ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 5, Issue 8, August 2016

# Comparative Analysis of Liver Disorder Diagnosis Data Using Classification Techniques and Naïve Bayesian Algorithm of Neural Networks

**Pardeep Singh[1] (Student), Gagandeep Kaur[2] (Assistant Professor)**
[1](ECE, ACET Bhawanigarh/I.K.Gujral Punjab Technical University,Jalandhar, INDIA))
[2](ECE, ACET Bhawanigarh/I.K.Gujral Punjab Technical University,Jalandhar, INDIA))

***Abstract : has identified different association rules by using Apriori and made performance analysis with Naïve Bayesian of Neural Network algorithm for Liver Disorder Detection. There are two liver patients' data sets, USA patients and Indian patients. On the basis of common attributes experiments are conducted on data sets. Firstly, to check significance difference, experiments of ANOVA and MANOVA are conducted for the two different populations. Value of significance as null hypotheses is defined as 0.05 at 95% level of confidence. Then, Apriori and Naïve Bayesian algorithms are applied to the two data sets. During analysis of the two techniques, association rules generated by Apriori and confusion matrices are generated by Naïve Bayesian. At last performance of both is compared with each other and Neural Networks Provide more accurate results.***

***Keywords: ANOVA, MANOVA, Apriori, Naïve Bayesian, Neural Network, Liver Disorder.***

## I. INTRODUCTION

Patients with Liver diseases are increasing continuously day by day. These are caused by the too much use of alcohol; breathe in of injurious gases, eating of unhygienic foodstuff, pickles and drugs. Automatic tools are used to classify diseases. These tools may reduce burden on doctors. There are number of different algorithms that are used for the classification of different liver patient datasets [14]. Previously, sickness analysis uses arithmetical methods for modeling. In statistical methods, there are number of suppositions are used to evaluate linear data. So they are less competent to use in case of very big and complex nonlinear and reliant data. There are two data sets of Liver patients one is from US and other is from INDIA having different attributes. There is evaluation of frequent patterns by using Boolean association rules that can help for more accurate detection how many patients are the there. Applied methods are listed as below:

- ANOVA and MANOVA analysis of combined data set.
- ANOVA and MANOVA analysis of Liver Patient of UCI and India data set.
- ANOVA and MANOVA analysis of Liver Non Patient of UCI and India data set.
- Apriori algorithm.
- Naïve Bayesian algorithm.

## II. DATA SETS

There are two data sets that are in use from University of California at Irvine (UCI) Machine Learning Repository. USA data set contains 345 records of Liver patients with six attributes. India data set contains 583 records of Liver patient records taken from India with ten attributes. There are three familiar attributes (ALKPHOS, SGPT and SGOT) in both the datasets. These three attributes are used for the intention of contrast between both the data sets. Firstly, typical arithmetical methods one-way Analysis of Variance (ANOVA) and Multivariate Analysis of Variance (MANOVA) are applied to evaluate considerable difference between two populations for the categorization. After this, Apriori and Naïve Bayesian algorithms are applied to find strongly associated rules for the different values of minimum support and confidence.

## III. RELATED WORK

Mireille Tohm´et al [7] proposed an alternative to usual multiclass multivariate group comparison tests such as Hypothesis tests are used to compare and show the efficiency of drugs. Junning Li et al.[8] proposed a Dynamic Bayesian Networks (DBN)-based groupanalysis which combines the DBN approach and the multivariate analysis of variance (MANOVA). Neven Cukrov et al.[9] was applied multivariate statistical analysis to the measured physico-chemical parameters to estimate anthropogenic and natural influences to water system of the Krka River. Z. Haddi et al.[10] proposed Multivariate Analysis of Variance (MANOVA) to test the significance of the differences between cheeses groups. Z. A. Dastgheib et al. [11] applied multivariate analysis of variance (MANOVA) to select pairs of features showing the most

significant differences between the groups to get more classifier accuracy. S. Dimitrova [12] conducted MANOVA to check the significance of the influence of three different factors namely 1 planetary gcomagnctic activity Ievel estimated by Ap-index and divided into five levels, 2. gender - males and females and 3. the presence of medication. Paulo Ricardo Galhanone et al. [13] applied MANOVA and Discriminate analysis to Spectral analysis of the multichannel EEG of neonates is carried out with a view to determining differences in characteristics of High-Voltage-Slow, Low-Voltage-Irregular and Mixed EEG patterns. Diego Moitre, and Fernando Magnago [14] presented the application of themethodology of analysis of variance of multivariate data (MANOVA) to detect the impact of the fuel consumption on the market price. B.Surendiran et al.[15] proposed an Univariate Analysis of Variance (ANOVA) and Discriminate Analysis (DA) classifier for classifying the masses present in mammogram. Martha L. Zequera et al. [16] was designed to assess the effect of time on the repeatability of the LorAn pressure distribution measurement system, and evaluate the variability of plantar pressure and postural balance, during barefoot standing in diabetic and non-diabetic subjects, for future diabetic foot clinical evaluation. Benjamin F et al. [17] presented Directed canonical analysis as an extension of the general form of canonical analysis, which is a method for reducing the dimensionality of multivariate data sets with minimum loss of discriminatory variance. Aleksandar Jeremic et al. [18] developed a frequency-domain channel estimation algorithm for single-user multiantenna orthogonal frequency division multiplexing (OFDM) wireless systems in the presence of synchronous interference.

## IV. RESULTS AND DISCUSSION

### A. *ANOVA and MANOVA analysis of combined data set*

In this, we have all records of patients of the two populations. There are 345 records in UCI data set and 583 records in Indian data set. So, total numbers of records in this data set are 928. Firstly, Descriptive statistics of each individual attribute is done. Group 1 is used to denote UCI dataset and Group 2 is used to denote India data set.

Table 1: Descriptive Statistics of ALKPHOS

| ALKPHOS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 345 | 69.8696 | 18.34767 | .98781 | 67.9267 | 71.8125 | 23.00 | 138.00 |
| 2 | 583 | 2.9058E2 | 242.93799 | 10.061 | 270.8151 | 310.3375 | 63.00 | 2110.00 |
| Total | 928 | 2.0852E2 | 220.38146 | 7.23438 | 194.3271 | 222.7224 | 23.00 | 2110.00 |

Table 2: Descriptive Statistics of SGPT

| SGPT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 345 | 30.4058 | 18.51231 | 1.05051 | 27.3396 | 31.4720 | 4.00 | 155.00 |
| 2 | 583 | 80.7136 | 182.62036 | 7.56336 | 65.8587 | 95.5684 | 10.00 | 2000.00 |
| Total | 928 | 61.0108 | 146.21187 | 4.83247 | 51.5269 | 70.4946 | 4.00 | 2000.00 |

Table 3: Descriptive Statistics of SGOT

| SGOT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 345 | 24.6435 | 10.06449 | .54185 | 23.5777 | 25.7092 | 5.00 | 82.00 |
| 2 | 583 | 1.0991E2 | 288.91853 | 11.96578 | 86.4094 | 133.4122 | 10.00 | 4929.00 |
| Total | 928 | 77.2112 | 231.69093 | 7.63845 | 62.2205 | 91.2019 | 5.00 | 4929.00 |

Table 1, Table 2 and Table 3 shows the explanatory statistics for all the individual attributes ALKPHOS, SGPT and SGOT respectively.

After this, one-way ANOVA is applied for the three attributes ALKPHOS, SGPT and SGOT. The results of one way ANOVA are shown as below:

Table 4: One Way ANOVA on ALKPHOS between UCI and INDIA datasets

| ALKPHOS | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 1.0557739946177348E7 | 1 | 1.0557739946177348E7 | 283.665 | .000 |
| Within Groups | 3.446478348377956E7 | 926 | 37218.988643390454 | | |
| Total | 4.502252342995690 6E7 | 927 | | | |

Table 5: One way ANOVA on SGPT between UCI and INDIA

| SGPT | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 548541.540885325 | 1 | 548541.540885325 | 25.994323345805746 | .000 |
| Within Groups | 1.954078435135608E7 | 926 | 21102.359 | | |
| Total | 2.0089325892241407E7 | 927 | | | |

Table 6: One way ANOVA on SGOT between UCI and INDIA datasets

| SGOT | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 1575814.0937010911 | 1 | 1575814.0937010911 | 30.0144788928219 | .000 |
| Within Groups | 4.8616664509747185E7 | 926 | 52501.798 | | |
| Total | 5.019247860344828E7 | 927 | | | |

Significant values of Table 4, Table 5 and Table 6 is 0.0. So, null hypothesis is safely rejected. There is more significant difference between the two groups.

Now, descriptive statistics is calculated for the different combination of attributes ALKPHOS,SGPT, ALKPHOS,SGOT , SGPT,SGOT and ALKPHOS,SGPT,SGOT. Results are recorded as shown in tables below:

Table 7: Descriptive Statistics of ALKPHOS and SGPT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| ALKPHOS | 1 | 68.8696 | 17.34767 | 345 |
| | 2 | 2.9058E2 | 242.93799 | 583 |
| | Total | 2.0852E2 | 220.38146 | 928 |
| SGPT | 1 | 30.4058 | 19.51231 | 345 |
| | 2 | 79.7136 | 181.62036 | 583 |
| | Total | 61.0108 | 146.21187 | 928 |

Table 8: Descriptive Statistics of ALKPHOS and SGOT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| ALKPHOS | 1 | 68.8696 | 17.34767 | 345 |
| | 2 | 2.9058E2 | 242.93799 | 583 |
| | Total | 2.0852E2 | 220.38146 | 928 |
| SGOT | 1 | 30.4058 | 19.51231 | 345 |
| | 2 | 80.7136 | 182.62036 | 583 |
| | Total | 62.0108 | 147.21187 | 928 |

Table 9: Descriptive Statistics of SGPT and SGOT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| SGOT | 1 | 23.6435 | 11.06449 | 345 |
| | 2 | 1.0991E2 | 288.91853 | 583 |
| | Total | 77.2112 | 231.69093 | 928 |
| SGPT | 1 | 29.4058 | 18.51231 | 345 |
| | 2 | 79.7136 | 181.62036 | 583 |
| | Total | 61.0108 | 146.21187 | 928 |

Table 10: Descriptive Statistics of ALKPHOS, SGPT and SGOT

| | GROUP | Mean | Std. Deviation | N |
|---|---|---|---|---|
| SGOT | 1 | 24.6435 | 10.06449 | 345 |
| | 2 | 1.0991E2 | 288.91853 | 583 |
| | Total | 77.2112 | 231.69093 | 928 |
| SGPT | 1 | 31.4058 | 18.51231 | 345 |
| | 2 | 79.7136 | 181.62036 | 583 |
| | Total | 63.0108 | 146.21187 | 928 |
| ALKPHOS | 1 | 69.8696 | 18.34767 | 345 |
| | 2 | 2.9058E2 | 242.93799 | 583 |
| | Total | 2.0852E2 | 220.38146 | 928 |

Table 7, Table 8, Table 9 and Table 10 shows the descriptive statistics for the different combinations of attributes ALKPHOS,SGPT, ALKPHOS,SGOT , SGPT,SGOT and ALKPHOS,SGPT,SGOT respectively.

Multivariate Tests are applied for the combination of attributes ALKPHOS,SGPT, ALKPHOS,SGOT , SGPT,SGOT and ALKPHOS,SGPT,SGOT. The results of Multivariate tests are reported in tabular form as below:

Table 11: Multivariate Tests on ALKPHOS and SGPT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .469 | 4.088E2a | 2.0 | 925.0 | 0.0 | .469 | 817.698 | 1.0 |
| | Wilks' Lambda | .531 | 4.088E2a | 2.0 | 925.0 | 0.0 | .469 | 817.698 | 1.0 |
| | Hotelling's Trace | .884 | 4.088E2a | 2.0 | 925.0 | 0.0 | .469 | 817.698 | 1.0 |
| | Roy's Largest Root | .884 | 4.088E2a | 2.0 | 925.0 | 0.0 | .469 | 817.698 | 1.0 |
| Group | Pillai's Trace | .239 | 1.462E2a | 2.0 | 925.0 | 0.0 | .239 | 291.410 | 1.0 |
| | Wilks' Lambda | .759 | 1.462E2a | 2.0 | 925.0 | 0.0 | .239 | 291.410 | 1.0 |
| | Hotelling's Trace | .315 | 1.462E2a | 2.0 | 925.0 | 0.0 | .239 | 291.410 | 1.0 |
| | Roy's Largest Root | .315 | 1.462E2a | 2.0 | 925.0 | 0.0 | .239 | 291.410 | 1.0 |

Significant value for table 11 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, two populations differ a lot on ALKPHOS and SGPT.

Table 12: Multivariate Tests on ALKPHOS and SGOT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .455 | 3.863E2a | 2.0 | 925.0 | 0.0 | .455 | 772.615 | 1.0 |
| | Wilks' Lambda | .545 | 3.863E2a | 2.0 | 925.0 | 0.0 | .455 | 772.615 | 1.0 |
| | Hotelling's Trace | .835 | 3.863E2a | 2.0 | 925.0 | 0.0 | .455 | 772.615 | 1.0 |
| | Roy's Largest Root | .835 | 3.863E2a | 2.0 | 925.0 | 0.0 | .455 | 772.615 | 1.0 |
| Group | Pillai's Trace | .238 | 1.453E2a | 2.0 | 925.0 | 0.0 | .238 | 289.655 | 1.0 |
| | Wilks' Lambda | .760 | 1.453E2a | 2.0 | 925.0 | 0.0 | .238 | 289.655 | 1.0 |
| | Hotelling's Trace | .313 | 1.453E2a | 2.0 | 925.0 | 0.0 | .238 | 289.655 | 1.0 |
| | Roy's Largest Root | .313 | 1.453E2a | 2.0 | 925.0 | 0.0 | .238 | 289.655 | 1.0 |

Significant value for table 12 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, two populations differ a lot on ALKPHOS and SGOT.

Table 13: Multivariate Test on SGPT and SGOT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .121 | 63.431a | 2.0 | 925.0 | 0.0 | .121 | 126.861 | 1.0 |
| | Wilks' Lambda | .879 | 63.431a | 2.0 | 925.0 | 0.0 | .121 | 126.861 | 1.0 |
| | Hotelling's Trace | .137 | 63.431a | 2.0 | 925.0 | 0.0 | .121 | 126.861 | 1.0 |
| | Roy's Largest Root | .137 | 63.431a | 2.0 | 925.0 | 0.0 | .121 | 126.861 | 1.0 |
| Group | Pillai's Trace | .032 | 14.775a | 2.0 | 925.0 | 0.0 | .032 | 30.549 | 1.0 |
| | Wilks' Lambda | .966 | 14.775a | 2.0 | 925.0 | 0.0 | .032 | 30.549 | 1.0 |
| | Hotelling's Trace | .033 | 14.775a | 2.0 | 925.0 | 0.0 | .032 | 30.549 | 1.0 |
| | Roy's Largest Root | .033 | 14.775a | 2.0 | 925.0 | 0.0 | .032 | 30.549 | 1.0 |

Significant value for table 13 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, two populations differ a lot on SGOT and SGPT.

Table 14: Multivariate Test on ALKPHOS, SGPT and SGOT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .473 | 2.761E2a | 3.0 | 924.0 | 0.0 | .473 | 827.245 | 1.0 |
| | Wilks' Lambda | .527 | 2.761E2a | 3.0 | 924.0 | 0.0 | .473 | 827.245 | 1.0 |
| | Hotelling's Trace | .897 | 2.761E2a | 3.0 | 924.0 | 0.0 | .473 | 827.245 | 1.0 |
| | Roy's Largest Root | .897 | 2.761E2a | 3.0 | 924.0 | 0.0 | .473 | 827.245 | 1.0 |
| Group | Pillai's Trace | .241 | 96.462a | 3.0 | 924.0 | 0.0 | .240 | 291.386 | 1.0 |
| | Wilks' Lambda | .761 | 96.462a | 3.0 | 924.0 | 0.0 | .240 | 291.386 | 1.0 |
| | Hotelling's Trace | .317 | 96.462a | 3.0 | 924.0 | 0.0 | .240 | 291.386 | 1.0 |
| | Roy's Largest Root | .317 | 96.462a | 3.0 | 924.0 | 0.0 | .240 | 291.386 | 1.0 |

Significant value for table 14 is 0.0. That is less than 0.05 ($F < 0.05$). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, two populations differ a lot on ALKPHOS, SGPT and SGOT.

All significant values are less than the value defined at null hypothesis for four different multivariate tests for all the combination of attributes. This indicates that there is an important consequence of the independent variables on all of the dependent variables considered as a group.

### *B. ANOVA and MANOVA analysis of Liver Patient of UCI and India data set*

In this, there are records of only liver patients of the two populations. There are 145 records in UCI data set and 416 records in Indian data set. So, total numbers of records in this data set are 561. Firstly, Descriptive statistics of each individual attribute is done. Group 1 is used to denote UCI dataset and Group 2 is used to denote India data set.

Table 15: Descriptive Statistics of ALKPHOS

| ALKPHOS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 145 | 72.9793 | 17.59079 | 2.54388 | 67.9277 | 74.0309 | 23.00 | 138.00 |
| 2 | 416 | 3.1901E2 | 268.30791 | 13.15488 | 293.1487 | 344.8657 | 63.00 | 2110.00 |
| Total | 561 | 2.5516E2 | 255.25397 | 10.77683 | 233.9907 | 276.3266 | 23.00 | 2110.00 |

Table 16: Descriptive Statistics of SGPT

| SGPT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 145 | 30.2069 | 14.77793 | 2.31029 | 27.6170 | 32.7968 | 10.00 | 103.00 |
| 2 | 416 | 99.6058 | 212.76847 | 10.43183 | 79.1000 | 120.1116 | 12.00 | 2000.00 |
| Total | 561 | 80.9269 | 184.77111 | 8.84326 | 65.5211 | 96.3327 | 10.00 | 2000.00 |

Table 17: Descriptive Statistics of SGOT

| SGOT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 145 | 22.7862 | 7.73806 | .64261 | 21.5160 | 24.0564 | 5.00 | 57.00 |
| 2 | 416 | 1.3770E2 | 337.38998 | 16.54190 | 105.1832 | 170.2159 | 11.00 | 4929.00 |
| Total | 561 | 1.0800E2 | 294.80242 | 12.44657 | 83.5506 | 132.4459 | 5.00 | 4929.00 |

Table 15, Table 16 and Table 17 shows the descriptive statistics for the individual attributes ALKPHOS, SGPT and SGOT . Then, one-way ANOVA is applied for the attributes ALKPHOS, SGPT and SGOT. The results of one way ANOVA are shown as below:

Table 18: One Way ANOVA on ALKPHOS between UCI and INDIA datasets

| ALKPHOS | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 6561309.964273992 | 1 | 6561309.964273992 | 123.564 | .00 |
| Within Groups | 2.992525991629642E7 | 559 | 53533.560 | | |
| Total | 3.648656888057041E7 | 560 | | | |

Table 19: One Way ANOVA on SGPT between UCI and INDIA datasets

| SGPT | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 503032.8643077679 | 1 | 503032.8643077679 | 14.93886620254279 | .00 |
| Within Groups | 1.882E7 | 559 | 33672.761 | | |
| Total | 1.933E7 | 560 | | | |

Table 20: One Way ANOVA on SGOT between UCI and INDIA datasets

| SGOT | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 1419839.1858998295 | 1 | 1419839.1858998295 | 16.798064599907626 | .000 |
| Within Groups | 4.725E7 | 559 | 84523.975 | | |
| Total | 4.867E7 | 560 | | | |

Significant values of Table 18, Table 19 and Table 20 is 0.0. Null hypothesis is safely rejected. Hence, the two populations differ a lot for all the three attributes (ALKPHOS, SGPT and SGOT).

Now, descriptive statistics is calculated for the different combination of attributes ALKPHOS,SGPT, ALKPHOS,SGOT , SGPT,SGOT and ALKPHOS,SGPT,SGOT. Results are recorded as shown in tables below

Table 21: Descriptive Statistics of ALKPHOS and SGPT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| ALKPHOS | 1 | 70.9793 | 17.59079 | 145 |
| | 2 | 3.1901E2 | 268.30791 | 416 |
| | Total | 2.5516E2 | 255.25397 | 561 |
| SGPT | 1 | 31.2069 | 15.77793 | 145 |
| | 2 | 98.6058 | 211.76847 | 416 |
| | Total | 80.9269 | 184.77111 | 561 |

Table 22: Descriptive Statistics of ALKPHOS and SGOT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| ALKPHOS | 1 | 70.9793 | 17.59079 | 145 |
| | 2 | 3.1901E2 | 268.30791 | 416 |
| | Total | 2.5516E2 | 255.25397 | 561 |
| SGOT | 1 | 22.7862 | 7.73806 | 145 |
| | 2 | 1.3770E2 | 337.38998 | 416 |
| | Total | 1.0800E2 | 294.80242 | 561 |

Table 23: Descriptive Statistics of SGOT and SGPT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| SGOT | 1 | 21.7862 | 6.73806 | 145 |
| | 2 | 1.3770E2 | 337.38998 | 416 |
| | Total | 1.0800E2 | 294.80242 | 561 |
| SGPT | 1 | 31.2069 | 15.77793 | 145 |
| | 2 | 98.6058 | 211.76847 | 416 |
| | Total | 80.9269 | 184.77111 | 561 |

Table 24: Descriptive Statistics of ALKPHOS, SGPT and SGOT

| | GROUP | Mean | Std. Deviation | N |
|---|---|---|---|---|
| SGOT | 1 | 22.7862 | 7.73806 | 145 |
| | 2 | 1.3770E2 | 337.38998 | 416 |
| | Total | 1.0800E2 | 294.80242 | 561 |
| SGPT | 1 | 31.2069 | 15.77793 | 145 |
| | 2 | 98.6058 | 211.76847 | 416 |
| | Total | 80.9269 | 184.77111 | 561 |
| ALKPHOS | 1 | 71.9793 | 18.59079 | 145 |
| | 2 | 3.1901E2 | 268.30791 | 416 |
| | Total | 2.5516E2 | 255.25397 | 561 |

Table 21, Table 22, Table 23 and Table 24 shows the descriptive statistics for the combination attributes ALKPHOS,SGPT, ALKPHOS,SGOT, SGPT,SGOT and ALKPHOS,SGPT,SGOT respectively.

Multivariate Tests are applied for the combination of attributes ALKPHOS,SGPT, ALKPHOS,SGOT , SGPT,SGOT and ALKPHOS,SGPT,SGOT. The results of Multivariate tests are reported in tabular form as below:

Table 25: Multivariate Tests on ALKPHOS and SGPT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .378 | 1.698E2a | 2.0 | 558.0 | 0.0 | .378 | 339.623 | 1.0 |
| | Wilks' Lambda | .622 | 1.698E2a | 2.0 | 558.0 | 0.0 | .378 | 339.623 | 1.0 |
| | Hotelling's Trace | .609 | 1.698E2a | 2.0 | 558.0 | 0.0 | .378 | 339.623 | 1.0 |
| | Roy's Largest Root | .609 | 1.698E2a | 2.0 | 558.0 | 0.0 | .378 | 339.623 | 1.0 |
| Group | Pillai's Trace | .188 | 64.173a | 2.0 | 557.0 | 0.0 | .188 | 129.346 | 1.0 |
| | Wilks' Lambda | .810 | 64.173a | 2.0 | 557.0 | 0.0 | .188 | 129.346 | 1.0 |
| | Hotelling's Trace | .233 | 64.173a | 2.0 | 557.0 | 0.0 | .188 | 129.346 | 1.0 |
| | Roy's Largest Root | .233 | 65.173a | 2.0 | 557.0 | 0.0 | .188 | 129.346 | 1.0 |

Significant value for table 25 is 0.0. That is less than 0.05 ($F < 0.05$). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, two populations differ a lot on ALKPHOS and SGPT.

Table 26: Multivariate Tests on ALKPHOS and SGOT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .362 | 1.584E2a | 2.0 | 558.0 | 0.0 | .362 | 316.802 | 1.0 |
| | Wilks' Lambda | .638 | 1.584E2a | 2.0 | 558.0 | 0.0 | .362 | 316.802 | 1.0 |
| | Hotelling's Trace | .568 | 1.584E2a | 2.0 | 558.0 | 0.0 | .362 | 316.802 | 1.0 |
| | Roy's Largest Root | .568 | 1.584E2a | 2.0 | 558.0 | 0.0 | .362 | 316.802 | 1.0 |
| Group | Pillai's Trace | .186 | 63.337a | 2.0 | 558.0 | 0.0 | .186 | 127.673 | 1.0 |
| | Wilks' Lambda | .812 | 63.337a | 2.0 | 558.0 | 0.0 | .186 | 127.673 | 1.0 |
| | Hotelling's Trace | .230 | 63.337a | 2.0 | 558.0 | 0.0 | .186 | 127.673 | 1.0 |
| | Roy's Largest Root | .230 | 63.337a | 2.0 | 558.0 | 0.0 | .186 | 127.673 | 1.0 |

Significant value for table 26 is 0.0. That is less than 0.05 ($F < 0.05$). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, two populations differ a lot on ALKPHOS and SGOT.

Table 27: Multivariate Tests on SGOT and SGPT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .089 | 27.283a | 2.0 | 558.0 | 0.0 | .089 | 54.566 | 1.0 |
| | Wilks' | .911 | 27.283a | 2.0 | 558.0 | 0.0 | .089 | 54.566 | 1.0 |
| | Hotelling's Trace | .098 | 27.283a | 2.0 | 558.0 | 0.0 | .089 | 54.566 | 1.0 |
| | Roy's Largest Root | .098 | 27.283a | 2.0 | 558.0 | 0.0 | .089 | 54.566 | 1.0 |
| Group | Pillai's Trace | .030 | 8.921a | 2.0 | 557.0 | 0.0 | .030 | 16.841 | .973 |
| | Wilks' | .968 | 8.921a | 2.0 | 557.0 | 0.0 | .030 | 16.841 | .973 |
| | Hotelling's Trace | .031 | 8.921a | 2.0 | 557.0 | 0.0 | .030 | 16.841 | .973 |
| | Roy's Largest Root | .031 | 8.921a | 2.0 | 557.0 | 0.0 | .030 | 16.841 | .973 |

Significant value for table 27 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, two populations differ a lot on SGPT and SGOT.

Table 28: Multivariate Tests on ALKPHOS, SGPT and SGOT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .381 | 1.145E2a | 3.0 | 556.0 | 0.0 | .380 | 342.461 | 1.0 |
| | Wilks' | .619 | 1.145E2a | 3.0 | 556.0 | 0.0 | .380 | 342.461 | 1.0 |
| | Hotelling's Trace | .616 | 1.145E2a | 3.0 | 556.0 | 0.0 | .380 | 342.461 | 1.0 |
| | Roy's Largest Root | .616 | 1.145E2a | 3.0 | 556.0 | 0.0 | .380 | 342.461 | 1.0 |
| Group | Pillai's Trace | .189 | 42.446a | 3.0 | 556.0 | 0.0 | .191 | 131.339 | 1.0 |
| | Wilks' | .809 | 42.446a | 3.0 | 556.0 | 0.0 | .191 | 131.339 | 1.0 |
| | Hotelling's Trace | .233 | 42.446a | 3.0 | 556.0 | 0.0 | .191 | 131.339 | 1.0 |
| | Roy's Largest Root | .233 | 42.446a | 3.0 | 556.0 | 0.0 | .191 | 131.339 | 1.0 |

Significant value for table 28 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, two populations differ a lot on ALKPHOS, SGPT and SGOT.

*C. ANOVA and MANOVA analysis of Non Liver Patient of UCI and India data set*

In this, there are records of only non liver patients of two data sets. There are 200 records in UCI data set and 167 records in Indian data set. So, total numbers of records in this data set are 367. Firstly, Descriptive statistics of each individual attribute is done.

Table 29: Descriptive Statistics of ALKPHOS

| ALKPHOS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 200 | 68.3400 | 18.06199 | 1.27718 | 65.8215 | 70.8585 | 37.00 | 134.00 |
| 2 | 167 | 2.1975E2 | 140.98626 | 10.90984 | 198.2146 | 241.2944 | 90.00 | 1580.00 |
| Total | 367 | 1.3724E2 | 122.03879 | 6.37037 | 124.7127 | 149.7669 | 37.00 | 1580.00 |

Table 30: Descriptive Statistics of SGPT

| SGPT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 200 | 29.8250 | 21.84492 | 1.54467 | 26.7790 | 32.8710 | 4.00 | 155.00 |
| 2 | 167 | 33.6527 | 25.06039 | 1.93923 | 29.8240 | 37.4814 | 10.00 | 181.00 |
| Total | 367 | 30.5668 | 22.40824 | 1.22190 | 28.1639 | 32.9696 | 4.00 | 181.00 |

Table 31: Descriptive Statistics of SGOT

| SGOT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 200 | 25.9900 | 11.28880 | .79824 | 24.4159 | 27.5641 | 8.00 | 82.00 |
| 2 | 167 | 40.6886 | 36.41162 | 2.81762 | 35.1256 | 46.2516 | 10.00 | 285.00 |
| Total | 367 | 31.6785 | 25.91344 | 1.40487 | 28.9158 | 34.4411 | 8.00 | 285.00 |

Table 29, Table 30 and Table 31 shows the descriptive statistics for the individual attributes ALKPHOS, SGPT and SGOT respectively.

Then, ANOVA is applied for the attributes ALKPHOS, SGPT and SGOT is applied. The results of one way ANOVA are shown as below:

Table 32: One way ANOVA on ALKPHOS between UCI and INDIA datasets

| ALKPHOS | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 2086485.085050825 | 1 | 2086485.085050825 | 226.35210749432147 | .000 |
| Within Groups | 3364523.814131736 | 365 | 9217.873463374619 | | |
| Total | 5451998.899 | 366 | | | |

Table 33: One Way ANOVA on SGPT between UCI and INDIA datasets

| SGPT | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 1333.3831539917421 | 1 | 1333.3831539917421 | 2.4430163776633638 | 0.11891559975864789 |
| Within Groups | 199214.7312874252 | 365 | 545.7937843491102 | | |
| Total | 200548.11444141695 | 366 | | | |

Table 34: One Way ANOVA on SGOT between UCI and INDIA datasets

| SGOT | | | | | |
|---|---|---|---|---|---|
| | Sum of Squares | Df | Mean Square | F | Sig. |
| Between Groups | 19662.27156227057 | 1 | 19662.27156227057 | 29.239807483019625 | .000 |
| Within Groups | 245443.78838323356 | 365 | 672.4487352965303 | | |
| Total | 265106.0599455041 | 366 | | | |

Significant values of Table 32, Table 33 and Table 34 is 0.0. That means significant value is less than 0.05. So, null hypothesis is safely rejected. There is more significant difference between the two groups. Hence, the two populations differ a lot for all the three attributes (ALKPHOS, SGPT and SGOT).

Now, descriptive statistics is calculated for the different combination of attributes ALKPHOS,SGPT, ALKPHOS,SGOT , SGPT,SGOT and ALKPHOS,SGPT,SGOT. Results are recorded as shown in tables below:

Table 35: Descriptive Statistics of ALKPHOS and SGPT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| ALKPHOS | 1 | 67.3400 | 17.06199 | 200 |
| | 2 | 2.1975E2 | 140.98626 | 167 |
| | Total | 1.3724E2 | 122.03879 | 367 |
| SGPT | 1 | 29.8250 | 21.84492 | 200 |
| | 2 | 32.6527 | 24.06039 | 167 |
| | Total | 30.5668 | 22.40824 | 367 |

Table 36: Descriptive Statistics of ALKPHOS and SGOT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| ALKPHOS | 1 | 67.3400 | 17.06199 | 200 |
| | 2 | 2.1975E2 | 140.98626 | 167 |
| | Total | 1.3724E2 | 122.03879 | 367 |
| SGOT | 1 | 25.9900 | 11.28880 | 200 |
| | 2 | 39.6886 | 35.41162 | 167 |
| | Total | 31.6785 | 25.91344 | 367 |

Table 37: Descriptive Statistics of SGPT and SGOT

| | Group | Mean | Std. Deviation | N |
|---|---|---|---|---|
| SGOT | 1 | 24.9900 | 10.28880 | 200 |
| | 2 | 39.6886 | 35.41162 | 167 |
| | Total | 31.6785 | 25.91344 | 367 |
| SGPT | 1 | 28.8250 | 20.84492 | 200 |
| | 2 | 32.6527 | 24.06039 | 167 |
| | Total | 30.5668 | 22.40824 | 367 |

Table 38: Descriptive Statistics of ALKPHOS, SGPT and SGOT

| | GROUP | Mean | Std. Deviation | N |
|---|---|---|---|---|
| SGOT | 1 | 25.9900 | 11.28880 | 200 |
| | 2 | 40.6886 | 36.41162 | 167 |
| | Total | 31.6785 | 25.91344 | 367 |
| SGPT | 1 | 28.8250 | 20.84492 | 200 |
| | 2 | 32.6527 | 24.06039 | 167 |
| | Total | 30.5668 | 22.40824 | 367 |
| ALKPHOS | 1 | 68.3400 | 18.06199 | 200 |
| | 2 | 2.1975E2 | 140.98626 | 167 |
| | Total | 1.3724E2 | 122.03879 | 367 |

Table 35, Table 36, Table 37 and Table 38 shows the descriptive statistics for the combination attributes ALKPHOS,SGPT, ALKPHOS,SGOT, SGPT,SGOT and ALKPHOS,SGPT,SGOT respectively.

Multivariate Tests are applied for the different combination of attributes ALKPHOS,SGPT, ALKPHOS,SGOT , SGPT,SGOT and ALKPHOS,SGPT,SGOT. The results of Multivariate tests are reported in tabular form as below:

Table 39: Multivariate Tests on ALKPHOS and SGPT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .762 | 5.815E2a | 2.0 | 363.0 | 0.0 | .761 | 1162.006 | 1.0 |
| | Wilks' | .238 | 5.815E2a | 2.0 | 363.0 | 0.0 | .761 | 1162.006 | 1.0 |
| | Hotelling's Trace | 3.194 | 5.815E2a | 2.0 | 363.0 | 0.0 | .761 | 1162.006 | 1.0 |
| | Roy's Largest Root | 3.194 | 5.815E2a | 2.0 | 363.0 | 0.0 | .761 | 1162.006 | 1.0 |
| Group | Pillai's Trace | .392 | 1.167E2a | 2.0 | 363.0 | 0.0 | .392 | 232.442 | 1.0 |
| | Wilks' | .608 | 1.167E2a | 2.0 | 363.0 | 0.0 | .392 | 232.442 | 1.0 |
| | Hotelling's Trace | .642 | 1.167E2a | 2.0 | 363.0 | 0.0 | .392 | 232.442 | 1.0 |
| | Roy's Largest Root | .642 | 1.167E2a | 2.0 | 363.0 | 0.0 | .392 | 232.442 | 1.0 |

Significant value for table 39 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Two populations differ a lot on ALKPHOS and SGPT.

Table 40: Multivariate Tests on ALKPHOS and SGOT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .757 | 5.661E2a | 2.0 | 363.0 | 0.0 | .756 | 1131.186 | 1.0 |
| | Wilks' | .243 | 5.661E2a | 2.0 | 363.0 | 0.0 | .756 | 1131.186 | 1.0 |
| | Hotelling's Trace | 3.121 | 5.661E2a | 2.0 | 363.0 | 0.0 | .756 | 1131.186 | 1.0 |
| | Roy's Largest Root | 3.121 | 5.661E2a | 2.0 | 363.0 | 0.0 | .756 | 1131.186 | 1.0 |
| Group | Pillai's Trace | .384 | 1.141E2a | 2.0 | 363.0 | 0.0 | .384 | 227.271 | 1.0 |
| | Wilks' | .614 | 1.141E2a | 2.0 | 363.0 | 0.0 | .384 | 227.271 | 1.0 |
| | Hotelling's Trace | .626 | 1.141E2a | 2.0 | 363.0 | 0.0 | .384 | 227.271 | 1.0 |
| | Roy's Largest Root | .626 | 1.141E2a | 2.0 | 363.0 | 0.0 | .384 | 227.271 | 1.0 |

Significant value for table 40 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, populations differ a lot on ALKPHOS and SGOT.

Table 41: Multivariate Tests on SGPT and SGOT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .682 | 3.911E2a | 2.0 | 363.0 | 0.0 | .681 | 781.195 | 1.0 |
| | Wilks' | .318 | 3.911E2a | 2.0 | 363.0 | 0.0 | .681 | 781.195 | 1.0 |
| | Hotelling's Trace | 3.149 | 3.911E2a | 2.0 | 363.0 | 0.0 | .681 | 781.195 | 1.0 |
| | Roy's Largest Root | 3.149 | 3.911E2a | 2.0 | 363.0 | 0.0 | .681 | 781.195 | 1.0 |
| Group | Pillai's Trace | .088 | 16.344a | 2.0 | 363.0 | 0.0 | .086 | 33.689 | 1.0 |
| | Wilks' | .915 | 16.344a | 2.0 | 363.0 | 0.0 | .086 | 33.689 | 1.0 |
| | Hotelling's Trace | .097 | 16.344a | 2.0 | 363.0 | 0.0 | .086 | 33.689 | 1.0 |
| | Roy's Largest Root | .097 | 16.344a | 2.0 | 363.0 | 0.0 | .086 | 33.689 | 1.0 |

Significant value for table 41 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. There is more significant difference between the two populations. Hence, populations differ a lot on SGPT and SGOT.

Table 42: Multivariate Tests on ALKPHOS, SGPT and SGOT between UCI and INDIA datasets

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .772 | 4.092E2a | 3.0 | 362.0 | 0.0 | .771 | 1226.699 | 1.0 |
| | Wilks' | .228 | 4.092E2a | 3.0 | 362.0 | 0.0 | .771 | 1226.699 | 1.0 |
| | Hotelling's Trace | 3.382 | 4.092E2a | 3.0 | 362.0 | 0.0 | .771 | 1226.699 | 1.0 |
| | Roy's Largest Root | 3.382 | 4.092E2a | 3.0 | 362.0 | 0.0 | .771 | 1226.699 | 1.0 |
| Group | Pillai's Trace | .407 | 82.103a | 3.0 | 362.0 | 0.0 | .406 | 248.308 | 1.0 |
| | Wilks' | .593 | 82.103a | 3.0 | 362.0 | 0.0 | .406 | 248.308 | 1.0 |
| | Hotelling's Trace | .687 | 82.103a | 3.0 | 362.0 | 0.0 | .406 | 248.308 | 1.0 |
| | Roy's Largest Root | .687 | 82.103a | 3.0 | 362.0 | 0.0 | .406 | 248.308 | 1.0 |

Significant value for table 42 is 0.0. That is less than 0.05 (F < 0.05). It means null hypothesis is safely rejected. It means there is more significant difference between the two populations. Hence, populations differ a lot on ALKPHOS, SGPT and SGOT.

### D. APRIORI ALGORITHM

Apriori is a standard algorithm for repeated item set mining. In this, different association rules are learned for different transactional databases. It proceeds by identifying the common individual items in the database and extending these item sets to larger and larger item sets as long as those item sets appear adequately often in the database. The frequent item sets are determined by Apriori can be used to determine association rules which highlight general trends in the database.

Firstly data preprocessing is done by following the different steps as:

1. Load arff file of data set containing records of patients.
2. Applying filter to remove unnecessary attributes from the dataset like Group,
3. Then apply discretize filter to make nominal data type of all the other attributes (ALKPHOS, SGPT and SGOT) of data set.
4. During discretization of attributes five number of bins with equal frequency.
5. After preprocessing of original data set new data set is saved with new name.
6. Normalized form of data set is shown as below

```
@relation CombinedData-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Discretize-F-B5-M-1.0-R1-
weka.filters.unsupervised.attribute.Discretize-F-B5-M-1.0-R2-
weka.filters.unsupervised.attribute.Discretize-F-B5-M-1.0-R3


@attribute alkphos {'\'[0-68.5]\'','\'(68.5-136]\'','\'(136-
190.5]\'','\'(190.5-273.5]\'','\'(273.5-max]\''}

@attribute sgpt {'\'[0-19.5]\'','\'(19.5-25.5]\'','\'(25.5-34.5]
\'','\'(34.5-56.5]\'','\'(56.5-max]\''}

@attribute sgot {'\'[0-19.5]\'','\'(19.5-24.5]\'','\'(24.5-33.5]
\'','\'(33.5-65.5]\'','\'(65.5-max]\''}

@attribute selector {yes,no}


@data

'\'(68.5-136]\'','\'(34.5-56.5]\'','\'(24.5-33.5]\'',yes
'\'[0-68.5]\'','\'(56.5-max]\'','\'(24.5-33.5]\'',no
'\'[0-68.5]\'','\'(25.5-34.5]\'','\'[0-19.5]\'',no
'\'(68.5-136]\'','\'(25.5-34.5]\'','\'(19.5-24.5]\'',no
'\'(68.5-136]\'','\'[0-19.5]\'','\'(24.5-33.5]\'',no
'\'[0-68.5]\'','\'[0-19.5]\'','\'[0-19.5]\'',no
'\'[0-68.5]\'','\'(19.5-25.5]\'','\'[0-19.5]\'',yes
'\'[0-68.5]\'','\'(19.5-25.5]\'','\'[0-19.5]\'',yes
'\'[0-68.5]\'','\'(19.5-25.5]\'','\'(19.5-24.5]\'',yes
'\'[0-68.5]\'' '\'(19.5-25.5]\'' '\'[0-19.5]\'' yes
```

Figure 1: Normalized form of data set

7. Then before applying Apriori algorithm different attributes of Apriori algorithm are set as lowerBoundMinSupport 0.01 – 0.10 numRules as 100 and upperBoundMinSupport as 1.0.
8. Best Rules found from Apriori algorithm for all five data sets

### E. NAÏVE BAYESIAN ALGORITHM

The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all

attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems.

Preprocessing of data set before applying Naïve Bayesian algorithm steps followed as below:

1. Load arff file of data set containing records of patients.
2. Applying filter to remove unnecessary attributes from the dataset like Group.
3. Then apply NumericToNominal filter data type to change data type of all the other attributes (ALKPHOS, SGPT and SGOT) of data set.
4. Then save normalized dataset as shown below

```
@relation CombinedData-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
@attribute Group {1,2}
@attribute alkphos {23,35,36,37,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,
55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,
82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102,103,105,106,107,
108,109,110,114,115,116,117,119,120,122,123,125,127,128,130,134,135,137,138,140,142,
143,144,145,146,147,148,149,150,151,152,153,154,155,156,157,158,159,160,161,162,163,164,
165,166,167,168,169,170,171,172,173,174,175,176,177,178,179,180,181,182,183,184,185,186,
187,188,189,190,191,192,193,194,195,196,197,198,199,200,201,202,204,205,206,208,209,210,
211,212,214,215,216,218,219,220,224,225,226,227,228,230,231,232,234,235,236,237,238,239,
240,243,245,246,247,248,250,251,253,254,256,257,258,259,260,262,263,265,268,269,270,271,
272,275,276,279,280,282,285,286,289,290,292,293,298,300,302,305,308,309,310,312,314,315,
316,320,326,331,332,335,340,342,348,349,350,352,356,358,360,365,367,374,375,380,386,388,390,
392,395,400,401,405,406,410,415,418,430,450,458,460,462,466,470,480,482,486,490,498,500,505,509,
512,515,518,527,538,542,554,555,558,560,562,574,575,580,588,592,599,610,612,614,621,630,650,661,664,
670,680,686,690,699,719,750,768,802,805,850,859,862,901,915,950,962,1020,1050,1100,1110,1124,1350,1420,1550,
@attribute sgpt {4,5,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,
37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,67,68,69,70,71,72,
74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,93,94,95,96,97,99,102,103,107,110,112,113,114,115,
116,118,119,120,123,126,131,132,133,137,139,140,141,142,148,149,152,154,155,157,159,160,166,168,173,178,179,
181,189,190,194,196,198,205,213,220,230,232,233,284,308,321,322,349,378,382,390,404,407,412,425,440,482,509,
622,779,790,875,950,1250,1350,1630,1680,2000}
@attribute sgot {5,8,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,
33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,
64,65,66,67,68,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,95,97,98,
99,100,101,102,103,104,105,108,110,111,113,114,116,125,126,127,130,134,135,138,139,140,141,142,
143,145,148,149,150,152,155,156,161,168,176,178,180,181,185,186,187,188,190,200,202,220,221,230,
231,232,233,235,236,245,247,248,250,275,285,298,330,348,350,367,368,384,397,400,401,405,406,441,
497,500,511,540,562,576,602,623,630,731,794,844,850,950,960,1050,1500,1600,2946,4929}
@attribute selector {yes,no}
@data1,92,45,27,yes
1,64,59,32,no
1,54,33,16,no
1,78,34,24,no
1,70,12,28,no
1,55,13,17,no
1,62,20,17,yes
1,67,21,11,yes
1,54,22,20,yes
1,60,25,19,yes
1,52,13,24,yes
1,62,17,17,yes
1,64,61,32,yes
1,77,25,19,yes
1,67,29,20,yes
1,78,20,31,yes
1,67,23,16,yes
1,79,17,17,yes
1,107,20,20,yes
1,116,11,33,yes
1,59,35,13,yes
```

Figure 2: Normalized form of dataset for Naïve Bayesian

5. Then Apply Naïve Bayesian Algorithm for given training set, Cross Validation and percentage split for three datasets (Combined, India and US Dataset).
6. As a result we get confusion matrices for all datasets.

### F. Bar Graph for Data Sets

For the Combined dataset in which all Patients that means both liver and non liver patients of UCI and India data set. UCI data set contains 345 patient records and India data set contains 583 patient records. Total records are 928.

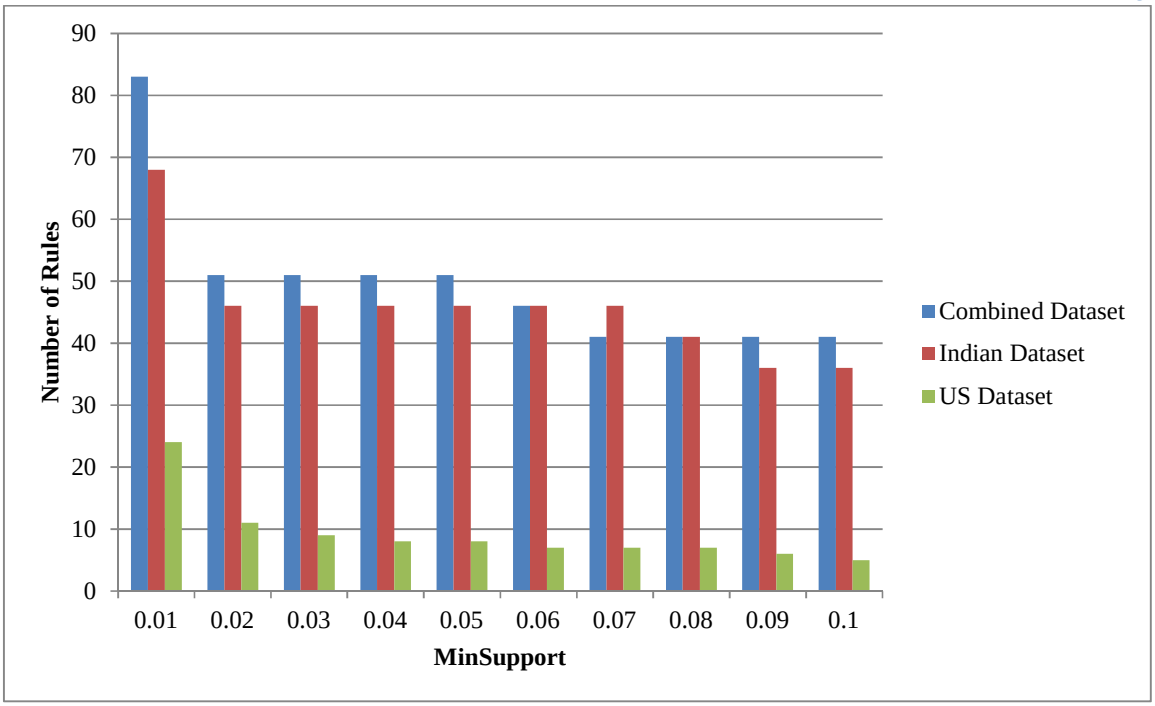


Figure 3: Bar Graph Between number of rules vs MinSupport for three datasets

Figure 3 shows the bar graph the number of rules found for Apriori algorithm. In this Graph shows that for Apriori number of rules are decreasing as the value of MinSupport is increasing. At value of MinSupport 0.01 - 0.10 number of rules for Combined dataset and Indian dataset are much more as compared to that of US dataset. At the values of MinSupport 0.06 and 0.08 number of rules for combined data set and Indian dataset are equal. But for US dataset Number of rules is very less as compared to other two dataset. As accuracy of Aprioi is above 80% but MinSupport is very less and on the other hand according to the rules of US dataset is very less.

Table 43. Classification Accuracy for Two classes for three datasets using Naïve Bayesian

| **Classification Accuracy** | | | |
|---|---|---|---|
| | Naïve Bayesian | | |
| | Combined dataset | Indian Dataset | US Dataset |
| **Yes** | 60.45% | 71.36% | 42.03% |
| **No** | 39.55% | 28.64% | 57.97% |

Above table shows classification accuracy of Naïve Bayesian algorithm for two classes. It shows that 71.36% data of Indian dataset is highly accurate and US dataset is least accurate. Naïve Bayesian shows actual accuracy of classification as ANOVA, MNOVA and Apriori fails to show actual results of classification. Also execution time of Naïve Bayesian is very less as compared to all other methods (ANOVA, MANOVA) that are applied to different datasets that is in milli seconds.

## V. CONCLUSION

In this paper, medical data of liver patients have analyzed. There is large amount of data in any hospital. As day by day medical transactions are becoming large and complex. So, it's very difficult to access data of particular patient. To make quick or easy decisions, there is requirement of Medical Decision Support System (MDS). It is concluded is that when ANOVA and MANOVA are applied for first two data sets there is more significant difference between two populations. In third data set, analysis on SGPT between non liver patients of USA and India data sets, there is no significant difference between the two populations. So, further experiments of Apriori and Naïve Bayesian Algorithms are conducted. These algorithms have generated different association rules of each. Number of association rules of both Apriori and Naïve Bayesian are almost same for all five categories of data sets. But in case of combined data set numbers of common rules are very less. From this it is shown that there is not much more difference between the association rules of the two algorithms. But execution time of Naïve Bayesian less as compared to all other methods (ANOVA, MANOVA) that are applied to different data sets that is in milli seconds.

## VI. FUTURE SCOPE

- In our research, we have diagnosed Liver data sets in the same way by using different algorithms for other diseases like breast cancer, kidney disorder etc.
- On the basis of best rules found, we can develop an automated medical diagnosis system and need for its localization settings based on the geographical region.
- There are many other algorithms (K – Optimal Pattern Discovery, Sequential Pattern mining, Quantitative Association Mining, High – Order pattern Discovery etc.) that can be used for data mining.
- Number of attributes can change for different data sets.
- There are different exploratory methods (Neural Networks, Multivariate Exploratory Techniques etc.).

## REFERENCES

[1] Anil Kumar Tiwari,Lokesh Kumar Sharma, G. Rama Krishna, "Comparative Study of Artificial Neural Network based Classification for Liver Patient", Journal of Information Engineering and Applications ISSN 2224-5782 (print) ISSN 2225-0506 (online), Vol.3, No.4, 2013

[2] B. Surendiran, Y. Sundaraiah, A. Vadivel "Classifying Digital Mammogram Masses using Univariate ANOVA Discriminant Analysis", In Proceedings of the IEEE International Conference on Advances in Recent Technologies in Communication and Computing, pages 175-177,2009.

[3] Bendi Venkata Ramana, Prof. M. Surendra Prasad Babu, Prof. N. B. Venkateswarlu "A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012.

[4] Benjamin F. Merembeck and Brian J. Tuner: "Directed Canonical Analysis and Performance of Classifiers Under Its Associated Linear Transformation" ,In Proceedings of the IEEE Transactions on Geooscience and Remote Sensing, Vol Ge-18, No. 2, pages 190 -196, April 1980.

[5] C.K. Loo, "Accurate and Reliable Diagnosis and Classification Using Probabilistic Ensemble Simplified Fuzzy ARTMAP", IEEE Transactions on Knowledge and Data Engineering, Vol 17, No.11, 2005.

[6] Junning Li, Z. Jane Wang and Martin J. McKeown, "A Framework for Group Analysis of FMRI Data using Dynamic Bayesian Networks" , In Proceedings of the 29th IEEE International Conference on EMBS pages 5991-5994, August 2007.

[7] Hian Chye Koh and Gerald Tan, "Data Mining Applicationsin Healthcare", Journal of Healthcare Information Management — Vol. 19, No. 2

[8] Hongjun Lu, Rudy Setiono, and Huan Liu,"Effective Data Mining Using Neural Networks", IEEE Transactions On Knowledge And Data Engineering, Vol. 8, No. 6, December 1996

[9] " http://consumersmedical.com/Medical-Decision-Support.html "

[10] "http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Min ing/The_Apriori_Algorithm.

[11] "http://en.wikipedia.org/wiki/Association_rule_learning#FP-growth_algorithm"

[12] "https://archive.ics.uci.edu/ml/datasets.html"

[13] Hyontai Sug," Effective Data Mining of Integrated Data Sets Using Decision Trees", International Journal of Mathematics and Computers In Simulation, Issue 3, Volume 7, 2013

[14] Jaiwei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", second edition, San Francisco, USA, ISBN 1-55860-901-6, 2006.

[15] Junning Li, Z. Jane Wang and Martin J. MCKeown," A Framework for Group Analysis of FMRI Data using Dynamic Bayesian Networks" , In Proceedings of 29th IEEE international Conference on EMBS pages 5991-5994, August 2007.

[16] Liver Disorder Data Set Available: "http://archive.ics.uci.edu/ml/datasets".

[17] M. S. Chen, J. Han, and P. S. Yu.,"Data mining: An overview from a database perspective", IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[18] Mark L. Berenson Levin, "Basic Business Statistics: Concepts and Applications", 7th edition,USA, ISBN:0137956185, 1998.

[19] Martha L. Zequera, Leonardo Garavito, William Sandham, Jorge A. Alvarado, Angela odriguez, Carlos A. Wilches, Ana c. Villa, Shirley V. Quintero and Juan C. Bernal, "Assessment of the effect of time in the repeatability of the stabilometric parameters in diabetic and non-diabetic subjects during bipedal standing using the LorAn pressure distribution measurement system", In Proceedings of the 33rd IEEE International Conference on EMBS, pages 8531-8534, September 2011.

[20] Mehdi Neshat and Abas E.Zadeh,"Hopfield Neural Network and Fuzzy Hopfield Neural Network for Diagnosis of Liver Disorders", 978-1-4244-5164-7/10 ©2010 IEEE

[21] Mehdi neshat, Dr.Mehdi yaghobi and Dr.Mohammad naghibi, "Designingan Expert System Of Liver Disorders By Using Neural Networkand Comparing It With Parametric And Nonparametric System", 2008 5th International Multi-Conference on Systems, Signals and Devices

[22] Mireille Tohm'e, R'rgis Lengell'e and Virginie Freytag" A multi class multivariate Mireille group comparison test, Application to drug safety", In Proceedings of the 32nd IEEE International Conference on EMBS, 2006, pages 4711-4714, Sepetember4, 2011.

[23] Neven Cukrov, Natasa Tepi, Dario Omanovi, Sonja Lojen, Elvira Bura-Naki, Vjerocka Vojvodi and Ivanka Pizeta "Anthropogenic and Natural Influences on the Krka River (Croatia) Evaluated by Multivariate Staristical Analysis" ,In Proceedings of the 31st IEEE International Conference on Information Technology Interfaces pages 219-224, June 2009.

[24] Parisa Tavakkoli_, Davood M. Sourany, Saeed Tavakkoliz, Majid Hatamianx, Armin Mehrabian, Valentina E. Balas," Classification of the Liver Disorders Data Using Multi-Layer Adaptive Neuro-Fuzzy Inference System", IEEE – 35239 6th ICCCNT – 2015 July 13 - 15, 2015,Denton, U.S.A

[25] Paulo Ricardo Galhanone, David Martin Simpson, Antonio Fernando C. Infantosi Eduardo Faveret, Maria Alice Genofre, Helio Bello and Leonard de Azevedo "Multivariate Analysis of Neonatal EEG in different Sleep Stages: Methods and Preliminary Results" , In

Proceedings of the 17th IEEE International Conference on Engineering in Medicine and Biology Society, vol 2, pages 1021-1022,1995.

***1)*** *[26] Ricardo Ribeiro, Rui Tato Marinho, Jos´e Velosa, Fernando Ramalho, J. Miguel Sanches and Jasjit S. Suri, "The Usefulness of Ultrasound in the Classification of Chronic Liver Disease", Engineering in Medicine and Biology Socity, EMBC, 2011 Annual International Conference of the IEEE, page no.5132-5135, 2011.*

[27] S. Dimitrova " Investigations of Some Human Physiological Parameters in Relation of Geomagnetic Variations of Solar Origin and Meteorological Factors" , In Proceedings of the 2nd IEEE International Conference on Recent Advances in Space Technologies, pages 728-733,2005.

[28] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus, "Knowledge Discoveryin Databases:An Overview", AI Magazine Volume 13 Number 3 (1992).

[29] Z. A. Dastgheib, B. Lithgowand Z. Moussavi " Application of Fractal Dimension on Vestibular Response Signals for diagnosis of Parkinson's Disease" , In Proceedings of the 33rd IEEE International Conference on EMBS, pages 7892-7895, September 2011.