly predicting a person having or getting Diabetes

For reference, visit the link provided below:
Github Link:https://github.com/legendslayer01/Diabetes-IV-viz
Refer the link for a more clear view of viz. Pictures and see the code.