APPLIED STATISTICAL METHODS
ASSIGNMENT REPORT

by Achyuta Krishna V
2018A7PS0165H

Aim :

The aim of the project is to analyse a given dataset using the statistical techniques learnt and make valid inferences.

Dataset used :

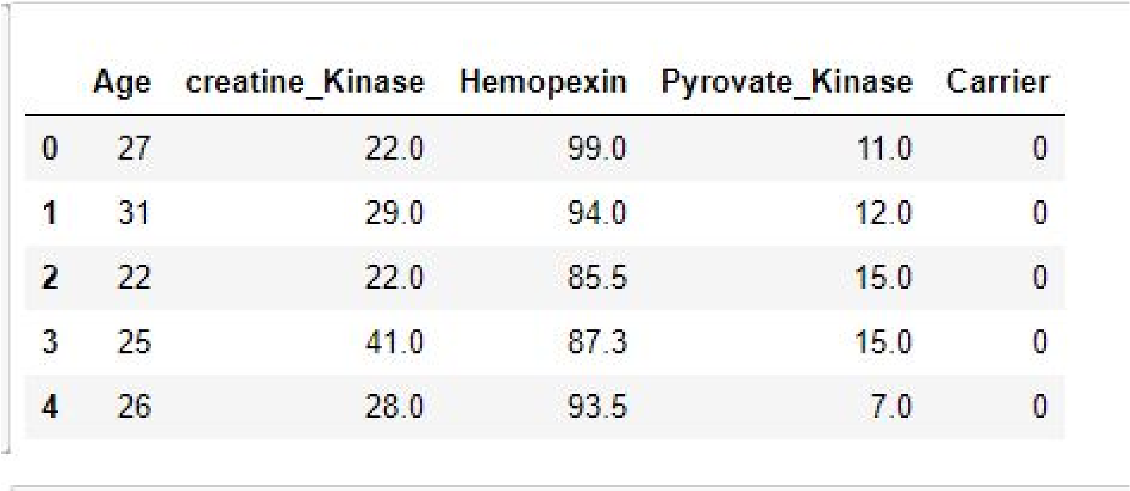
The dataset used was the ‘Muscular Data_Project_10.xlsx’ corresponding to Project 10.

Description of the dataset :

The dataset consists of 5 attributes, which are, 'Age', 'creatine_Kinase', 'Hemopexin', 'Pyrovate_Kinase' and ‘Carrier’.

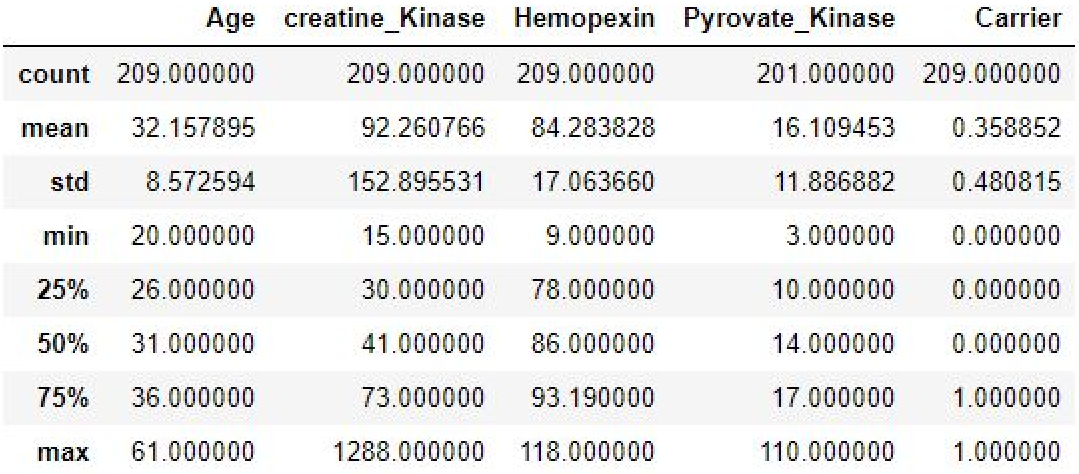
'Age', 'creatine_Kinase', 'Hemopexin' and 'Pyrovate_Kinase' are numerical attributes while ‘Carrier’ is a binary categorical attribute taking values ‘1’ (Carrier) or ‘0’ (not a carrier).

The dataset consists of 209 data points, with 134 data points having a value of ‘0’ for the attribute ‘Carrier’ and the remaining 75 having a value of ‘1’ for ‘Carrier’.



	Age	creatine_Kinase	Hemopexin	Pyrovate_Kinase	Carrier
0	27	22.0	99.0	11.0	0
1	31	29.0	94.0	12.0	0
2	22	22.0	85.5	15.0	0
3	25	41.0	87.3	15.0	0
4	26	28.0	93.5	7.0	0

Fig 1: Table consisting of the first 5 rows of the dataset

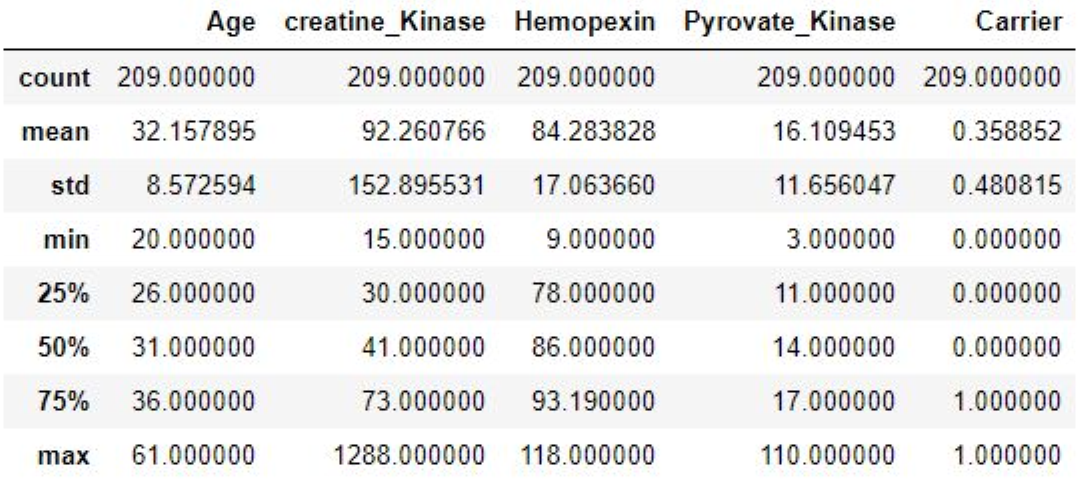


	Age	creatine_Kinase	Hemopexin	Pyrovate_Kinase	Carrier
count	209.000000	209.000000	209.000000	201.000000	209.000000
mean	32.157895	92.260766	84.283828	16.109453	0.358852
std	8.572594	152.895531	17.063660	11.886882	0.480815
min	20.000000	15.000000	9.000000	3.000000	0.000000
25%	26.000000	30.000000	78.000000	10.000000	0.000000
50%	31.000000	41.000000	86.000000	14.000000	0.000000
75%	36.000000	73.000000	93.190000	17.000000	1.000000
max	61.000000	1288.000000	118.000000	110.000000	1.000000

Fig 2: Table showing the summary statistics of all the attributes of the dataset. The statistical quantities shown are the count, mean, standard deviation, minimum value, 25th percentile value, 50th percentile value, 75th percentile value and the maximum value for each attribute respectively.

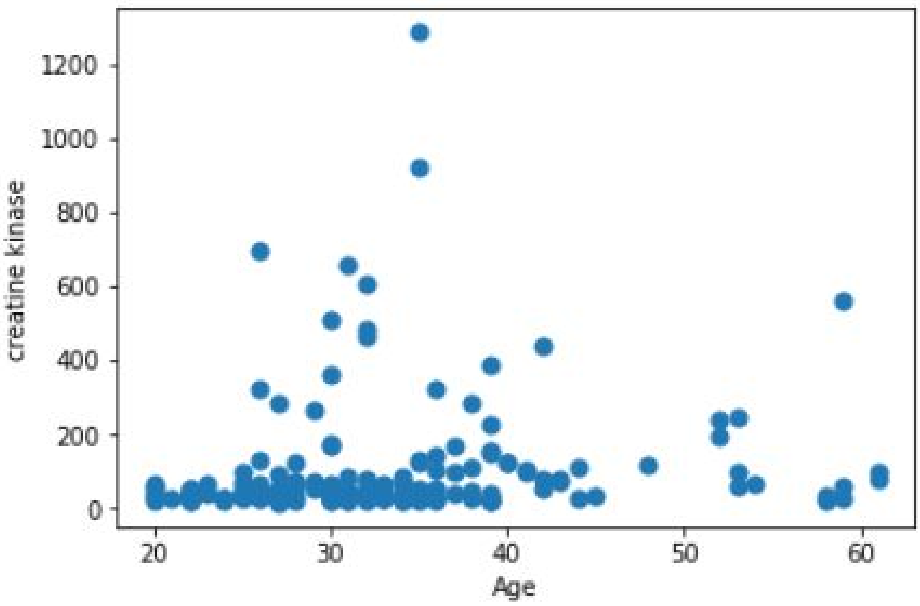
Data Preprocessing :

The data summarization process revealed that the attribute ‘Pyrovate_Kinase’ has only 201 values (ie.) 8 values were missing. The missing values were filled with the mean of the remaining values in that column of the dataset.

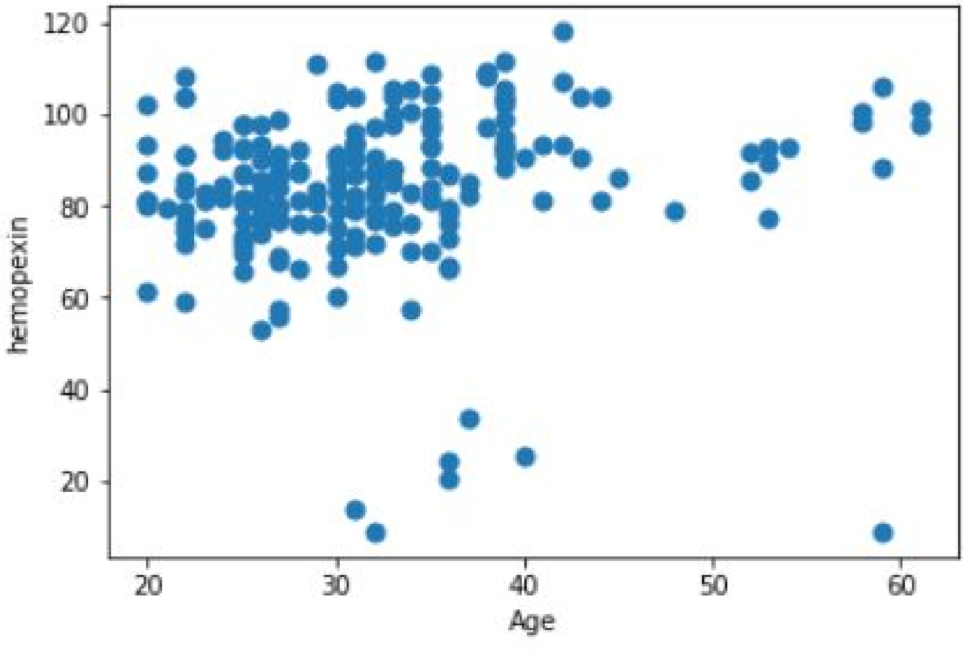


	Age	creatine_Kinase	Hemopexin	Pyrovate_Kinase	Carrier
count	209.000000	209.000000	209.000000	209.000000	209.000000
mean	32.157895	92.260766	84.283828	16.109453	0.358852
std	8.572594	152.895531	17.063660	11.656047	0.480815
min	20.000000	15.000000	9.000000	3.000000	0.000000
25%	26.000000	30.000000	78.000000	11.000000	0.000000
50%	31.000000	41.000000	86.000000	14.000000	0.000000
75%	36.000000	73.000000	93.190000	17.000000	1.000000
max	61.000000	1288.000000	118.000000	110.000000	1.000000

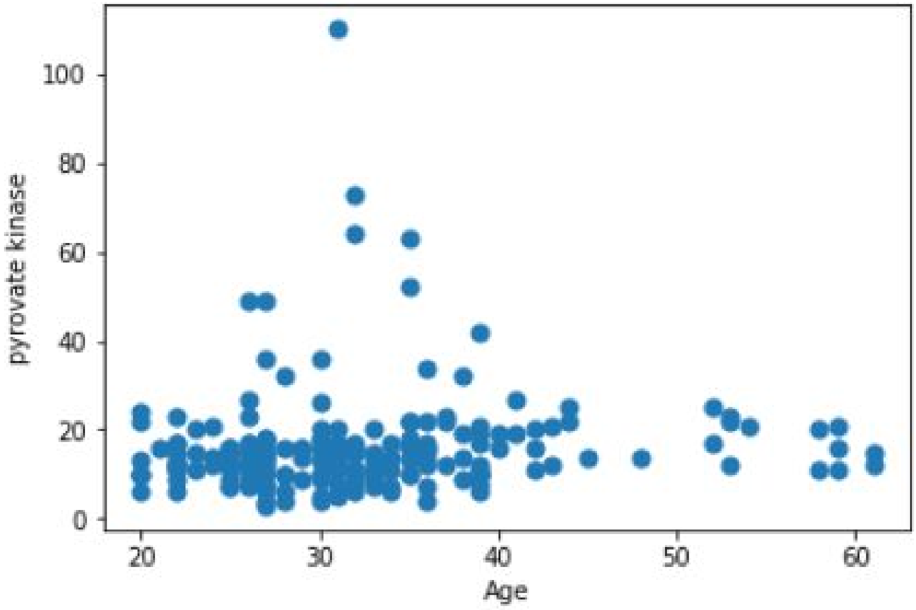
Fig 3: Summary statistics of the dataset after filling the missing values with the corresponding column mean. It is to be noted that the count statistic for the attribute ‘Pyrovate_Kinase’ has changed from 201 to 209.



Plot of Age vs creatine kinase



Plot of Age vs hemopexin



Plot of Age vs pyrovate kinase

Statistical tests performed :

1. **Student’s t-test**

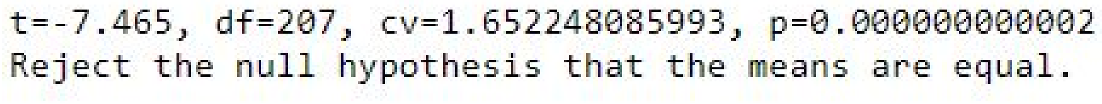
   The student’s t-test is a hypothesis test that tests the equality of the means of an independent attribute across 2 groups. It in turn tests the significance of that independent attribute in deciding which of the 2 groups the data point belongs to. In our dataset, the 2 groups are Carrier = 0 and Carrier = 1. The independent variables whose significance is to be tested are 'Age', 'creatine_Kinase', 'Hemopexin' and 'Pyrovate_Kinase'. The t-test is performed separately considering each independent variable separately.

   The t-test is typically used to test for the equality of means across 2 groups. If 3 or more groups exist, the Analysis of Variance (ANOVA) is used. Since there are only 2 categories for the attribute ‘Carrier’, the t-test is used.

   a) t-test with ‘Age’ as the independent attribute

   Let μ_1 be the mean age of data points in group 1 (Carrier = 0)
   μ_2 be the mean age of data points in group 2 (Carrier = 1)

   Null hypothesis $H_0 : \mu_1 = \mu_2$
   Alternative hypothesis $H_a : \mu_1 \neq \mu_2$
   Level of significance $\alpha = 0.05$



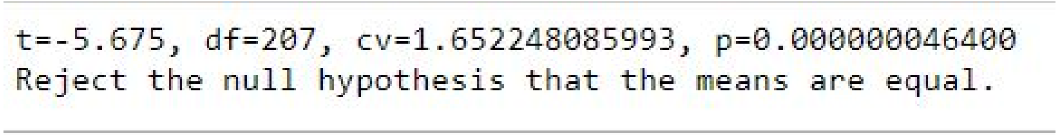
```
t=-7.465, df=207, cv=1.652248085993, p=0.000000000002
Reject the null hypothesis that the means are equal.
```

The result of the t-test obtained is that the null hypothesis is to be rejected. Hence we can conclude that the mean age differs significantly across the 2 groups and hence age is a significant attribute in determining whether a person is a carrier or not.

b) t-test with ‘creatine_Kinase’ as the independent attribute

Let μ_1 be the mean value of creatine kinase of data points in group 1 (Carrier = 0)
μ_2 be the mean value of creatine kinase of data points in group 2 (Carrier = 1)

Null hypothesis	$H_0 : \mu_1 = \mu_2$
Alternative hypothesis	$H_a : \mu_1 \neq \mu_2$
Level of significance	$\alpha = 0.05$



```
t=-5.675, df=207, cv=1.652248085993, p=0.000000046400
Reject the null hypothesis that the means are equal.
```

The result of the t-test obtained is that the null hypothesis is to be rejected. Hence we can conclude that the mean value of creatine kinase differs significantly across the 2 groups and hence the value of creatine kinase is a significant attribute in determining whether a person is a carrier or not.

c) t-test with ‘Pyrovate_Kinase’ as the independent attribute

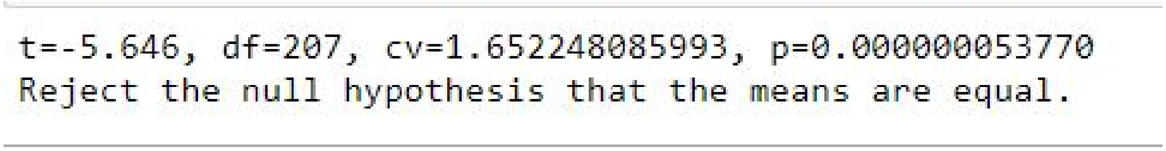
Let μ_1 be the mean value of pyrovate kinase of data points in group 1 (Carrier = 0)

μ_2 be the mean value of pyrovate kinase of data points in group 2 (Carrier = 1)

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_a : \mu_1 \neq \mu_2$

Level of significance $\alpha = 0.05$



```
t=-5.646, df=207, cv=1.652248085993, p=0.000000053770
Reject the null hypothesis that the means are equal.
```

The result of the t-test obtained is that the null hypothesis is to be rejected. Hence we can conclude that the mean value of pyrovate kinase differs significantly across the 2 groups and hence the value of pyrovate kinase is a significant attribute in determining whether a person is a carrier or not.

d) t-test with ‘Pyrovate_Kinase’ as the independent attribute

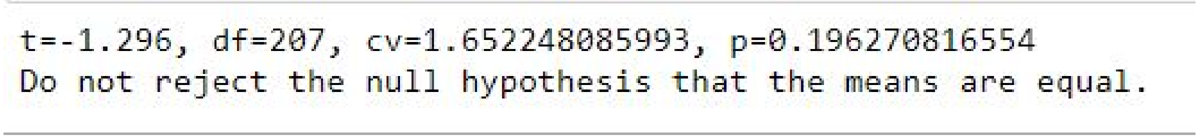
Let μ_1 be the mean value of hemopexin of data points in group 1 (Carrier = 0)

μ_2 be the mean value of hemopexin of data points in group 2 (Carrier = 1)

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_a : \mu_1 \neq \mu_2$

Level of significance $\alpha = 0.05$



```
t=-1.296, df=207, cv=1.652248085993, p=0.196270816554
Do not reject the null hypothesis that the means are equal.
```

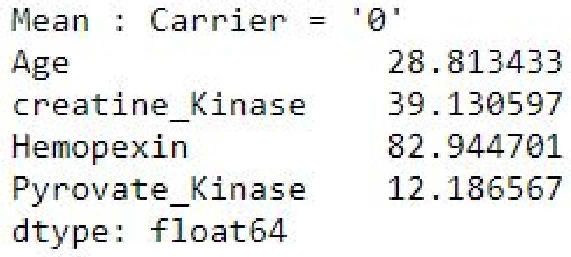
The result of the t-test obtained is that the null hypothesis is not to be rejected. Hence we can conclude that the mean value of hemopexin does not differ significantly across the 2 groups and hence the value of hemopexin is not a significant attribute in determining whether a person is a carrier or not.

2. Hotelling T-2 test

The Hotelling T-2 test is the multivariate counterpart of the t-test. A two-sample Hotelling's T^2 test is used to test for significant differences between the mean vectors (multivariate means) of two multivariate data sets (groups).

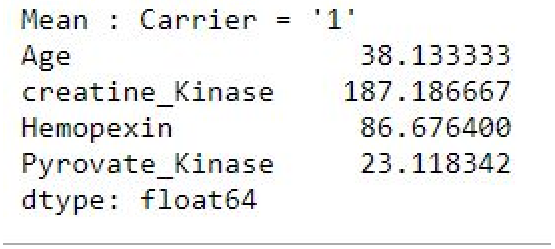
Let μ_1 be the population vector of means of group 1 (Carrier = 0)
μ_2 be the population vector of means of group 2 (Carrier = 1)

Null hypothesis $H_0 : \mu_1 = \mu_2$
Alternative hypothesis $H_a : \mu_1 \neq \mu_2$
Level of significance $\alpha = 0.05$



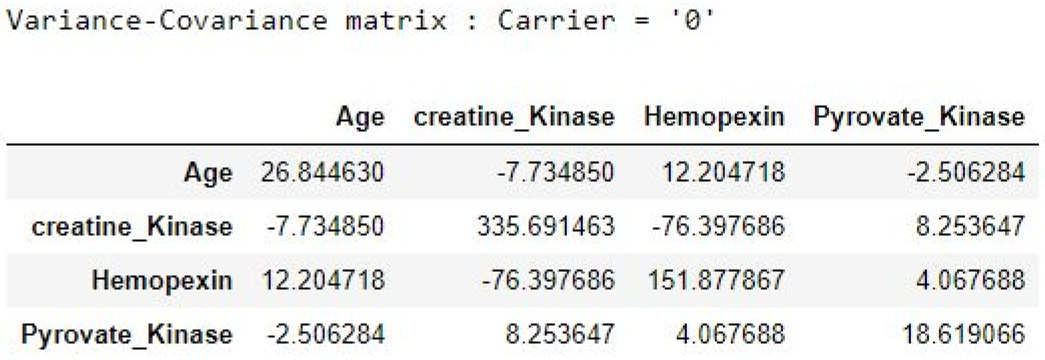
```
Mean : Carrier = '0'
Age                  28.813433
creatine_Kinase      39.130597
Hemopexin            82.944701
Pyrovate_Kinase      12.186567
dtype: float64
```

x_1-bar : The vector of means of data points in group 1 (Carrier = 0)



```
Mean : Carrier = '1'
Age                   38.133333
creatine_Kinase      187.186667
Hemopexin             86.676400
Pyrovate_Kinase       23.118342
dtype: float64
```

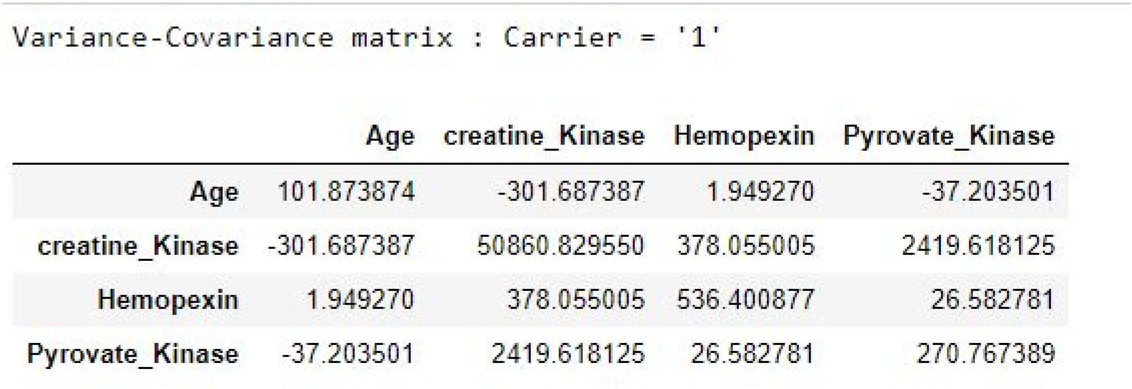
x_2-bar : The vector of means of data points in group 2 (Carrier = 1)



Variance-Covariance matrix : Carrier = '0'

	Age	creatine_Kinase	Hemopexin	Pyrovate_Kinase
Age	26.844630	-7.734850	12.204718	-2.506284
creatine_Kinase	-7.734850	335.691463	-76.397686	8.253647
Hemopexin	12.204718	-76.397686	151.877867	4.067688
Pyrovate_Kinase	-2.506284	8.253647	4.067688	18.619066

S_1 : Variance-Covariance matrix for group 1 (Carrier = 0)

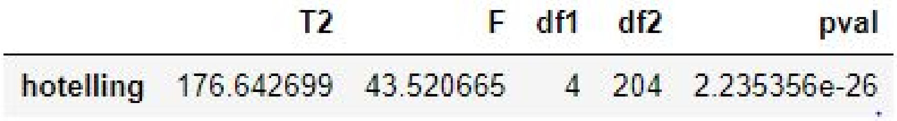


```
Variance-Covariance matrix : Carrier = '1'
```

	Age	creatine_Kinase	Hemopexin	Pyrovate_Kinase
Age	101.873874	-301.687387	1.949270	-37.203501
creatine_Kinase	-301.687387	50860.829550	378.055005	2419.618125
Hemopexin	1.949270	378.055005	536.400877	26.582781
Pyrovate_Kinase	-37.203501	2419.618125	26.582781	270.767389

S_2 : Variance-Covariance matrix for group 2 (Carrier = 1)

Results obtained on performing Hotelling T-2 test at α = 0.05 level of significance :



	T2	F	df1	df2	pval
hotelling	176.642699	43.520665	4	204	2.235356e-26

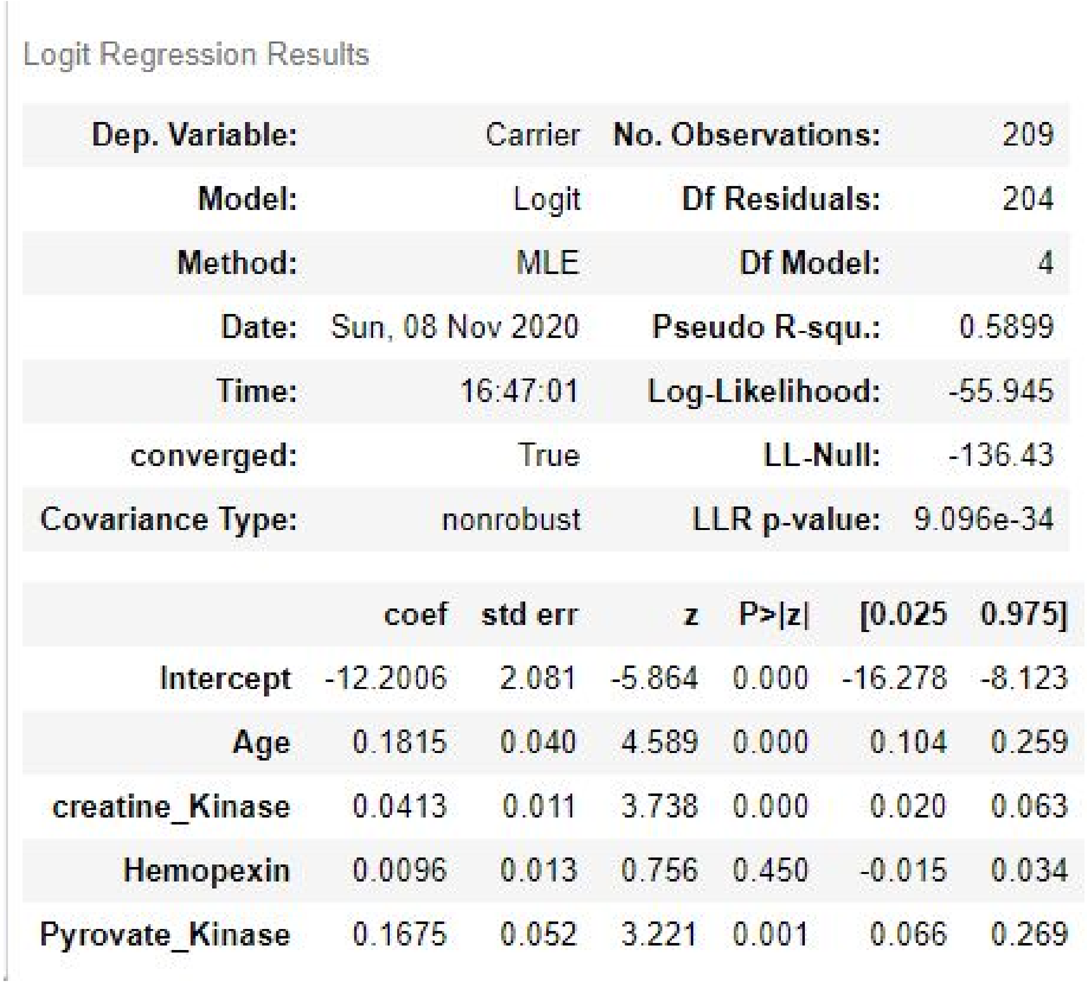
Since the p value obtained < 0.05 (level of significance α), the null hypothesis that the population mean vectors for the two groups are equal is rejected. Hence it can be concluded that there is a significant difference between carriers and non-carriers with respect to the combination of attributes ‘Age’, ‘creatine_Kinase’, ‘Pyrovate_Kinase’ and ‘Hemopexin’ (the 2 groups differ at least in one of the attributes). Using the results of the t-tests performed earlier it can be concluded that the difference between carriers and non-carriers occurs mainly upon the attributes ‘Age’, ‘creatine_Kinase’ and ‘Pyrovate_Kinase’.

3. Logistic regression

Logistic regression is a supervised learning algorithm that is used to perform binary classification on data. It is based on the probability of an event happening versus it not happening (odds of an event happening).

In our dataset, we want to predict whether a person is a carrier or not based on values of the other independent attributes age, creatine kinase, pyrovate kinase and hemopexin.

The logistic regression algorithm was applied and the following results were obtained.



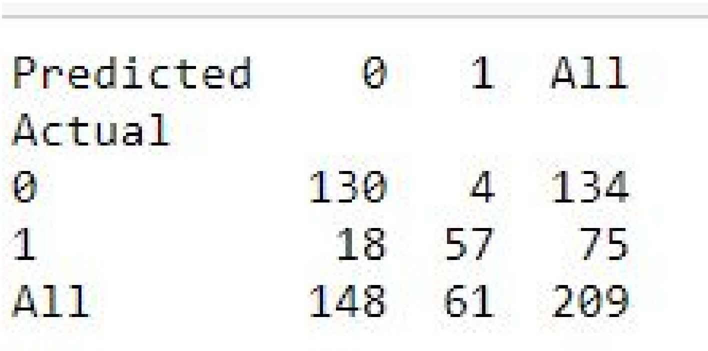
Logit Regression Results

Dep. Variable:	Carrier	**No. Observations:**	209
Model:	Logit	**Df Residuals:**	204
Method:	MLE	**Df Model:**	4
Date:	Sun, 08 Nov 2020	**Pseudo R-squ.:**	0.5899
Time:	16:47:01	**Log-Likelihood:**	-55.945
converged:	True	**LL-Null:**	-136.43
Covariance Type:	nonrobust	**LLR p-value:**	9.096e-34

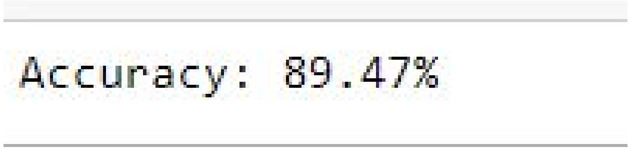
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-12.2006	2.081	-5.864	0.000	-16.278	-8.123
Age	0.1815	0.040	4.589	0.000	0.104	0.259
creatine_Kinase	0.0413	0.011	3.738	0.000	0.020	0.063
Hemopexin	0.0096	0.013	0.756	0.450	-0.015	0.034
Pyrovate_Kinase	0.1675	0.052	3.221	0.001	0.066	0.269

The result table obtained depicts the coefficients of all the independent attributes in the model, the standard error in calculating the coefficients, the z-statistic , the corresponding p-value and the lower and upper limits of a 95% confidence interval.

From the p-value obtained, it can be concluded that at a significance level of 0.05, the attributes ‘Age’, ‘creatine_Kinase’ and ‘Pyrovate_Kinase’ are significant in predicting the value of the attribute ‘Carrier’ whereas the attribute ‘Hemopexin’ is not significant. Thus, the results obtained from logistic regression are in accordance with our previous results.



```
Predicted    0   1  All
Actual
0          130   4  134
1           18  57   75
All        148  61  209
```



```
Accuracy: 89.47%
```

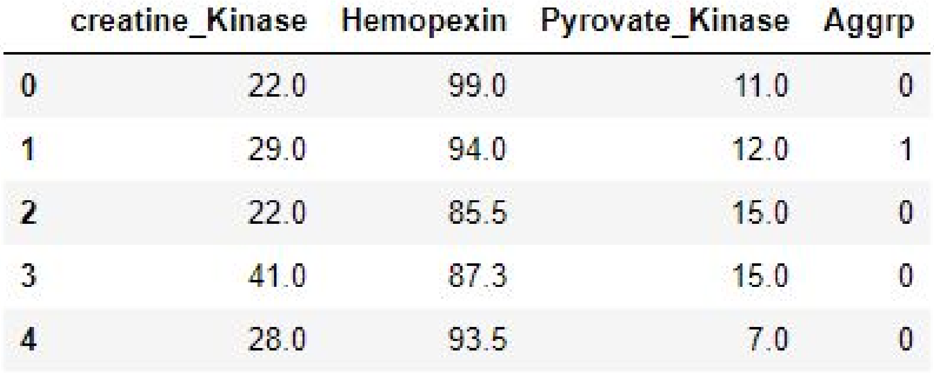
The model obtained was used on the dataset and a confusion matrix was generated. The confusion matrix shows us that 130 out of 134 data points with Carrier = 0 were classified correctly. Similarly, 57 out of the 75 data points with Carrier = 1 were classified correctly. Thus, the accuracy of the model on our dataset is 89.47%.

4. Linear Discriminant Analysis

Fischer’s Linear Discriminant Analysis is a classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes’ rule.

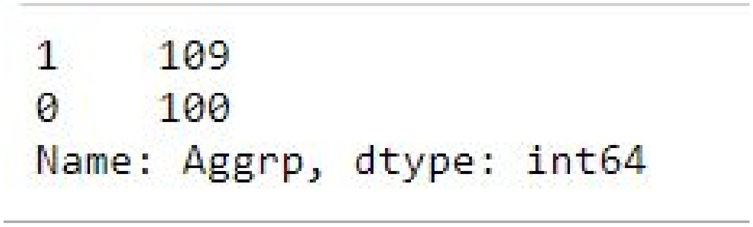
In this project, we intend to use the LDA to predict the age-category of a person given the values for the attributes creatine kinase, hemopexin and pyrovate kinase. In order to convert the age attribute into a categorical attribute age group (aggrp), the ages are divided into 2 categories, age <= 30 (category 0) and age > 30 (category 1).

The dataset is thus modified by including a new attribute aggrp denoting the age group and dropping the attributes Age and Carrier.



	creatine_Kinase	Hemopexin	Pyrovate_Kinase	Aggrp
0	22.0	99.0	11.0	0
1	29.0	94.0	12.0	1
2	22.0	85.5	15.0	0
3	41.0	87.3	15.0	0
4	28.0	93.5	7.0	0

The first five rows of the modified table.

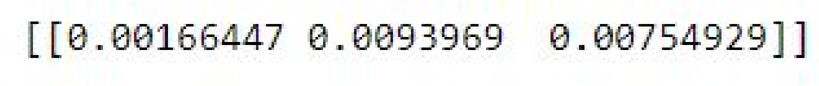


```
1    109
0    100
Name: Aggrp, dtype: int64
```

An analysis of the modified dataset shows that there are 100 data points belonging to group 0 and 109 data points belonging to group 1.

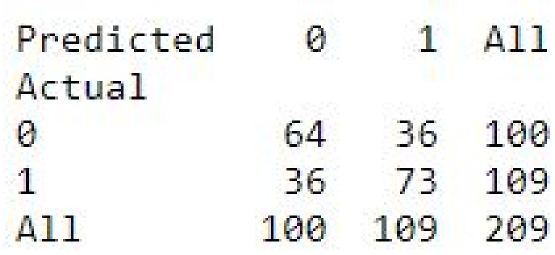
The LDA is applied on this modified dataset and the following results were obtained.

The model coefficients :



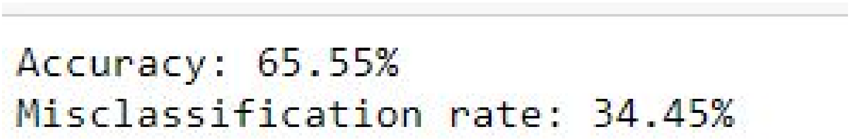
```
[[0.00166447 0.0093969  0.00754929]]
```

The confusion matrix



```
Predicted    0    1  All
Actual
0           64   36  100
1           36   73  109
All        100  109  209
```

The confusion matrix gives us the conclusion that 64 out of 100 data points with Aggrp = 0 were classified correctly. Similarly, 73 out of the 109 data points with Aggrp = 1 were classified correctly.



```
Accuracy: 65.55%
Misclassification rate: 34.45%
```

The accuracy of the model in predicting the age group of a person give the values for the attributes creatine kinase, pyrovate kinase and hemopexin is 65.55%. Hence the misclassification rate obtained is 34.45%. This high value of misclassification rate suggests that age groups are not a very effective indicator of the creatine kinase, pyrovate kinase and hemopexin levels.

5. K-means clustering algorithm

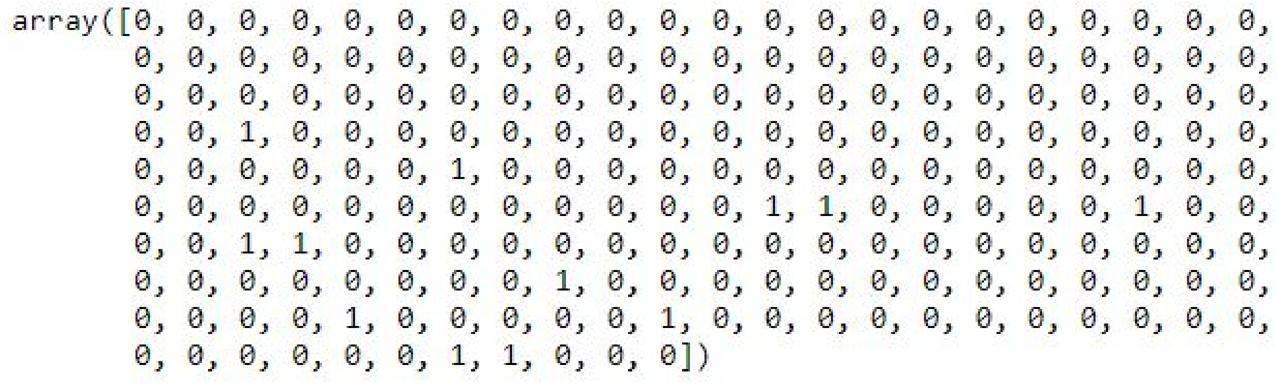
K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

In this project, we intend to use k-means clustering to check the following

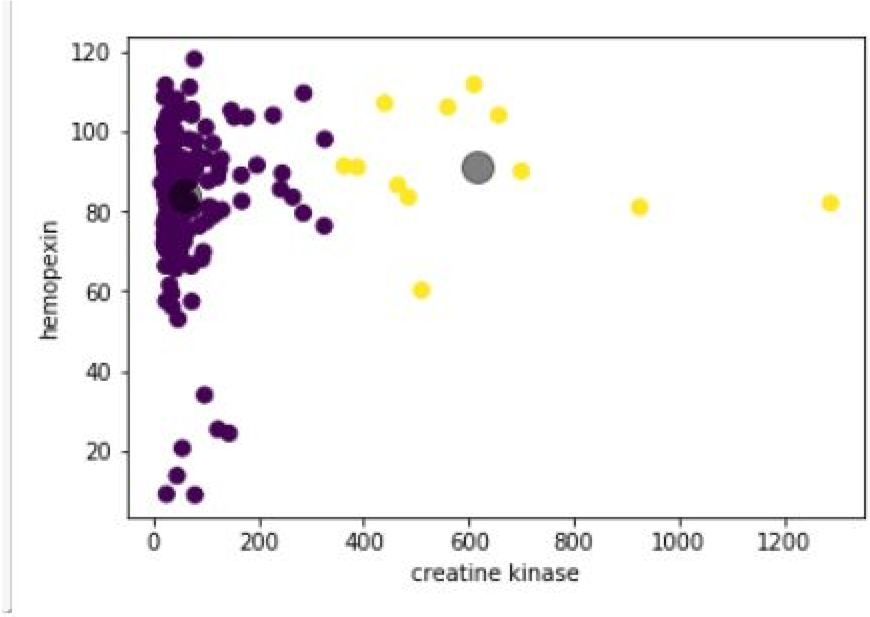
a) Whether age has an effect on creatine kinase, pyrovate kinase and hemopexin levels.
b) Whether k-means is able to detect the groups corresponding to Carrier = 0 and Carrier = 1 in the dataset consisting of creatine kinase, pyrovate kinase, hemopexin and age.

a) Analysing age groups

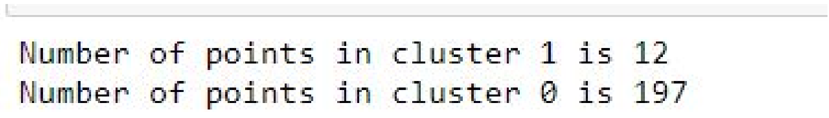
The results of the k-means clustering are as follows.



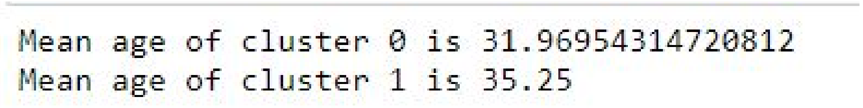
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0])
```



A sample plot of creatine kinase vs hemopexin showing the 2 clusters



```
Number of points in cluster 1 is 12
Number of points in cluster 0 is 197
```

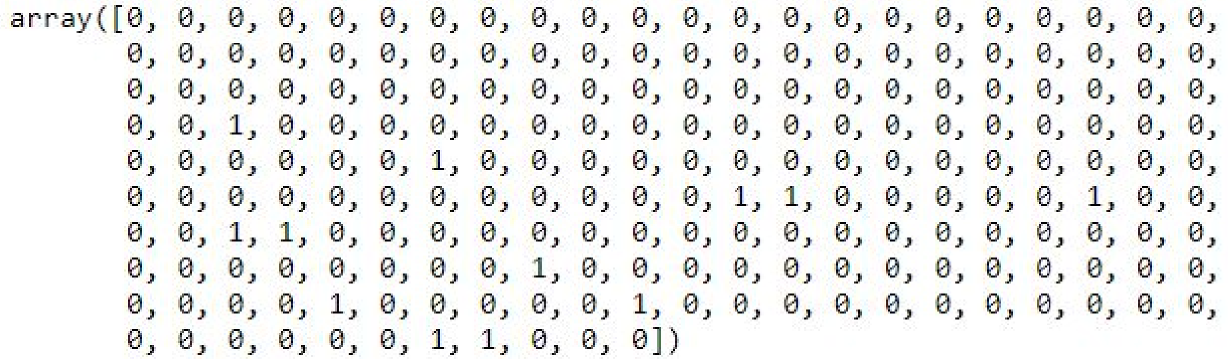


```
Mean age of cluster 0 is 31.96954314720812
Mean age of cluster 1 is 35.25
```

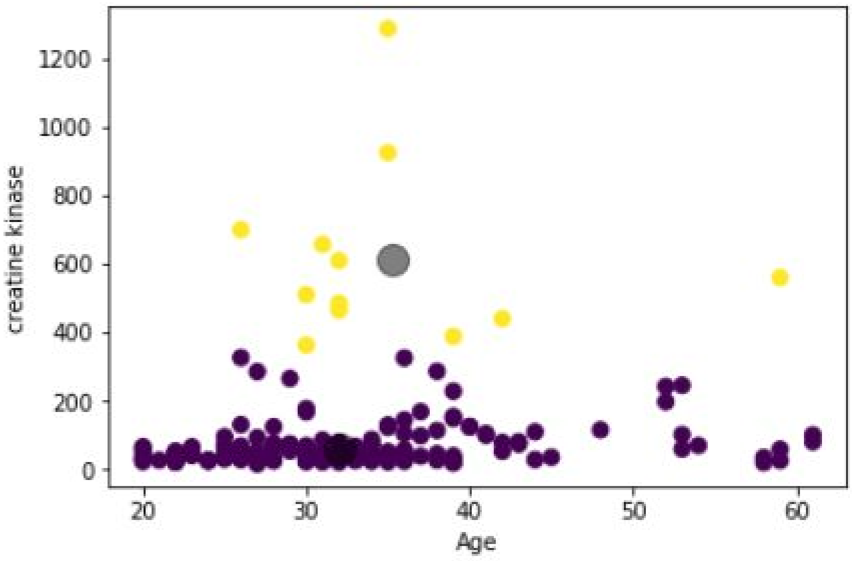
The 2 clusters obtained have 12 and 197 points respectively. It was calculated that the average age of the 2 clusters are 31.96 and 35.25 respectively. Since the mean age of all the identified clusters does not have a significant difference, it can be concluded that age does not have a significant effect on the levels of creatine kinase, pyrovate kinase and hemopexin which is in agreement with our previously obtained results.

b) Analysing whether a person is a carrier or not

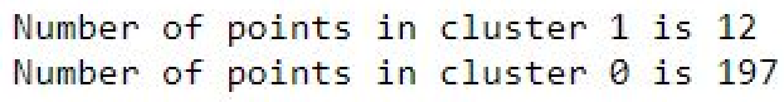
The results of the k-mean clustering algorithm are as follows.



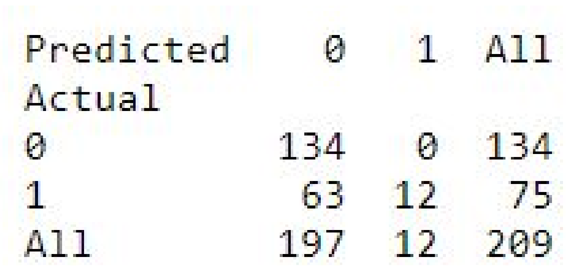
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0])
```



A sample plot of Age vs creatine kinase showing the 2 clusters



```
Number of points in cluster 1 is 12
Number of points in cluster 0 is 197
```



```
Predicted    0   1  All
Actual
0          134   0  134
1           63  12   75
All        197  12  209
```

The 2 clusters obtained have 12 and 197 points respectively. If we consider cluster 0 to represent the case Carrier = 0 and cluster 1 to Represent the case Carrier = 1, the confusion matrix obtained shows that all actual cases of Carrier = 0 is predicted as 0. However, among the 75 cases of Carrier = 1, only 12 were predicted to be 1. Hence the k-means does not detect cases of Carrier = 1 accurately.

Conclusion :

The dataset was analysed using Student’s t-test, Hotelling T-2 test, Logistic regression, Linear Discriminant analysis and K-means clustering.

The results of the Student’s t-test indicated that the attributes ‘Age’, ‘Creatine kinase’ and ‘pyrovate kinase’ are significant in determining whether a person is a carrier or not and the attribute ‘hemopexin’ is not significant for the same. The Hotelling T-sq test helps us to conclude that all the independent attributes are significant in determining the value of carrier attribute when taken together. Thus the combination of the independent attributes can be used to predict whether a person is a carrier or not.

Linear discriminant analysis was used to test whether the age group of a person is significant determining factor for the values of ‘Creatine kinase’ , ‘hemopexin’ and ‘pyrovate kinase. After acquiring a high misclassification rate of 34.45%, it was concluded that age group is not a very good determining factor of the values of the attributes e age group of a person is significant determining factor for the values of ‘Creatine kinase’ , ‘hemopexin’ and ‘pyrovate kinase.

K-means clustering was used to test two assumptions, whether age is significant in determining the values of Creatine kinase’ , ‘hemopexin’ and pyrovate kinase, and whether clustering is able to predict if a person is a carrier or not. The results obtained led us to conclude that age is not a very significant factor in determining the values of the various muscle enzymes. Also, it was found that clustering could not yield good results in classifying the data points into the 2 categories based on the ‘Carrier’ attribute. The reasons for it can be that the density spread of data points across the data space is different and the data points follow non-convex shapes