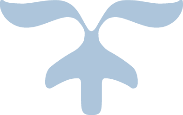


Exploring Indian Liver Patient dataset

CST8390\_23W Assignment 1



Hoang, John - 040896360  
Kingdom, Neil - 040967309  
  
Prepared for:

Dr. Anu Thomas

February 4, 2023

Algonquin College

Computer Engineering Technology  
Business Intelligence and Data Analytics

Table of Contents

[Introduction 2](#_Toc126436973)

[Information On Yeast Data Set 2](#_Toc126436974)

[Information on the Indian Liver Patient Dataset 2](#_Toc126436975)

[Dataset Selection 2](#_Toc126436976)

[Data Understanding 3](#_Toc126436977)

[Load of Database 3](#_Toc126436978)

[Description of Features 3](#_Toc126436979)

[Data Preparation 4](#_Toc126436980)

[Preprocessing – Correcting Missing Information 4](#_Toc126436981)

[Preprocessing – Removing Duplicates 5](#_Toc126436982)

[Preprocessing – Cleaning and Filtering 5](#_Toc126436983)

[Tabulated Statistics 6](#_Toc126436984)

[Interesting Charts 6](#_Toc126436985)

[Total Proteins vs. Albumin 6](#_Toc126436986)

[Total Bilirubin vs. Aspartate Aminotransferase 7](#_Toc126436987)

[Albumin and Globulin Ratio vs. Alamine Animotransferase 8](#_Toc126436988)

[Modeling and Evaluation 8](#_Toc126436989)

[Percentage of Correctly Classified by KNN 8](#_Toc126436990)

[Classification using 10-fold Cross Validation with Default Seed of 1 8](#_Toc126436991)

[Classification with Percentage Split of 70% 9](#_Toc126436992)

[Classification Using 10-fold Cross Validation with Random Seed of 10 9](#_Toc126436993)

[Manual Testing 10](#_Toc126436994)

[Conclusion 11](#_Toc126436995)

[References 12](#_Toc126436996)

# Introduction

## Information On Yeast Data Set

The yeast dataset contains data attributed to experiments that classifies protein localization phenomena. The creator of the protein localization sites of yeast is Kenta Nakai of the Institute of Molecular and Cellular Biology of Osaka University and the dataset was donated by Paul Horton. The dataset contains computed data using 8 predictive methods to distinguish data providing 8 attributes including:

1. McGeoch's method (signal sequence recognition),
2. Von Heijne's method (signal sequence recognition),
3. Score of the ALOM membrane spanning region prediction program,
4. Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins,
5. Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen),
6. Paroxysmal targeting signal in the C-terminus,
7. Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins,
8. Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

In total, the dataset contains 10 attributes including 8 predictive attributes, a name and a class distribution label pertaining to 1484 instances.

## Information on the Indian Liver Patient Dataset

The Indian Liver Patient Dataset [1] contains data to 416 liver patients and 167 non liver patients collected from north east of Andhra Pradesh, India. Data were collected and donated by 3 institution's associates: Bendi Venkata Ramana, Associate Professor of Department of Information Technology and Professor N.B. Vernkateswarlu of Department of Computer Science and Engineering both at Aditya institute of Technology and Management and Professor M. Surendra Prasad Babu of Department of Computer Science & Systems Engineering at Andhra University College of Engineering. The database has a total of 583 instances generalized as 441 male patient and 142 female patients. There are 9 other attributes and a label for a total of 11 attributes including: Age (89+ as 90), Total Bilirubin, Direct Bilirubin, Alkphos Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, Albumin and Globulin Ratio and a selector label (split data into 2 sets).

## Dataset Selection

For an educational experience on completing practical Assignment 1, we are of the opinion that the Indian Liver Patient Dataset has the effect of providing a progressive and productive learning experience. Our reasons for this effect is that we have little experience on working with data, let alone computing predictive learning algorithm on data or grouping and splitting dataset for training or testing purposes. The Yeast Dataset contains data attributes obtained from predictive learning methods that requires research or knowledge in the area of biochemistry. We also find that the Cross-Industry Standard Process for Data Mining first step is to understand the business therefore, the Indian Liver Patient Dataset containing general properties like age and gender would be the better approach to comprehension.

# Data Understanding

## Load of Database

The database has a total of 583 instances generalized as 441 male patient and 142 female patients including 416 liver patients and 167 non liver patients collected from north east of Andhra Pradesh, India.

## Description of Features

The features of the data are prepared for Weka and is described as follows:

Age is a continuous datatype and recorded as numeric because the data collected can be any number.

Gender is described as nominal in the data set that follows biological distinction, it is recorded as binary male or female.

Total Bilirubin is recorded as a numeric real number and include 2 types of bilirubin. Unconjugated (indirect) bilirubin, created from red blood cell as it breakdown and travels in the blood to the liver. Conjugated (direct) bilirubin is the bilirubin once it reaches the liver and undergoes chemical change and moves to the intestines before being removed through stool. Total Bilirubin is recorded in milligrams per deciliter (mg/dl) [2].

Direct Bilirubin is recorded as numeric. It is conjugated bilirubin that undergo a change when in the liver and is part of the bile that the liver makes to help digest food. Normal results vary depending on your age, gender and health history and should be less than 0.3 mg/dl [3].

Alkphos Alkaline Phosphatase (ALP) is recorded as numeric since it is measured as International Unit per Liter (IU/L). High alkaline phosphatase may be sign of liver problem while moderately high levels are sign of different types of condition and low levels are less common and may be sign of lack of zinc, malnutrition, pernicious anemia, thyroid diseases, Wilson disease or hypophosphatasia (a rare genetic disease that affects bones and teeth). The normal range is 44 to 147 IU/L but it is also dependant on age, gender and health history and markedly high level can be 3350 IU/L [4] [5].

Alamine Aminotransferase (ALT) is recorded as numeric and data is collected as International units per liter(IU/L). ALT measurement shows signs of liver damaged or disease. Low level of ALT is normally found in the blood but when liver is damaged it releases ALT into the bloodstream making ALT level go up. Normal level of ALT in male can be 10-40 IU/L and 7-35 IU/L in female [6].

Aspartate Aminotransferase (AST) is recorded as numeric and data is collected as International units per liter(IU/L). AST determines the balance of proteins (amino acids) and the level of AST is used as sign of liver disease and other health problems. Normal healthy level of AST in blood range from 5 to 40 IU/L where low level is uncommon that indicate sign of vitamin B6 deficiency and high AST level is above 40 IU/L [7].

Total Proteins (TP) is numeric and is measured as grams per deciliter (g/dl). The total proteins measure 2 classes of proteins albumin and globulin where albumin prevent fluid leaking out of blood vessels and helps carry chemicals in the blood while globulins are part of the immune system. Normal range is 6.0 to 8.3 g/dl and high level can indicate chronic inflammation or infection (including HIV and hepatitis B or C), multiple myeloma and Waldenstrom disease. Whereas, lower than normal level can be due to malnutrition, bleeding (hemorrhage), liver disease and many other symptoms [8].

Albumin is numeric and measured in grams per deciliter (g/dl). Low albumin levels can be sign of liver or kidney disease and high level can be sign of dehydration or high protein diets. Albumin is a protein made in the liver that carries hormones, vitamins and enzymes throughout the body. Low level of albumin can cause fluid to leak out of blood and build up in the lungs, abdomen(belly) or other parts of your body. Normal range of albumin is 3.4 to 5.4 g/dl [9].

Albumin and Globulin Ratio (A/G ratio) is numeric and is measure as a ratio. It consists of the ratio of albumin and globulin where as albumin makes up 50 percent of serum proteins representing nutritional status and globulin make up around 48 percent of serum protein indicating the state of the immune function and the severity of any inflammation. Normal range for albumin to globulin is over 1 to 2 because there is usually more albumin than globulin in serum protein. Low A/G ratio indicate low albumin levels or high globulin and can be a sign of kidney disease, liver disease (including liver cancer), autoimmune disease and many other health condition. High A/G ratio means high albumin or low globulin and can be due to severe dehydration or diarrhea but can also occur during pregnancy [10].

Expert Selector Field split data into two set is a label type which is treated as binary nominal of 1 or 2. It seems that the records labeled as 1 are liver patient where there are 416 set and 2 indicate non liver patient with 167 set. This field is later changed to binary 0 for patient and 1 for non-patient.

# Data Preparation

## Preprocessing – Correcting Missing Information

On preparing the dataset, there were some missing data particularly on albumin to globulin ratio (A/G ratio) in 4 data set, 2 of male of age 27 (non-liver patient) and age 51 (liver patient) and on 2 of female of age 35 (non-liver patient) and age 45 (liver patient). Correction to the data can be made by simply using the Total proteins features of each instance [11]. Subtracting total protein by albumin to get globulin then finding the ratio of albumin to globulin by dividing albumin by globulin.

Example: Globulin = Total Proteins – Albumin  
 A/G ratio = Albumin / Globulin

This seems to be the case with many of the data set instances except some ratio seems to have calculated A/G ratio with truncated 2nd significant digit. For the purpose of this report for the missing data on A/G ratio the record is calculated using the method above.

The updated feature to A/G ratio on the missing records are as follows:

* Female age 45 a liver patient with total protein of 6.6 g/dl and 3.9 g/dl albumin with updated globulin of 2.7 g/dl for a ratio of 1.44 g/dl,
* Female age 35 a non liver patient with total protein of 5.2 g/dl and 2.7 g/dl albumin with updated globulin of 2.5 g/dl for a ratio of 1.08 g/dl,
* Male age 51 a liver patient with total protein 6.5 g/dl and 3.1 g/dl albumin with updated globulin of 3.4 g/dl for a ratio of 0.91 g/dl,
* Male age 27 a non liver patient with total protein 8.5 g/dl and 4.8 g/dl albumin with updated globulin of 3.7 g/dl for a ratio of 1.3 g/dl.

## Preprocessing – Removing Duplicates

On removing duplicates to remove biases in the data set, simple filtering using Weka to update the data set is used. This resulted in removal of 13 instances of which 11 males and 2 female and 10 liver patients and 3 non liver patients. The update result in a total of 570 instances with 430 male and 140 female patients including 406 liver patients and 164 non liver patients.

## Preprocessing – Cleaning and Filtering

For this exploratory report on the Indian Liver Patient Data set the focus is to build a predictive model by classification of the model to evaluate and test the predictive model. On preserving all the features of the data set to be used on building the predictive model the gender feature is converted to it’s binary nominal format 0 for male and 1 for female using the Weka built-in filtering tool “NominalToBinary”.

The Selector Label is used to classify instances as patients (1) or non-patients (2). Weka presumes that the type ought to be numeric for these attributes, but it was changed to be a nominal type so that data classification can be applied.

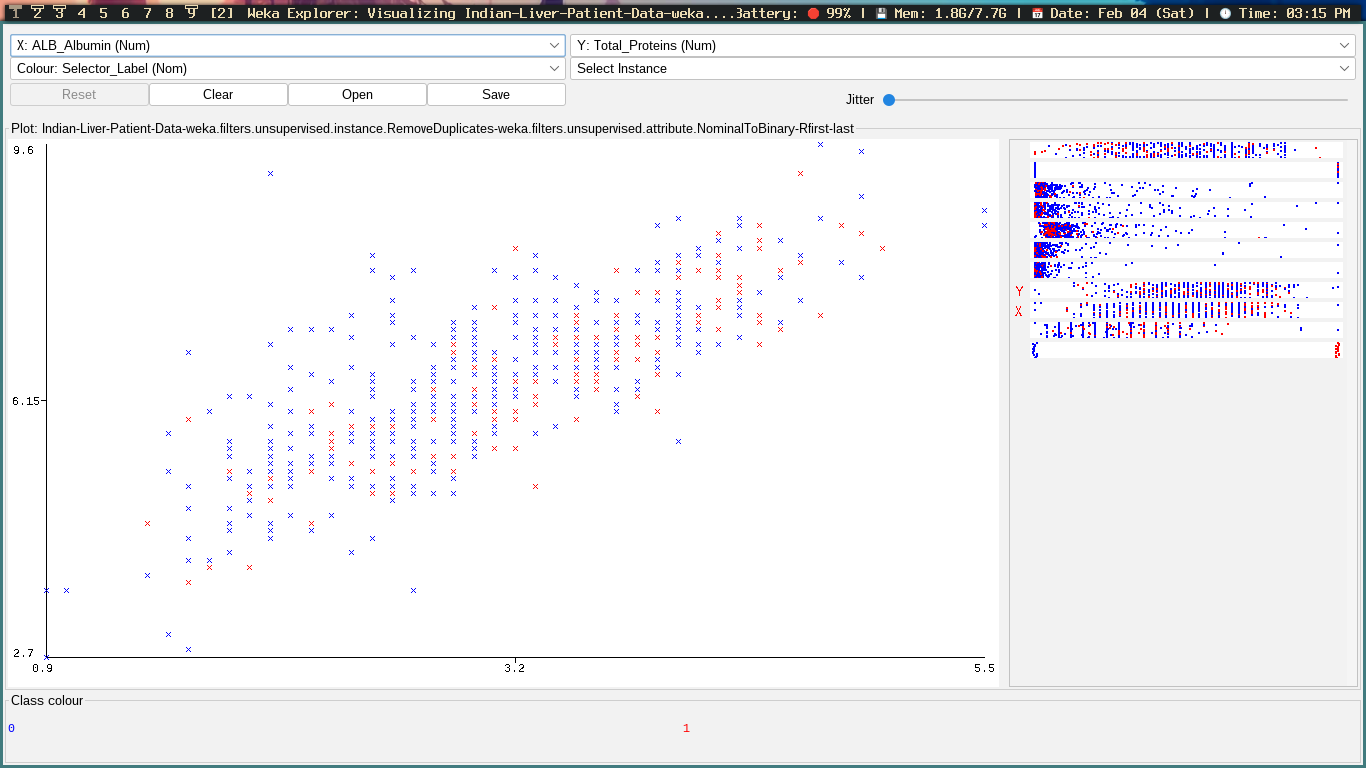
On cleaning up the data, 1 record stands out where Male of age 55 recorded as a patient has total bilirubin of 75 mg/dl which is way above the known and possible level (the usual range is between 20 to 25 mg/dl and may reach 50 mg/dl [12]). A bilirubin level below 50 mg/dl makes sense where all records from the Indian Liver Patient data set provided are under 50 mg/dl. Since total bilirubin is relative to direct bilirubin, looking at direct bilirubin level of the record in question its observed that the record indicates 3.6 mg/dl. Making relative comparison with other records consisting of direct bilirubin it can be speculated that there was an error in recording the record. As oppose to the impossible recording of 75 mg/dl of total bilirubin it makes more sense that the level is most likely around 7.5 mg/dl and thus, was updated to reflect our assumption.

## Tabulated Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Minimum** | **Maximum** | **Mean** | **StdDev** |
| Age | 4 | 90 | 44.849 | 16.242 |
| Gender | 0 | 1 | 0.246 | 0.431 |
| Total Bilirubin | 0.4 | 42.8 | 3.303 | 5.502 |
| Direct Bilirubin | 0.1 | 19.7 | 1.498 | 2.833 |
| ALP Alkaline\_P | 63 | 2110 | 291.751 | 245.292 |
| ALT Alamine\_A | 10 | 2000 | 79.728 | 181.472 |
| AST Aspartate\_A | 10 | 4929 | 109.381 | 290.881 |
| Total Protein | 2.7 | 9.6 | 6.496 | 1.088 |
| Albumin | 0.9 | 5.5 | 3.149 | 0.797 |
| A/G Ratio | 0.3 | 2.8 | 0.95 | 0.32 |
| Selector | 0 | 1 | 0.288 | 0.453 |

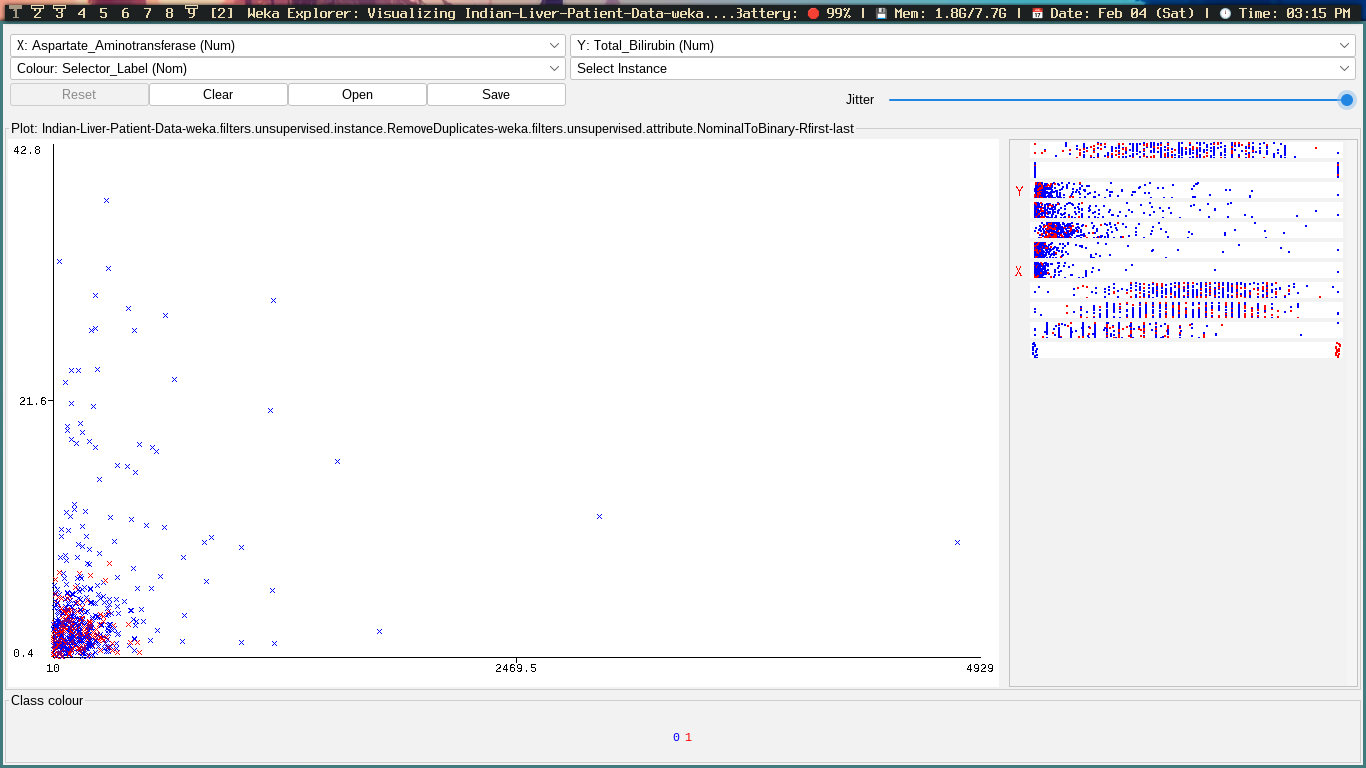
## Interesting Charts

### Total Proteins vs. Albumin



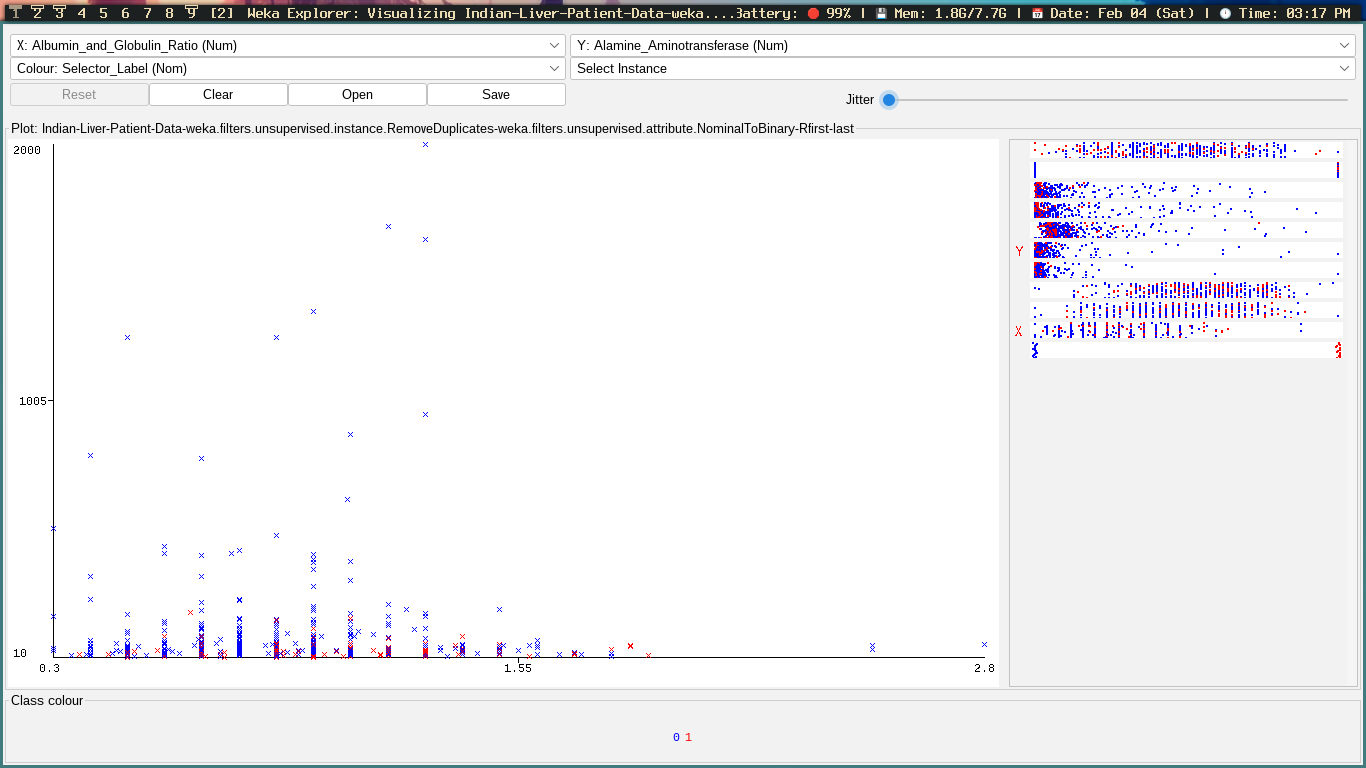
A positive correlation between the total proteins and Albumin levels can be seen in the plot above. This is to be expected, as Albumin is one of the two categories of proteins measured by total proteins (the other being globulin), but it is still worth seeing this correlation visually.

### Total Bilirubin vs. Aspartate Aminotransferase



An example of clustering between two unrelated attributes is the relationship between total Bilirubin and Aspartate Aminotransferase. There is a very big overlap between the two with a few outliers in non-patients in terms of both total Bilirubin and Aspartate Aminotransferase levels (Jitter function at maximum).

### Albumin and Globulin Ratio vs. Alamine Animotransferase



In the relationship above, it can be noted that there is a small amount of standard deviation where Alamine Aminotransferase is at its peak when the Albumin and Globulin Ratio is approximately 0.6%.

# Modeling and Evaluation

### Percentage of Correctly Classified by KNN

|  |  |
| --- | --- |
| **K** | **Percentage Correctly Classified** |
| 3 | 64.0351% |
| 5 | 65.4386% |
| 7 | 65.0877% |
| 9 | 65.9649% |
| 11 | 65.0877% |

### Classification using 10-fold Cross Validation with Default Seed of 1

#### Best K’s Accuracy at K = 9 Nearest Neighbour

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **True Positive Rate** | **False Positive Rate** | **#Misclassification** |
| **0** | 0.813 | 0.720 | 76 |
| **1** | 0.280 | 0.187 | 118 |

#### Worse K’s Accuracy at K = 3 Nearest Neighbour

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **True Positive Rate** | **False Positive Rate** | **#Misclassification** |
| **0** | 0.776 | 0.695 | 91 |
| **1** | 0.305 | 0.224 | 114 |

### Classification with Percentage Split of 70%

On 70% split classification the predictive model verifies on 171 instances the test model.

#### Percentage Correctly Classified with 70% split

|  |  |
| --- | --- |
| **K** | **Percentage Correctly Classified** |
| 3 | 66.0819% |
| 5 | 67.8363% |
| 7 | 69.0058% |
| 9 | 67.2515% |
| 11 | 68.4211% |

#### Best K’s Accuracy at K = 7 Nearest Neighbour

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **True Positive Rate** | **False Positive Rate** | **#Misclassification** |
| **0** | 0.869 | 0.755 | 16 |
| **1** | 0.245 | 0.131 | 37 |

#### Worse K’s Accuracy at K = 3 Nearest Neighbour

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **True Positive Rate** | **False Positive Rate** | **#Misclassification** |
| **0** | 0.811 | 0.714 | 23 |
| **1** | 0.286 | 0.189 | 35 |

### Classification Using 10-fold Cross Validation with Random Seed of 10

|  |  |
| --- | --- |
| **K** | **Percentage Correctly Classified** |
| 3 | 65.2632% |
| 5 | 65.614% |
| 7 | 65.2632% |
| 9 | 66.8421% |
| 11 | 67.193% |

#### Best K’s Accuracy at K = 11 Nearest Neighbour

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **True Positive Rate** | **False Positive Rate** | **#Misclassification** |
| **0** | 0.842 | 0.750 | 64 |
| **1** | 0.250 | 0.158 | 123 |

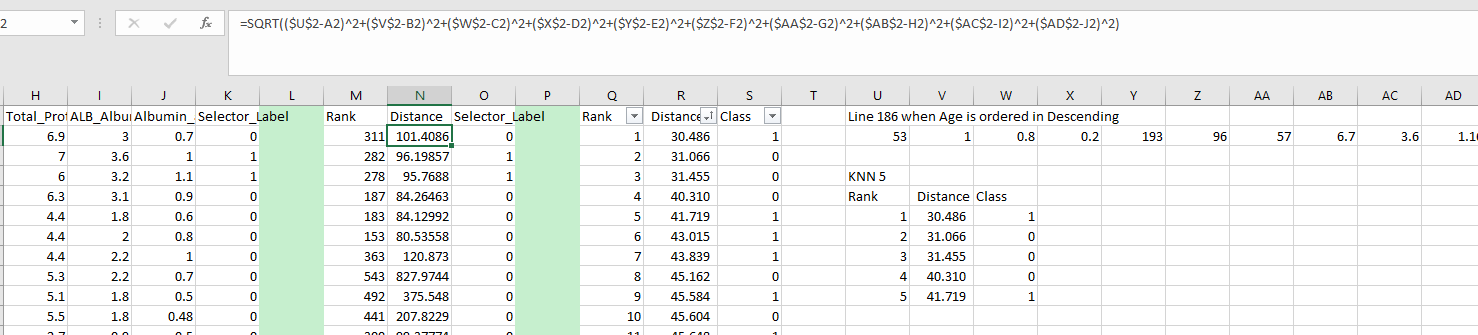
#### Worse K’s Accuracy - 2 Cases 1rst at K = 3

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **True Positive Rate** | **False Positive Rate** | **#Misclassification** |
| **0** | 0.778 | 0.659 | 90 |
| **1** | 0.341 | 0.222 | 108 |

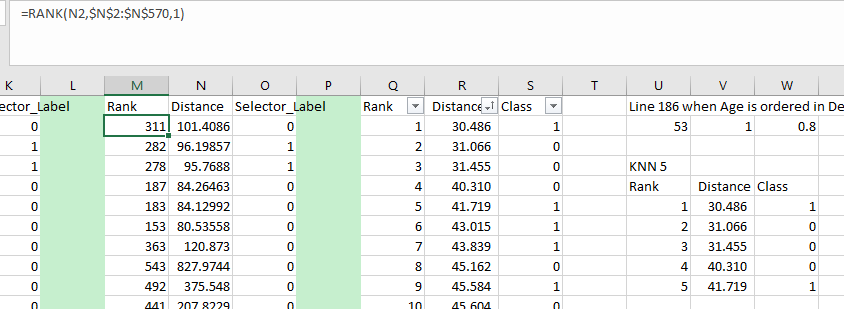
#### Worse K’s Accuracy - 2 Cases 2nd at K = 7

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **True Positive Rate** | **False Positive Rate** | **#Misclassification** |
| **0** | 0.800 | 0.713 | 81 |
| **1** | 0.287 | 0.200 | 117 |

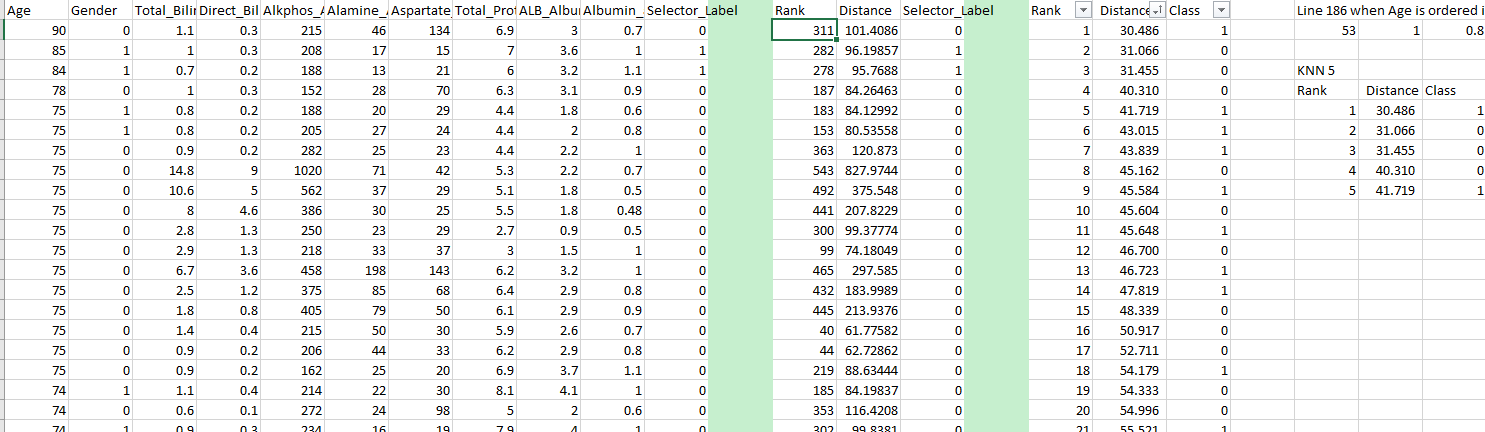
# Manual Testing



In order to perform manual testing, the preprocessed dataset is first imported into Excel. A Rank and Distance column are created to track the nearest Euclidian distance when using the nearest neighbour linear search algorithm for predicting the Sector Label (Class) of a randomly selected instance from the dataset. The randomly selected instance’s attributes can be seen layer out on row 2, columns U-AD. Euclidian distance is calculated using the Pythagorean Theorem to calculate the hypotenuse, where each attribute of the instance may be considered as the distance along a plane that exists within a single dimension. In practice, this means taking the difference between the selected instance’s attributes and the corresponding attributes of each alternative instance in the dataset. These distances are squared, and summed before the square root is applied to retrieve the distance of the hypotenuse - the Euclidian distance.



The Euclidian distances, once calculated, are ordered from least to greatest (ascending). For KNN where K = 5, the instances with the top 5 distances are considered. The Class of the selected instance is estimated by looking at the ratio of Classes amongst the top 5 ranked candidates and choosing the Class with a higher weight or probability (in this case, 0).



# Conclusion

Studying the Indian Liver Patient dataset was a fascinating venture into the domain of biology and cultural demographics. Although we did not conduct any studies for this assignment, the dataset has been used to conduct two separate peer-reviewed studies. The first study [13] looked at the correlation between liver patients in both India and the U.S.A, and the second [14] looked at methods for effective data classification to aid in diagnosing Liver Disease. The conclusion of the latter paper demonstrated that age, sex, SGOT, SGPT, ALP, total Bilirubin, direct Bilirubin, total proteins, and Albumin are all crucial attributes for determining the status of the liver. It is interesting that we were able to analyse the same dataset used in these two studies, and that we were able to use one of the recommended algorithms for data classification (KNN), in order to determine whether or not an instance was more or less likely to be considered a patient or non-patient, respectively.

# References

|  |  |
| --- | --- |
| [1] | D. Dua and C. Graff, "ILPD (Indian Liver Patient Dataset) Data Set;School of Information and Computer Science," UCI Machine Learning Repository;University of California, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29. [Accessed 25 January 2023]. |
| [2] | W. E. Contributors, "What Is a Bilirubin Test?;Total Bilirubin," WebMD, 6 February 2021. [Online]. Available: https://www.webmd.com/a-to-z-guides/bilirubin-test. [Accessed 30 January 2023]. |
| [3] | Chad Haldeman-Englert MD;Raymond Turley Jr PA-C;Tara Novick BSN MSN, "Health Encyclopedia;Direct Bilirubin," University of Rochester Medical Center, 2023. [Online]. Available: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=bilirubin\_direct. [Accessed 30 Januar 2023]. |
| [4] | L. MD, H. SM and L. DG, "Alkaline Phosphatase Pathophysiology with Emphasis on the Seldom-Discussed Role of Defective Elimination in Unexplained Elevations of Serum ALP – A Case Report and Literature Review," Dovepress, 7 December 2021. [Online]. Available: https://www.dovepress.com/alkaline-phosphatase-pathophysiology-with-emphasis-on-the-seldom-discu-peer-reviewed-fulltext-article-CEG. [Accessed 30 January 2023]. |
| [5] | MedlinePlus, "Alkaline Phosphatase," MedlinePlus, 3 August 2022. [Online]. Available: https://medlineplus.gov/lab-tests/alkaline-phosphatase/. [Accessed 30 January 2023]. |
| [6] | H. Staff, "Alanine Aminotransferase (ALT) Test," Health Link BC, 17 June 2021. [Online]. Available: https://www.healthlinkbc.ca/tests-treatments-medications/medical-tests/alanine-aminotransferase-alt. [Accessed 30 January 2023]. |
| [7] | B. Joe Cohen, "High/Low Aspartate Aminotransferase Levels + Function & Diseases," SelfDecode Labs, 15 December 2022. [Online]. Available: https://labs.selfdecode.com/blog/aspartate-aminotransferase-ast/. [Accessed 30 January 2023]. |
| [8] | M. Sinai, "Total protein," Mount Sinai, 2023. [Online]. Available: https://www.mountsinai.org/health-library/tests/total-protein. [Accessed 30 Januar 2023]. |
| [9] | MedlinePlus, "Albumin Blood Test," MedlinePlus, 7 June 2022. [Online]. Available: https://medlineplus.gov/lab-tests/albumin-blood-test/. [Accessed 30 January 2023]. |
| [10] | K. Nunez, "What Does a High A/G Ratio Test Result Mean?," HeathLine, 9 June 2022. [Online]. Available: https://www.healthline.com/health/a-g-ratio-high. [Accessed 30 January 2023]. |
| [11] | K. Maladkar, "5 Ways To Handle Missing Values In Machine Learning Datasets," AIM, 9 February 2018. [Online]. Available: https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/. [Accessed 30 January 2023]. |
| [12] | R. E. Pyeritz, B. R. Korf and W. W. Grody, "Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics," Science Direct, 2020. [Online]. Available: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/bilirubin-blood-level. [Accessed 30 January 2023]. |
| [13] | B. V. Ramana, P. M. S. P. Babu and P. N. B. Venkateswarlu, "A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis," International Journal of Computer Science, 2012. [Online]. Available: https://ijcsi.org/papers/IJCSI-9-3-2-506-516.pdf.. [Accessed 4 February 2023]. |
| [14] | B. V. Ramana, P. M. S. P. Babu and P. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," International Journal of Database Management Systems (IJDMS), 2011. [Online]. Available: https://www.airccse.org/journal/ijdms/papers/3211ijdms07.pdf. [Accessed 4 February 2023]. |