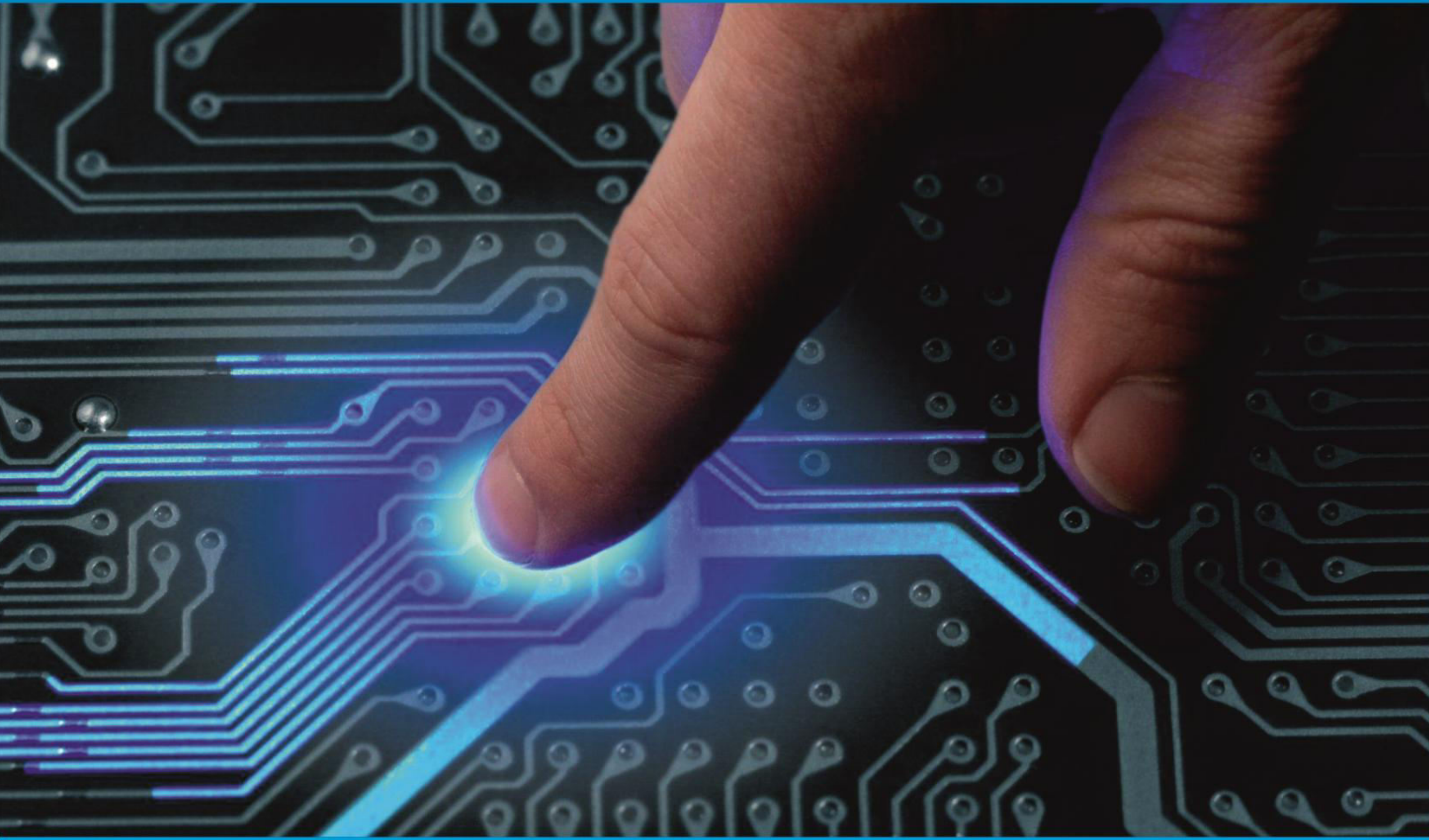
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 5, May 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

| DOI: 10.15680/IJIRCCE.2021.0905061 |

A Case Study on Liver Patient Data in India Using Statistics and Data Studio

Vikrant Thakur[1], **Siddharth Nanda**[2]

U.G Student, School of Engineering, Ajeenkya DY Patil University, Pune, Maharashtra, India [1]

Faculty, School of Engineering, Ajeenkya DY Patil University, Pune, Maharashtra, India [2]

ABSTRACT: This research paper is a case study on liver patient data from India and its analysis using statistical method by proper sampling and also by using Google data studio. The dataset we are working on is available on Kaggle website. You can find the same at the link given below the image after this paragraph. The dataset consists of many parameters which can be useful to determine liver disease in a human being. There was total 589 population or values out of which the ideal sample size was estimated to be 232. We are using systematic sampling method for selecting these 232 samples. Then we will divide the sample into two strata on the basis on disease confirmation. Then we calculate and compare them. The dataset used for the analysis is available on Kaggle. The dataset consists of many parameters like protein level, AG-ratio, Age and Albumin count. These are all important factors that may affect liver in some way. We will be studying which of the above chemicals are more harmful for liver disease.

KEYWORDS: Data studio, systematic sampling, sampling, liver**,** stratified sampling

I. INTRODUCTION

The liver is an important human organ. It sits just under your rib cage on the right side of your abdomen. The liver is important for the functioning of digestion and filtering toxic substance from body. Liver disease can be genetic which makes it more problematic. Some of the Symptoms of liver disease include dark in urine, Nausea, Itchy skin, yellowish appearance of skin and eyes and many such symptoms. More than 2 million people lose their lives every year due to liver related diseases. According to the latest WHO data published in 2017, liver disease deaths in India reached 259,749 or 2.95% of total deaths, accounting for one-fifth (18.3%) of all cirrhosis deaths globally.In this paper, we have using some concepts like systematic sampling for performing a statistical analysis on India liver patients. The parameters for the liver will be deeply studied and conclusion will be made about their effect on liver. We will create meaningful data visualization from the given dataset using google data studio.

II. LITERATURE SURVEY

1. Developing Sampling Frame for Case Study: Challenges and Conditions. By: Noriah Mohd Ishak [1] & Abu Yazid Abu Baka [2], Accessed on 02/04/2021

Description: Statistical analysis is very important in order to understand the data. But before any kind of analysis, it is very important that proper sample should be selected which can represent the whole population. If it is not suitable for analysis then results can get affected to some extent. So, Statistician where very much concerned about the need of proper sampling technique which led to creation of many sampling methods. These sampling methods are of two types Random sampling and Non-Random sampling. In this paper the author has discussed about the need of sampling and how we can develop one sampling framework for any case study.

2. Analysis of Liver Disorder Using Data Mining Algorithm by P. Rajeshwari [1] and G Sophia Reena [2]accessed on 02/4/2021.

Description: This paper discusses about important data mining algorithm which can be very useful for analysis of liver disorder and help for understanding its prevention and treatment methods. The study of liver development helps us to understand more details about the morphogenesis and differentiation of other human organs. Knowledge about the understanding and prevention of human congenital diseases. Significantly, much of understanding of organ development has arisen from analyses of patients with liver deficiencies. In this paper the data classification is based on liver disorder the training data set is developed by collecting data from UCI repository consists of 345 instances with 7 different attributes.

3. Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey. Shambel Kefelegn [1] and Pooja Kamat [2]

Description: This paper discusses about important data mining techniques which can be very useful for analysis of liver disorder and help for understanding its prevention and treatment methods. The study of liver development helps us to understand more details about the morphogenesis and differentiation of other human organs. Knowledge about the understanding and prevention of human congenital diseases. Liver disorder diseases one of the major diseases in the world, Liver is one of the huge important organs in the human body which is also considered as a gland because along with filtering of many substance in body it also secrets bile juice. The liver plays a vital role in many physical functions from protein building and blood clotting to fat, sugar and iron metabolism.

4. Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms by M. Banu Priya1, P. Laura Juliet2, P.R. Tamilselvi3

Description: This paper discusses about important machine learning techniques which can be very useful for analysis of liver disorder and help for understanding its prevention and treatment methods. The study of liver development helps us to understand more details about the morphogenesis and differentiation of other human organs. Knowledge about the understanding and prevention of human congenital diseases. Liver disorder diseases one of the major diseases in the world, Liver is one of the huge important organs in the human body which is also considered as a gland because along with filtering of many substance in body it also secrets bile juice. Itplays a vital role in many physical functions from protein building and blood clotting to fat, sugar and iron metabolism.

5. Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques by JagdeepSingh[a] ,SachinBagga[b] and RanjodhKaur[c]

Description: Today's health care is very important for everyone, so there is a need to provide affordable medical services to everyone. In this paper, the main focus is on predicting liver disease according to software engineering method using classification techniques and methods for selecting traits. The implementation of the proposed work was carried out on the Indian Liver Patient Dataset (ILPD) from the University of California, Irvine database. Various attributes such as age, direct bilirubin, sex, total bilirubin, Alkphos, sgpt, albumin, globulin and calf dosage, etc., of the liver patient database, are used to predict the degree of risk of liver disease. Various classification algorithms such as Logistic Regression, SMO, Random Forest algorithm, Naive Bayes, J48, and neighboring k (IBk) are used in the Liver Patient Database to obtain accuracy. The comparison of the dividing results of the classification is done by feature selection and without using the feature selection process. The development of intelligent liver disease software (ILDPS) is done using a selection of predictable features and strategies according to the software engineering model.

6. Analysis of classification algorithms for liver disease diagnosis by S.R GOSH1 and Sajjad waheed2

Description: These days liver disease is on the rise due to excessive drinking, smoking, drinking water contaminated with arsenic, obesity, low immune system, and heredity. Symptoms of liver cancer can include jaundice, abdominal pain, fatigue, nausea, vomiting, back pain, abdominal swelling, weight loss, generalized itching. Selected algorithms can be used in medical tools (e.g., CT scanner, MRI, ultra-sonic, ECG etc.) to reduce time and cost in hepatic disease screening. Here are other algorithms such as Naive Bayes classification (NBC), Bagging, KStar, Logistic, and REP tree used to test accuracy, accuracy, sensitivity, and specificity. These two sets of UCLA and AP data sets are considered to be the best algorithm. All analyses are performed using the software 3.3.10. Revealed, the KStar algorithm had high accuracy, precision, sensitivity, and specificity. On the other hand, less accuracy was found on NBC. Therefore, the K * algorithm can be used in diagnostic tools or in tools to quickly diagnose a particular liver problem.

7. LIVER DISEASE PREDICTION BY USING DIFFERENT DECISION TREE TECHNIQUES by Nazmun Nahar1 and Ferdous Ara2

Description: Early diagnosis of liver disease is very important to save a person's life and take appropriate measures to controldisease. Decision Tree algorithms have been used successfully in a variety of fields, especially in medicinescience. This research work examines the early prognosis of liver disease using a variety of decision therapiesstrategies. The hepatic data selected for this study contains attributes such as total valuebilirubin, direct bilirubin, age, gender, total protein, albumin and globulin ratio. The main purpose of thistask to calculate the effectiveness of the various methods of decision trees and compare their effectiveness.The decision tree strategies used in this study are J48, LMT, Random Forest, Random tree, REPTree,Decision Title, and Hoeffding tree. Analysis proves that Decision Stump offers the highest qualityaccuracy has other options.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905061 |

8. Performance Evolution of Different Machine Learning Algorithms for Prediction of Liver Disease by Muktevi Srivenkatesh1

Description: Liver disease is a general medical problem associated with distinct disorders and high mortality. It's for the basic importance is that the illness is detected before it is so serious the numbers of these lives can be saved. Stages of liver disease it is an important concept for focused therapy. It's horribledifficult performance of medical analysts to fore see disease within the early stages due to sensitivity manifestation. Often the side effects are obvious oncethe point is that there is no going back. To strike this down, we have the courage infection forecast. Liver disease can be classified ascountless programs, and these are classified as prediction of the use of numerical highlighting and division combinations. In this study, we used five types of distinctions namely Naïve Bayes, retreat, vector support systems, Unplanned Forest, K The nearest neighbour for testing liver disease. Separation shows tested at 5different from the magnitude of the murder, that is, to clarify, kappa, Maximum error (MAE), Root means square error(RMSE), and F measures. The purpose of this question function is tosee liver infection by reading a different machine and choosing a very efficient algorithm.

9. Stratified-sampling over social networks using MapReduce by Roy levin1 and Yaron kanza2

Description: Sampling is used in mathematical research to select a set of specific people to specific individuals, to measure human properties. In the divided sample, the tested subjects were divided into smaller groups and individuals were selected within smaller groups, reducing the sample size. In this paper we look at a large sample, which is distributed online, and shows how we can deal with cases where several similar studies are conducted-- in some studies it may be necessary to share them individually to reduce costs, while other studies, sharing should be reduced, e.g., to prevent test fatigue. A separate sample of multidisciplinary research is the task of selecting people from multidisciplinary research, similarly, according to sharing issues, without bias. In this paper, we introduce a seamless distributed algorithm, designed for the MapReduce framework, to answer sample queries used by the number of people in a communication network. We have also introduced a separate sample response algorithm for most studies, and show how we can use it using MapReduce. Experimental tests demonstrate the effectiveness of our algorithms and their performance for a separate sample for most studies.

10. Recent Developments in Systematic Sampling: A Review by Sayed mustafa1

Description: Systematic sampling is one of the most common methods of measurement. The popularity of a structured design is largely due to its functionality. Compared to a simple random sample, it is easier to draw a particularly structured sample when the selection of sample units is done in a field. In addition, systematic sampling may provide more accurate estimates than simple random sampling where explicit or implicit specifications are obtained in sample frames. However, formal structure has two major drawbacks. First, if the size of the population is not a component of the sample size you want, the actual sample size will be random. Second, a single systematic sample cannot provide an unbiased measurement of sample variability. Another limitation in formal construction is that for people with a specific structure, the efficiency of systematic sample speculators will depend largely on the relationship between the length of time and the measurement time. In the book, one can find that many attempts have been made to address one or more of these problems. This current paper provides an overview of the latest work in this area and provides recommendations for study staff using a structured design for a variety of sample conditions.

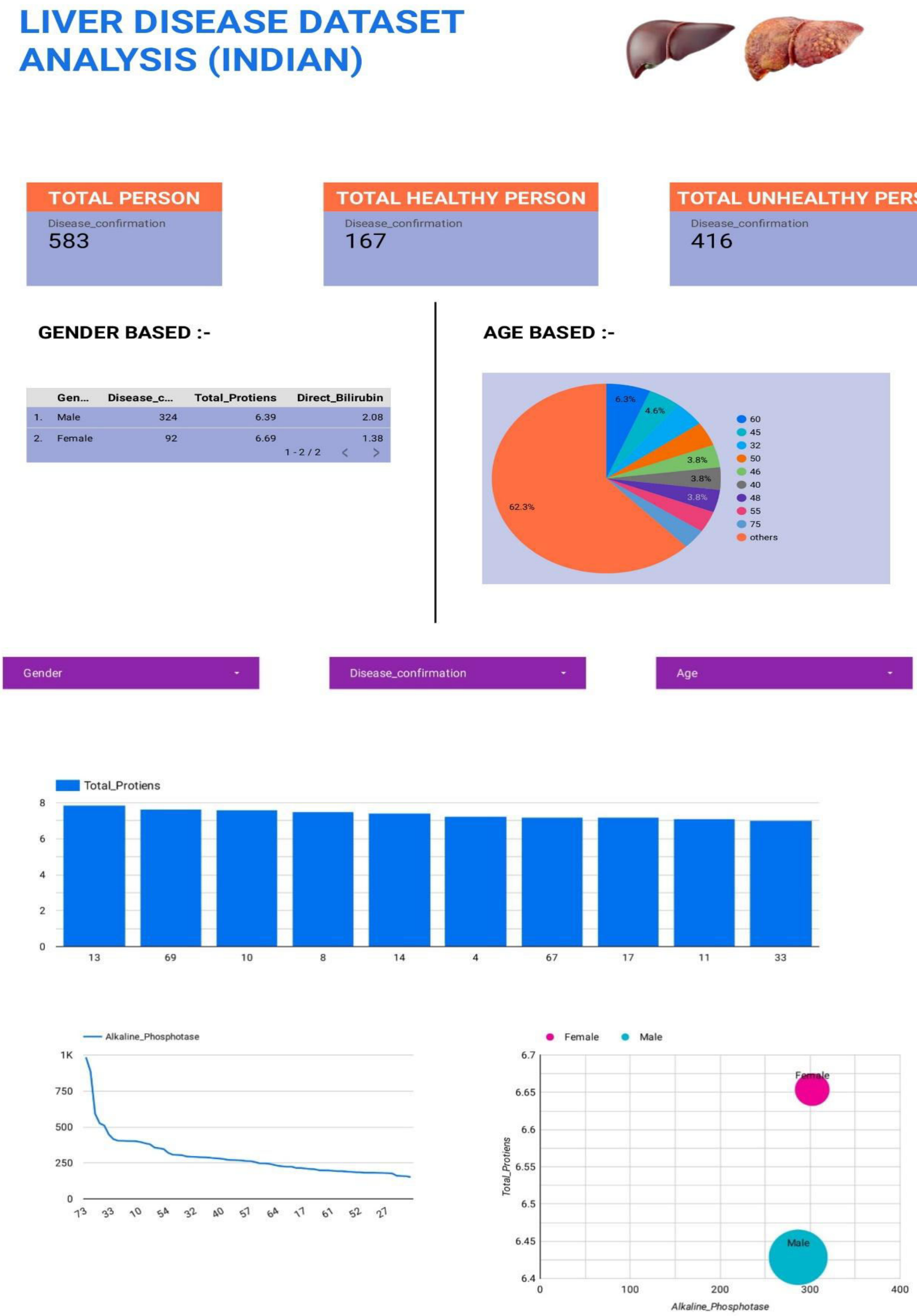
Analysis using Data Studio

About the data-

The dataset used for the analysis is available on Kaggle. The dataset consists of many parameters like protein level, AG-ratio, Age and Albumin count. These are all important factors that may affect liver in some way. We will be studying which of the above chemicals are more harmful for liver disease.

Link to dataset: - https://www.kaggle.com/sanjames/liver-patients-analysis-prediction-accuracy

The following are the graphs and chart created using data studio to understand the dataset in more detailed way.



LIVER DISEASE DATASET ANALYSIS (INDIAN)

TOTAL PERSON
Disease_confirmation
583

TOTAL HEALTHY PERSON
Disease_confirmation
167

TOTAL UNHEALTHY PERSON
Disease_confirmation
416

GENDER BASED :-

	Gen...	Disease_c...	Total_Protiens	Direct_Bilirubin
1.	Male	324	6.39	2.08
2.	Female	92	6.69	1.38

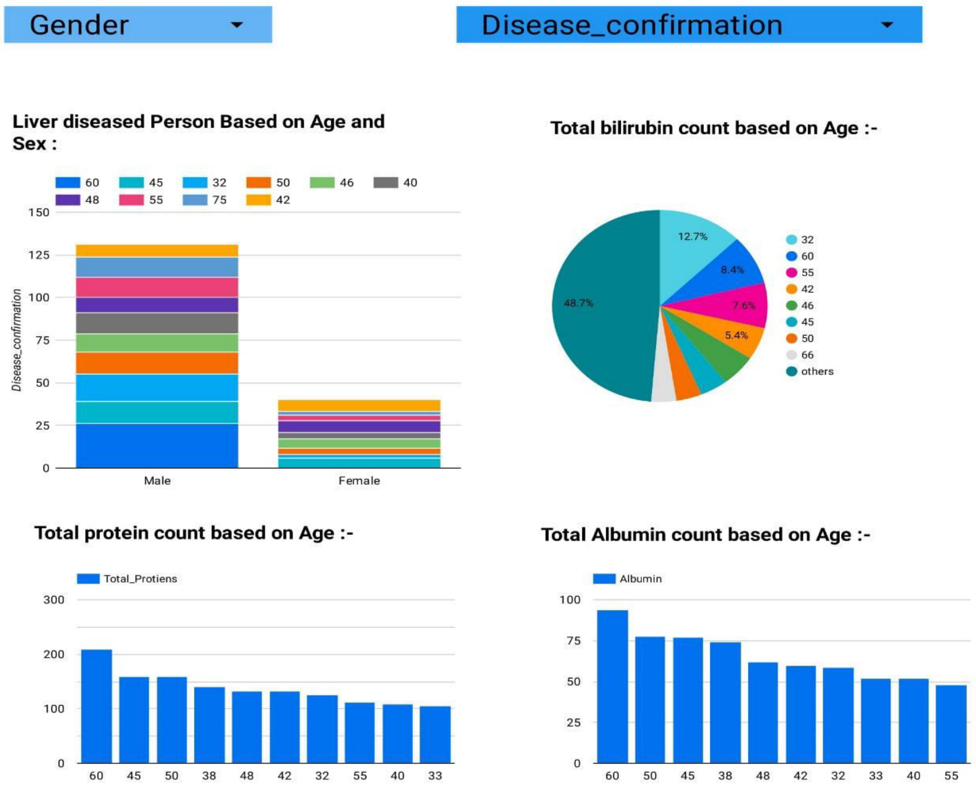
1 - 2 / 2 < >

AGE BASED :-

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905061 |



III. PROPOSED ANALYSIS APPROACH

In India, a test was conducted to know how much value of protein, albumin, albumin-globin ratio and age are there in normal and liver patient human. The data will be used to compare these parameters of liver disease and know which of them is best suitable parameter to know about liver disease. We will work on comparing both liver diseases confirm and healthy person values by using statistical approach.

Using sample data, estimate the mean protein count, AG-ratio, Albumin count, Age for both liver patients and healthy person and compare them. Find the margin of error and the confidence interval. Assume 95% confidence interval for each parameter.

1. Sample Parameter:

To compute the overall sample, mean, we need to compute the sample means for each stratum.

$$\bar{x} = \sum (x_i)/n$$

For Liver patients:

Mean (Age) = 2931/63 = 46.36

Mean (Total-Protein) = 390/63 = 6.20

Mean (Albumin) = 186/63 = 2.912

Mean (AG-Ratio) = 53/63 = 0.890

For Healthy People:

Mean (Age) = 2618/63 = 41.55

Mean (Total-Protein) = 407.8 /63 = 6.47

Mean (Albumin) = 212/63 = 3.36

Mean (AG-Ratio) = 65.77/63 = 1.068

2. Population Variance

Sample variance = $(x - x)^2/n - 1$

So, when we calculate the sample variance, we get following values: -

For Liver patients:

Age = 254.23

Total-Protein = 0.675

Albumin = 0.522

AG-Ratio = 0.097

For Healthy People:

Age = 290.15

Total-Protein = 1.10

Albumin = 0.6044

AG-Ratio = 0.07

3. Standard Error

The standard error measures the variability of our sample estimate of the population mean. We will use standard error to compute the margin of error and to define a confidence level.

- SE = SD/$\sqrt{n}$

When we calculate for all the values, we get the following results: -

Age = 2.00

Total-Protein = 0.10

Albumin = 2.91

AG-Ratio = 0.89

For Healthy People:

Age = 2.14

Total-Protein = 0.132

Albumin = 0.097

AG-Ratio = 0.034

We are working on 95% confidence level.

4. Confidence Level

There are two tests, t-test and z-test. T-test is used when we have a small sample size (<50) and when population variance is unknown whereas z-test is used when we have large sample size (>50) and population variance is known. As the sample size is large, thus we will use z-test. Standard Normal Distribution Table is used to find critical z-score. In this part of the analysis, usually researchers choose a confidence level and the most frequently chosen confidence level is 95%. Thus, we will use that only.

5. Critical Value

The critical value is a factor used to compute the margin of error. To find the critical value, we take these steps:

- Alpha(α):

  $\alpha = 1\text{-(confidence level/100)}$

  $\alpha = 1\text{-}(95/100)$

  $= 0.05$

- Critical Probability (p^*):

  $p^* = 1\text{- }(\alpha/2)$

  $= 1\text{-}(0.05/2)$

  $= 0.975$

Using Standard Normal Distribution Table, we can see that the critical z-score value is 1.96.

6. Margin of Error

ME = (Critical Value * Standard Error), where critical value = 1.96

When we calculate for all the values, we get the following results: -

Age = 1.96*2.00 = 3.92

Total-Protein = 1.96*0.10 = 0.196

Albumin = 1.96*2.91 = 5.70

AG-Ratio = 1.96*0.89 = 1.744

For Healthy People:

Age = 1.96*2.14 = 4.194

Total-Protein = 1.96*0.132 = 0.258

Albumin = 1.96*0.997 =1.95

| DOI: 10.15680/IJIRCCE.2021.0905061 |

AG-Ratio = 1.96*0.034 =0.066

7. Confidence Interval

The minimum and the maximum values of the confidence interval are:

CI_{min} = x – Standard Error * Critical Value

CI_{max} = x + Standard Error * Critical Value

When we calculate for all the values, we get the following results: -

1) Age: -
   CI_{min} = 46.36– (3.92) = 42.44
   CI_{max} = 46.36+ (3.92) = 50.28

2) Total-Protein: -
   CI_{min} = 6.20– (0.196) = 6.004
   CI_{max} = 6.20+ (0.196) = 6.396

3) Albumin: -
   CI_{min} = 0.2192 – (5.70) = -5.48
   CI_{max} = 0.2192 + (5.70) = 6.138

4) AG-Ratio: -
   CI_{min} = 0.890– (1.744) = -0.854
   CI_{max} = 0.890 + (1.744) = 2.634

For Healthy People:

1) Age: -
   CI_{min} = 41.55– (4.194) = 37.356
   CI_{max} = 41.55+ (4.194) = 45.744

2) Total-Protein: -
   CI_{min} = 6.47 – (0.258) = 6.22
   CI_{max} = 6.47 + (0.258) = 6.72

3) Albumin: -
   CI_{min} = 3.36– (1.97) = 1.39
   CI_{max} = 3.36 + (1.97) = 5.33

4) AG-Ratio: -
   CI_{min} = 1.068– (0.066) = 1.002
   CI_{max} = 1.068 + (0.066) = 1.134

IV. SUMMARY

Based on our sample data, we estimated the sample mean for liver patient and healthy patient are different. From the data studio report we can conclude that the dataset included a greater number of liver patient than healthy patient. By carefully observing the graphs we can conclude that the Albumin-Globulin ratio is very important factor which contributes to the dis-functioning of liver. We can observe that AG-Ratio for healthy patient is less than compared to liver patient. Same thing was with the protein level in healthy person it was low for them and very high for diseased people. So, we can say that protein level must be controlled in order to prevent us from liver dis-functioning.

V. FUTURE SCOPE & DISCUSSION

In this paper we have used stratified sampling and systematic sampling to analyse the data-set using statistical method. More accurate the sampling better are the results so in future we can also try the same analysis using different methods of sampling to understand the effect of using better sampling method.

Liver diseases can be inherited or caused by a variety of factors that damage the liver (virus, drugs or chemicals, obesity, diabetes, or an attack from own immune system), when the condition is left untreated, it can become life-threatening and can permanently damage the liver or the bile duct.

VI. CONCLUSION

In this research paper, we have taken a dataset consisting of different people with or without liver disease which also includes their details about count of protein, AG-Ratio, albumin count, and Age. We have used Stratified Random Sampling to create some strata within the dataset to have a uniform data. After creating those strata, we applied some statistics to it, to find the average value of each parameters neededto study their effect on contributing to liver disease as well as variance of every stratum. The majority of data consist of liver disease confirmed patients. Liver diseases can be inherited or caused by a variety of factors that damage the liver (virus, drugs or chemicals, obesity, diabetes, or an attack from own immune system), when the condition is left untreated, it can become life-threatening and can permanently damage the liver or the bile duct. This damage can then become malignant. The liver disease prognosis depends on how quickly the condition was diagnosed and treated. In the beginning stages, the liver disease usually responds to treatment, but in advanced liver disease, the damage was done by fibrosis, cirrhosis, and liver failure cannot be reversed. This advancedstage leads to eventual death. Data Analysis in health care management is unlike the other fields owing to the fact that the data present are heterogeneous and that certain ethical, legal, and social constraints apply to private medical information.

REFERENCES

1) Developing Sampling Frame for Case Study: Challenges and Conditions. By: Noriah Mohd Ishak [1] & Abu Yazid Abu Baka [2], Accessed on 02/04/2021
2) Analysis of Liver Disorder Using Data Mining Algorithm by P. Rajeshwari [1] and G Sophia Reena [2]accessed on 02/4/2021.
Link: -https://www.researchgate.net/publication/265082307_Analysis_of_Liver_Disorder_Using_Data_mining_Algorithm
3) Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms by M. Banu Priya1, P. Laura Juliet2, P.R. Tamilselvi3
Link: -irjet.net/archives/V5/i1/IRJET-V5I142.pdf
4) Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques by Jagdeep Singh[a] ,Sachin Bagga[b] and Ranjodh Kaur[c]
Link - https://www.sciencedirect.com/science/article/pii/S187705092030692X
5) Analysis of classification algorithms for liver disease diagnosis by S.R GOSH1 and Sajjad waheed2
Link- -https://www.researchgate.net/publication/319983998_Analysis_of_classification_algorithms_for_liver_disease_diagnosis
6) LIVER DISEASE PREDICTION BY USING DIFFERENT DECISION TREE TECHNIQUES Nazmun Nahar1 and Ferdous Ara2
Link - https://aircconline.com/ijdkp/V8N2/8218ijdkp01.pdf

7) Performance Evolution of Different Machine Learning Algorithms for Prediction of Liver Disease by Muktevi Srivenkatesh1
Link - https://www.ijitee.org/wp-content/uploads/papers/v9i2/L36191081219.pdf

8) Stratified-sampling over social networks using MapReduce by Roy levin1 and Yaron kanza2

Link-
https://www.researchgate.net/publication/266656479_Stratified-sampling_over_social_networks_using_MapReduce

9) Recent Developments in Systematic Sampling: A Review by Sayed mustafa1

Link- -
https://www.researchgate.net/publication/318409669_Recent_Developments_in_Systematic_Sampling_A_Review

10) H. Yuguang, and Lei Li. "Naive Bayes classification algorithm based on small sample set." In Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on, pp. 34-39. IEEE, 2011.

11) [5]. P. Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." International Journal of Computer Science and Applications 6, no. 2 (2013): 256-261.

12) [6]. K. Masud, and Rashedur M. Rahman. "Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing." Journal of Software Engineering and Applications 6, no. 04 (2013): 196.

13) [7]. D. K. SRIVASTAVA, and B. Lekha. "Data classification using support vector machine." Journal of Theoretical and Applied Information Technology 12, no. 1 (2010): 1-7.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING