Assignment 7

Data Analytics

Security Tools Lab 1

Gowtham Baskar
1006523

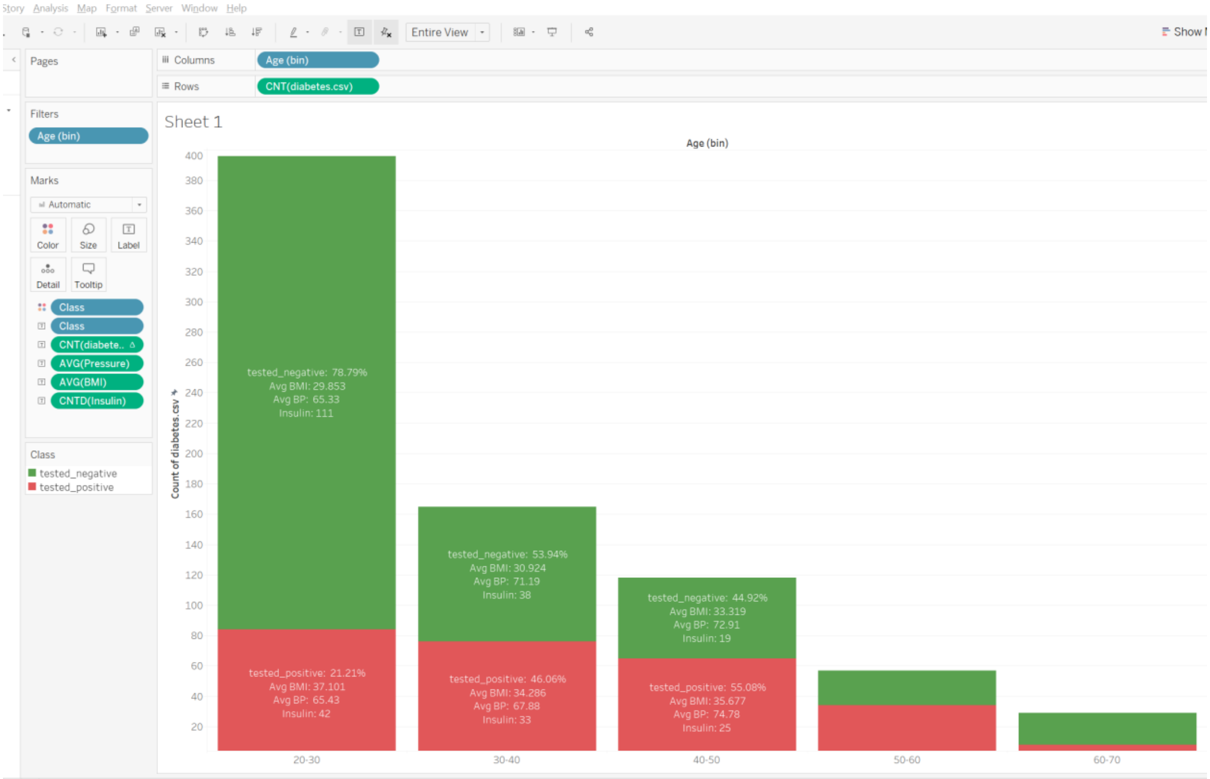
1. Data Visualization

Hypothesis

People with high blood pressure, high BMI, No Insulin intake are tested positive to diabetes more than other factors.

Stack Bar Chart

From the hypothesis above, we can acquire the resulted stack bar chart as below.



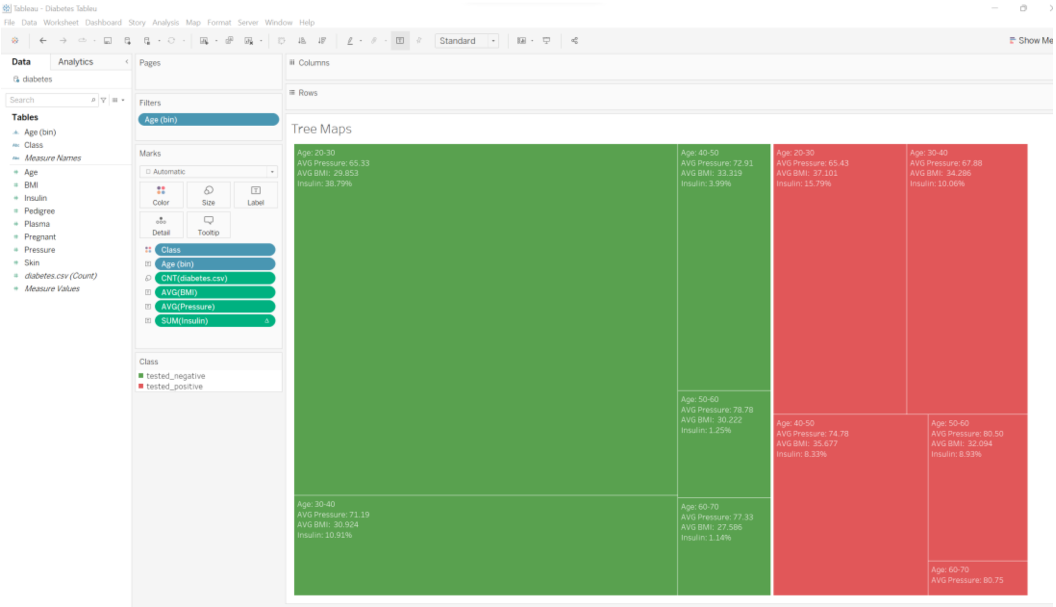
Columns: Age (bin)
Rows: CNT(diabetes.csv)
Filters: Age (bin)
Marks: Automatic
Class
Class
CNT(diabete.. Δ
AVG(Pressure)
AVG(BMI)
CNTD(Insulin)

From this chart, I've filtered age by removing any person above the age 70+ as the data were very less.

From the chart, we can conclude that as age progresses from 20 to 70, people tend to be more diabetic. Out of 396 people within age 20 - 30, only 21% are tested positive, whereas among people within age 30 – 40, 46% are tested positive and 55% tested +ve for people aged 40-50. Blood pressure and BMI also plays a major role to know the difference between those tested positive and those tested negative.

Tree Maps

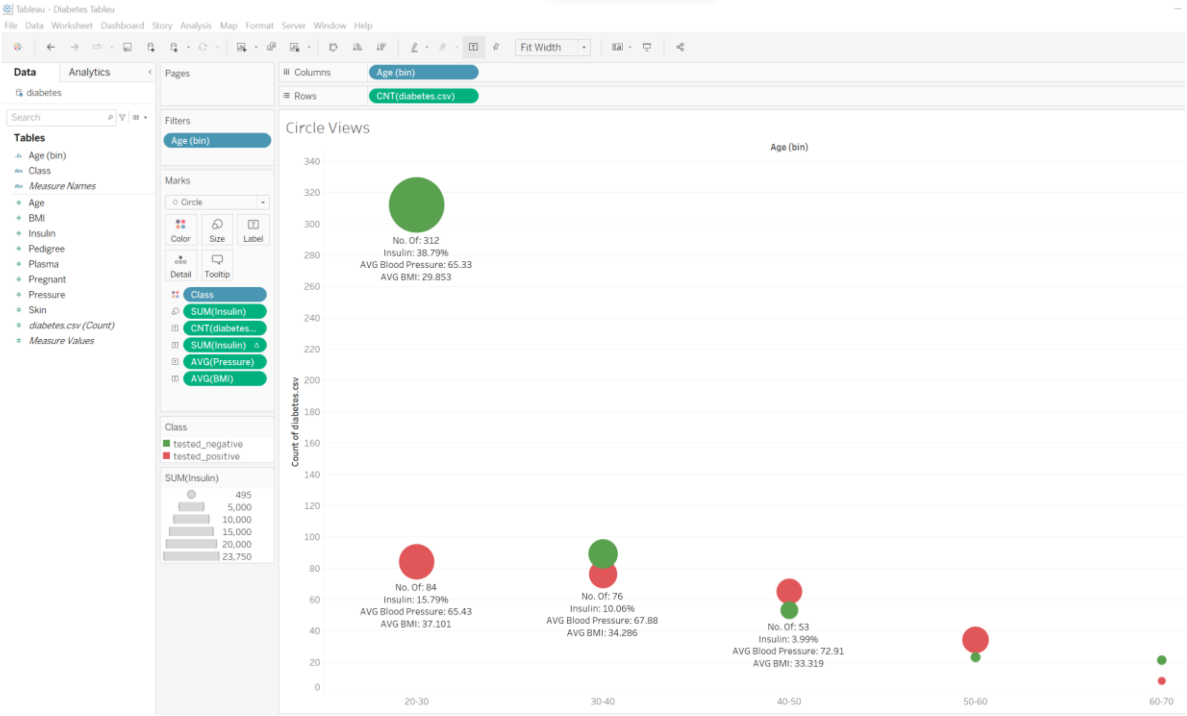
From the hypothesis above, we can acquire the resulted Tree Maps chart as below.



Same as stack chart, I have created a filter by removing the age of any person above the age 70+ as the data were inadequate. I have also added the marks with the given data as shown above to differentiate between those tested positive and negative with factors such as their age, BMI, Blood pressure and the insulin percentage.

Circle View

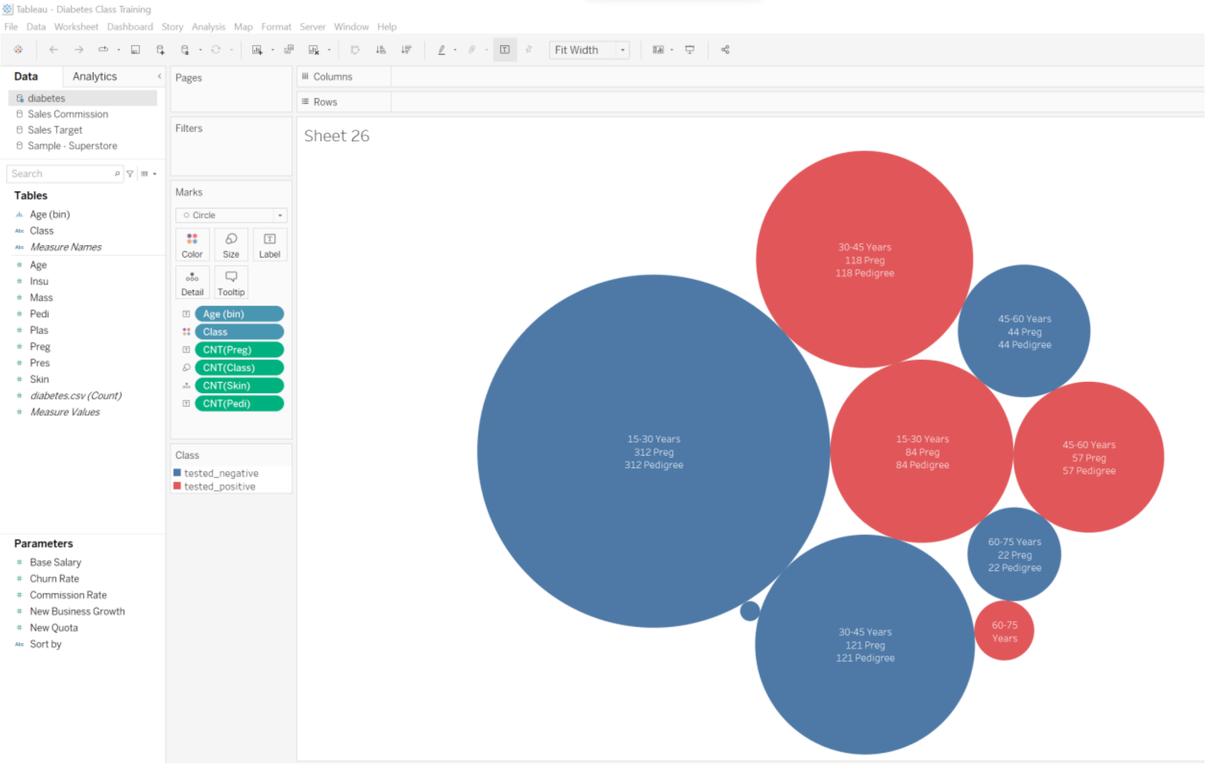
From the hypothesis above, we can acquire the resulted Circle View as below.



According to the above, I’ve used the sum of insulin as the circle size to calculate if insulin influences diabetes. From the above, we can see that, higher the insulin %, higher the no. of people tested negative or less diabetic. Other factors like BMI, Blood Pressure and count of people are input as an additional information.

Packed Bubble

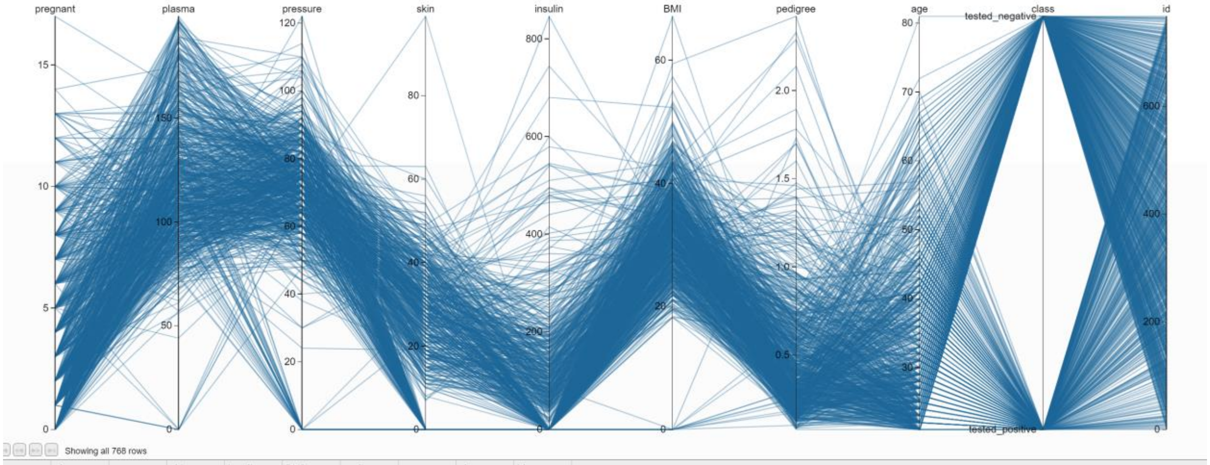
From the hypothesis above and additionally other factors were tested. We can acquire the resulted Packed Bubble View as below.



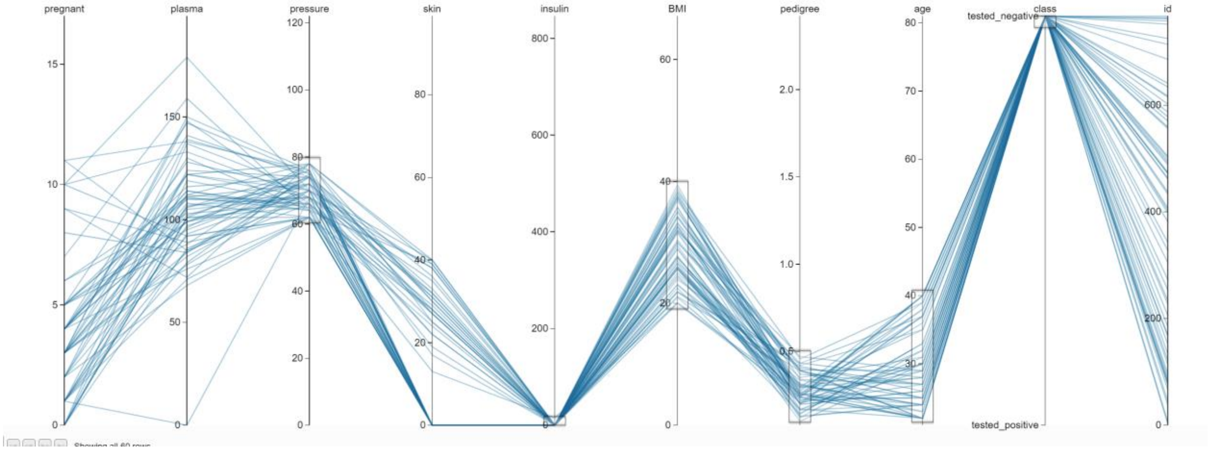
Additional factors apart from the hypothesis were tested to check if other factors such as being pregnant, or having thick skin, or having family pedigree history results in any change or make difference to our hypothesis. No Filters have been applied as the difference does not give us any trend to the chart.

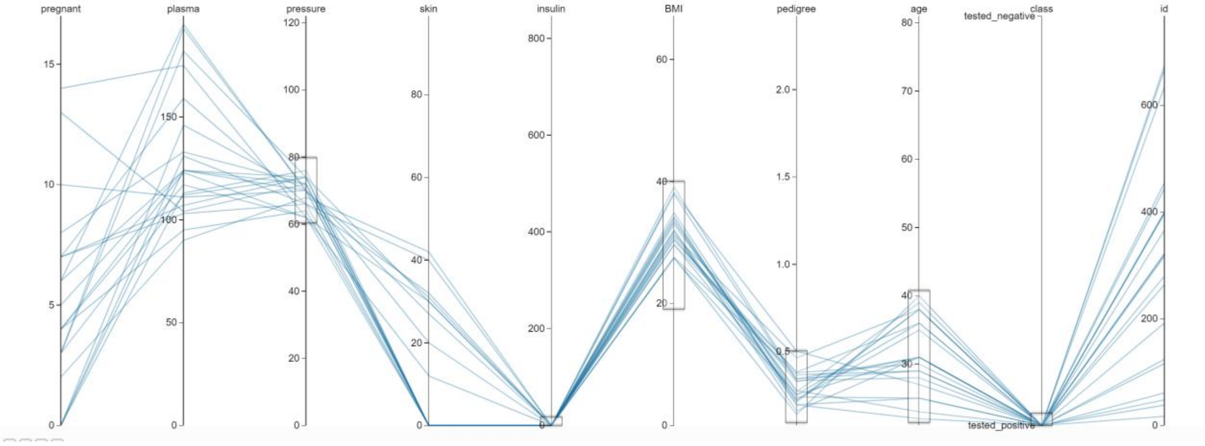
Parallel Coordinates

Overall data for Diabetes are attached in the below parallel coordinates



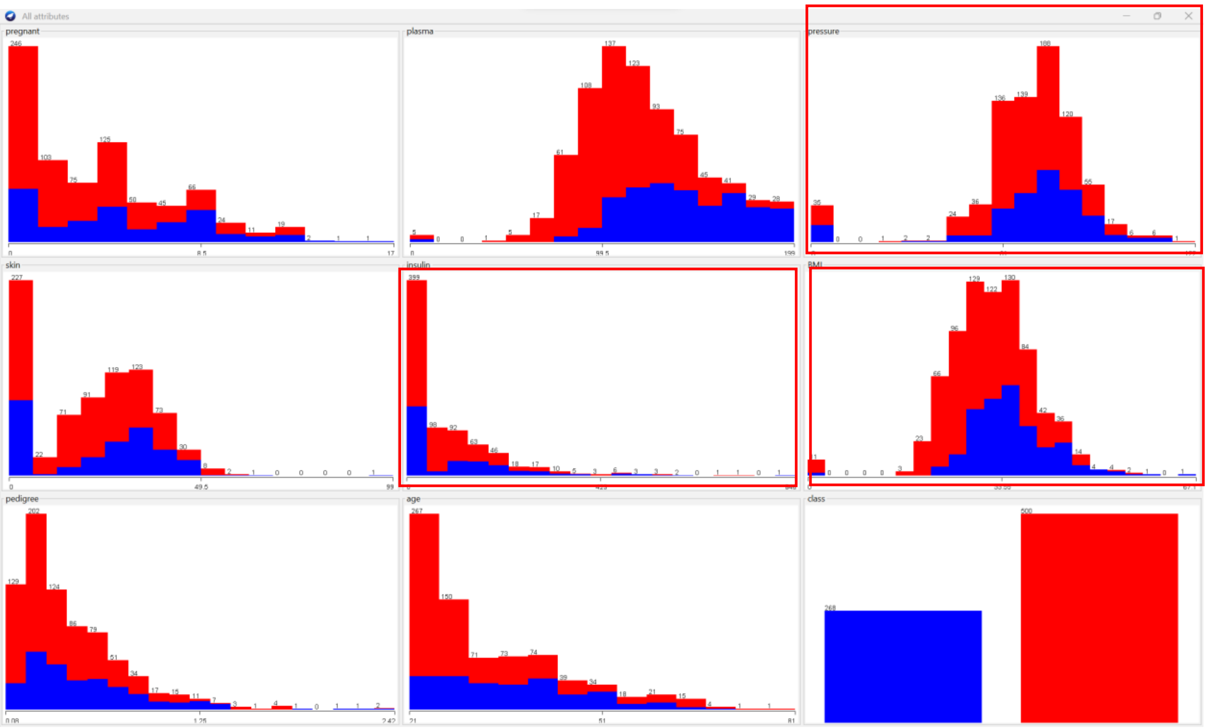
From this, I’ve analysed the data for both negative tested and positive tested along with age as 20 to 40, Family Pedigree details from 0 to 0.6, BMI from 20 to 40 with 0 insulin and 60 to 80 Bblood pressure and the resulted parallel coordinate graph is attached as below.





Weka

From the data acquired, I was able to identify multiple trends that people with higher BMI, higher Blood Pressure and people with no insulin are tested positive. Though we don't have enough data to conclude that with high BP or high BMI will be tested positive but from the data we are able to achieve the hypothesis we had earlier. Attached the weka chart below for reference.



2. Data Analytics & Ranking

Before all the analysis were performed, the data went through several data cleansing formats such as removing duplicates, removing missing values, removing numerical outliers. (None were found)

Additionally, I've added a value of tested positive '1' and tested negative as '2'.

Correlation Analysis

	pregnant	*plasma*	*pressure*	*skin*	*insulin*	*BMI*	*pedigree*	*age*	*class*
pregnant	1.00								
plasma	0.12	1.00							
pressure	0.14	0.15	1.00						
Skin	-0.08	0.06	0.21	1.00					
insulin	-0.07	0.33	0.09	0.44	1.00				
BMI	0.01	0.22	0.28	0.39	0.20	1.00			
pedigree	-0.03	0.14	0.04	0.18	0.19	0.14	1.00		
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	
Class	-0.22	-0.47	-0.07	-0.07	-0.13	-0.29	-0.17	-0.24	1.00

I have highlighted the good attributes in terms of correlation.

Covariance Analysis

	pregnant	plasma	pressure	skin	insulin	BMI	pedigree	age	class
pregnant	11.34								
plasma	13.93	1020.92							
pressure	9.20	94.31	374.16						
skin	-4.38	29.20	63.95	254.14					
insulin	-28.52	1219.35	198.12	801.93	13263.89				
BMI	0.47	55.65	42.95	49.31	179.54	62.08			
pedigree	-0.04	1.45	0.26	0.97	7.06	0.37	0.11		
age	21.54	98.95	54.45	-21.35	-57.07	3.36	0.13	138.12	
class	-0.36	-7.11	-0.60	-0.57	-7.17	-1.10	-0.03	-1.34	0.23

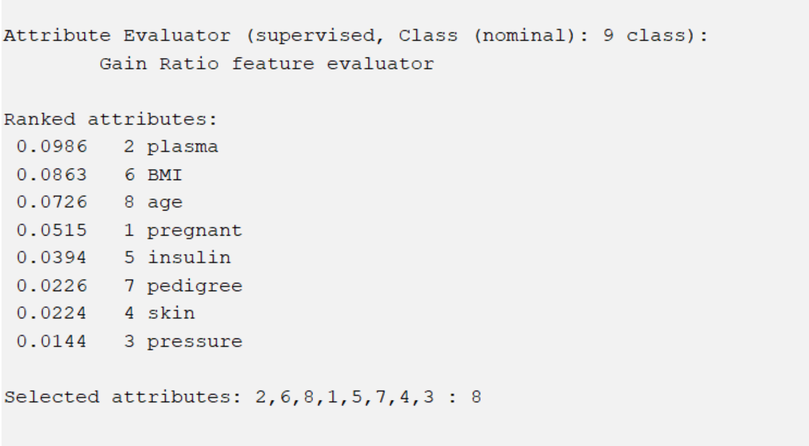
I have highlighted the good attributes in terms of covariance.

Difference between correlation and covariance.

Covariance indicates the direction of the linear relationship between variables whereas Correlation measures both the strength and direction of the linear relationship between two variables. Covariance shows you how the two variables differ whereas correlation shows you how the two variables are related.

Correlation is better than covariance analysis because correlation removes the effects of the variance of the variables as it provides a standardized absolute measure of the strength of the relationship bounded by -1.0 and 1.0.

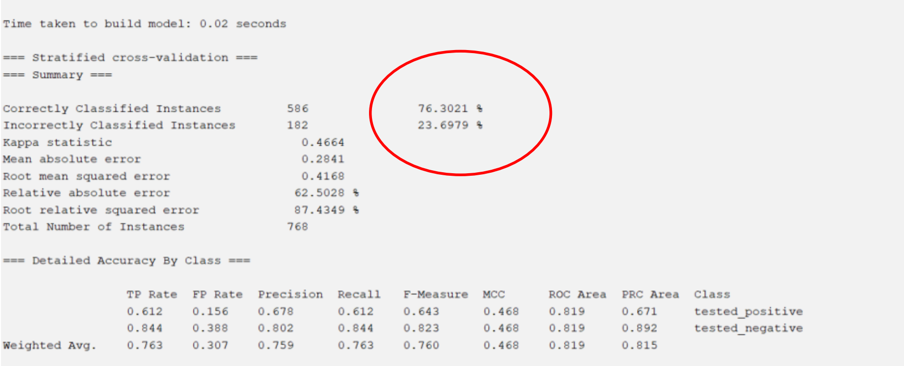
Weka



```
Attribute Evaluator (supervised, Class (nominal): 9 class):
        Gain Ratio feature evaluator

Ranked attributes:
 0.0986   2 plasma
 0.0863   6 BMI
 0.0726   8 age
 0.0515   1 pregnant
 0.0394   5 insulin
 0.0226   7 pedigree
 0.0224   4 skin
 0.0144   3 pressure

Selected attributes: 2,6,8,1,5,7,4,3 : 8
```



```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4664
Mean absolute error                      0.2841
Root mean squared error                  0.4168
Relative absolute error                 62.5028 %
Root relative squared error             87.4349 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.612    0.156    0.678      0.612    0.643      0.468    0.819     0.671     tested_positive
                 0.844    0.388    0.802      0.844    0.823      0.468    0.819     0.892     tested_negative
Weighted Avg.    0.763    0.307    0.759      0.763    0.760      0.468    0.819     0.815
```

I’ve analysed the data and ranked using WEKA. These ranked attributes based on the ranking gives me 76% correct ranking attribute. The ranking is as attached as above.

ANOVA

This test is based on the Anova Single Factor analysis to analyse the top Factor found in WEKA.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
insulin	400	34602	86.505	13907.87
class	400	600	1.5	0.250627

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1445170	1	1445170	207.8167	4.84499E-42	3.853138126
Within Groups	5549340	798	6954.06			
Total	6994510	799				

From this we can understand that, Insulin plays a major role in the data analysis as the F critical value is lower than the P-value as shown in the above data.

ANOVA Two Factor Analysis

Anova: Two-Factor With Replication

SUMMARY	insu	class	Total
72			
Count	200	200	400
Sum	20433	200	20633
Average	102.165	1	51.5825
Variance	18831.58	0	11957.19
76			
Count	200	200	400
Sum	14169	400	14569
Average	70.845	2	36.4225
Variance	8561.117	0	5457.708
Total			
Count	400	400	
Sum	34602	600	
Average	86.505	1.5	
Variance	13907.87	0.250627	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	45965.12	1	45965.12	6.712027	0.009752	3.853168
Columns	1445170	1	1445170	211.03	1.38E-42	3.853168
Interaction	52229.12	1	52229.12	7.626723	0.005883	3.853168
Within	5451146	796	6848.173			
Total	6994510	799				

I have done an ANOVA on insulin and Blood pressure comparing them with the Class (Positive and Negative). P value should be higher than the F critical value, however this was not achieved in the Anova above as compared to the charts.