[Healthcare Analytics Project]

- Hepatocellular Carcinoma Dataset -

by

Seongkyoung Ryu	0725164
Bhavya vinod	0735416
Amrutha Robins	0731913
Jils Joseph	0730354
Santhosh Addanki	0732646

TABLE OF CONTENTS

1. INTRODUCTION

The liver is a part of our digestive system and it is the largest solid organ in our body. Which is on the right side of the belly, it normally weighs around 3 pounds and in a reddish-brown color. The liver has two large sections, called the right and left lobes. The liver, pancreas, and intestines work together to digest, absorb, and process food. The liver's main job is to filter the blood coming from the digestive tract, before passing it to the body.

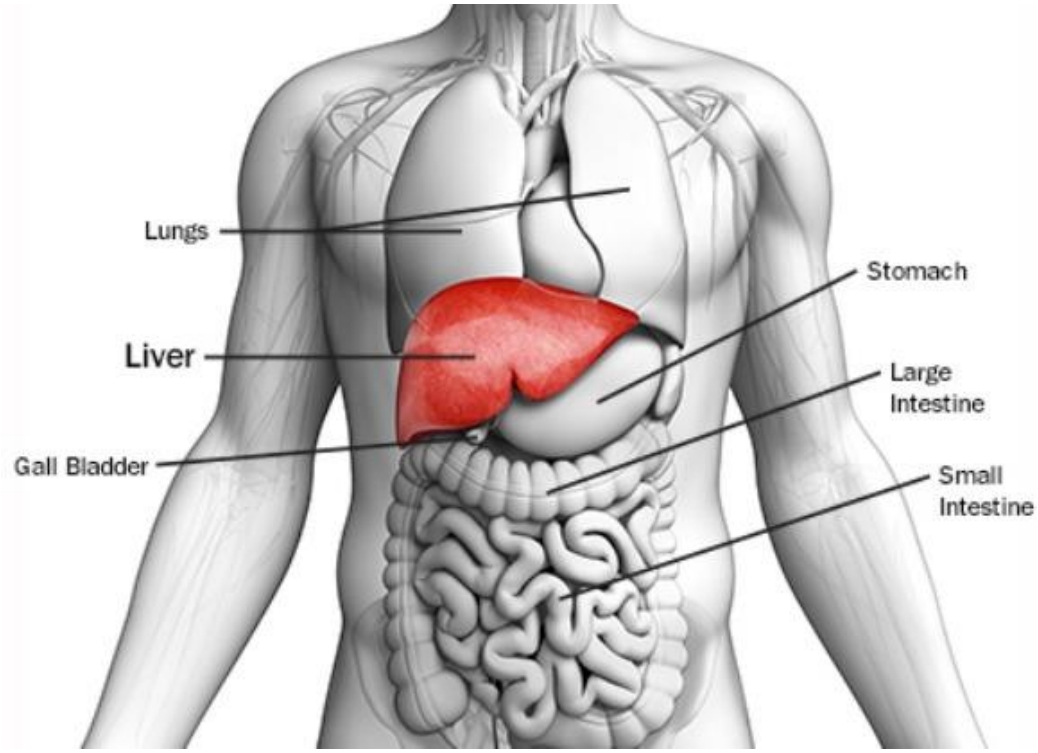

Fig1.1 Human digestive system

The liver makes bile, and it helps the small intestine to digest and absorb fats and it also helps to absorb cholesterol and some vitamins, especially vitamin K This vitamin plays a key role in blood clotting, bone metabolism and in the regulation of blood calcium levels. It also plays a major role in detoxification. Through our observation, we understood that liver diseases are inherited or happened because of various damages that happened to the liver or through absorbing a large amount of iron through your food.

Liver failure is a very serious condition and it requires immediate actions to save a life. Most of the time liver failure occurs gradually, and it will be taking many years. This mainly occurs when liver becomes damaged beyond repair and the liver can't work anymore. There are two types of liver failure: Acute and Chronic. In acute failure liver stops functioning in a matter of days or weeks. In chronic failure liver damage happens over time and gradually it stops working.

Some liver problems can be treated with lifestyle modifications, such as controlling the consumption of alcohol or by regulating obesity. Other liver problems need to be treated with medications and may require surgery. There are various methods are available to find out about liver damage like blood tests, imaging tests, and tissue analysis (biopsy) and based on the diagnosis, we need to confirm the treatment. Treatment for liver disease that causes or has led to liver failure may ultimately require a liver transplant.

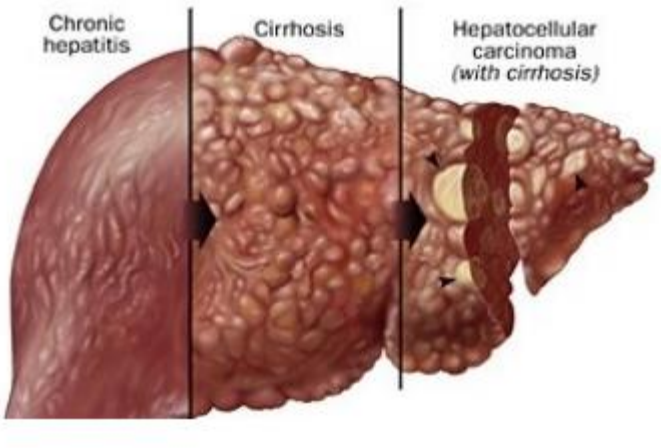

Fig1.2 different Liver conditions

To keep our liver healthy is very important for maintaining health and for that food plays a major role and inclusion of some foods into our diet can improve its healthy like coffee or tea as a beverage, grapefruit which contains many antioxidants and it can protect the liver from various injuries. Similarly, blueberries and cranberries can also improve the health of our liver. Making this food a regular part of your diet can gradually improve the immunity of our liver.

Approximately, 80% of liver cancers start in a type of liver cell called the hepatocytes. Most other liver cancers arise from cells of the bile ducts

There are three types of cancers

- Hepatocellular carcinoma
- Cholangiocarcinoma
- Hepatoblastoma

A major cause of liver cancer is the hepatitis virus, resulting from hepatitis B, C. When the liver tries to replace damaged tissues, it can lead to DNA replication and can lead to liver cancer.

HEPATOCELLULAR CARCINOMA (ICD-10: C22.0 LIVER CELL CARCINOMA)

Hepatocellular carcinoma (**HCC**) is the most common type of primary liver cancer in adults, and is the most common cause of death in people with cirrhosis.[1]

Fig 1.1 Shows worldwide HCC incident rate in 2018

It occurs in the setting of chronic liver inflammation and is most closely linked to chronic viral hepatitis infection (hepatitis B or C) or exposure to toxins such as alcohol or aflatoxin. Certain diseases, such as hemochromatosis and alpha 1-antitrypsin deficiency, markedly increase the risk of developing HCC. Metabolic syndrome and NASH are also increasingly recognized as risk factors for HCC.[2]

Metabolic syndrome, sometimes known by other names, is a clustering of at least three of the five following medical conditions: central obesity, high blood pressure, high blood sugar, high serum triglycerides, and low serum high-density lipoprotein (HDL).

HCC is one of the most prevalent cancers in the world. In some regions in the world, HCC is much more common than other type of cancer most likely due to increased frequency of hepatitis B infection or exposure to aflatoxins. One estimate places HCC as the fifth most common cancer that leads to death in the world. HCC accounts for about 85%-90% of all primary liver cancers. Studies are made to find the major causes of this cancer and stated that it is due to underlying liver disorder including infection with hepatitis B or C viruses, scarring of the liver (cirrhosis), non-alcoholic fatty liver disease, a condition in which fat builds up in the liver along with inflammation and liver cell damage, Heavy drinking, Cigarette[1].From the data collected from different patient we are trying to understand which factor is the most common cause, whether there is any more factor or is there any correlation between different factors.

Hepatocellular Carcinoma dataset (HCC dataset) was collected from a University Hospital in Portugal. It contains real clinical data of 165 patients diagnosed with HCC. The dataset contains 49 features selected according to the EASL-EORTC (European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer).

2. DATASET DESCRIPTION

HCC- DATA-COMPLETE BALANCED

1. HBsAg - + ve means you are infected and can spread hepatitis B virus to others through your blood. Hepatitis B infection of liver – liver failure and cancer.
2. HBeAg – Active viral replication – likely transmit the virus to another person.
3. HBcAb – You are infected and can spread the hepatitis B virus to others through your blood.
4. HCVAb – Hepatitis C – liver infection – Passed on through husing contaminated needles and syringes or sharing other items with infected blood on them. It's also sexually transmitted infection – less common.
5. Cirrhosis – Cirrhosis is a late state of scarring of the liver caused by many forms of liver diseases and conditions such as hepatitis and chronic alcoholism
6. Endemic – Particular people or in a certain area
7. CRI – Cancer Research Institute, they are working to advance immunotherapy as a viable treatment for people affected with these diseases.
8. Varices – Esophagel varices are abnormal, enlarged veins in the tube that connects the throat and stomach (esophagus). This occurs when normal blood flow is blocked because of the blood clot or scar tissue, because of the blockages the blood will flow through other vessels, which is not designed to carry a large amount of blood and it can lead to rupture and bleeding.
9. Spleno – Hepatosplenomegaly is a disorder where liver and spleen swell beyond their normal size. Spleen helps in identifying pathogens which are bacteria and microorganisms, it creates antibodies to fight with them. The main causes obesity, alcohol addiction, hepatitis, diabetes
10. NASH – Non-Alcoholic Steatohepatitis. Steatohepatitis means a fatty liver with inflammation and the damage is like that of alcoholic liver disease. But, in this case it is going to affect people with no history of alcoholism or to whom with moderate drinking habits. NASH is chronic yet silent disease, it's a slowly

progressive disease. NASH can progress to cirrhosis, liver cancer and then liver failure.

11.

i. **Causes**

A, Obesity

B, Genetic factors and drug consumption

C, Diabetes

12. Hemochromatosis - causes your body to absorb too much iron from the food you eat – too much iron – liver diseases, heart problems and diabetes.
    a. Hereditary – never have symptoms (joint pain, weakness, abdominal pain)
    b. Later sign - diabetes, heart failure, liver failure, loss of sex drive)
    c. Mutation – parents to children
        i. Mutations of hepcidin genes
            1. Autoimmune disease
13. PHT - Liver disease can cause what is known as “portal hypertension,” meaning increased blood pressure in the veins that enter the liver. This increased pressure causes blood to bypass the liver.
14. PVT - Portal vein thrombosis (PVT) is a blood clot of the portal vein, also known as the hepatic portal vein. This vein allows blood to flow from the intestines to the liver. A PVT blocks this blood flow. Although PVT is treatable, it can be life-threatening. There are number of risk factors for developing this condition. Such as cancer, liver disease, inflammation of the pancreas, appendicitis.
15. Encephalopathy - decline in brain function that occurs as a result of severe liver disease. When this happens liver will not be able to remove toxic contents and it can lead to brain damage.
16. Ascites - Ascites is the accumulation of protein-containing (ascitic) fluid within the abdomen. Many disorders can cause ascites, but the most common is high blood pressure in the veins that bring blood to the liver (portal hypertension), which is usually due to cirrhosis.
17. AFP - An AFP level of less than 10 ng/mL is normal for adults. An extremely high level of AFP in your blood—greater than 500 ng/mL—could be a sign

of liver tumor. High levels of AFP may mean other cancers, including Hodgkin disease, lymphoma, and renal cell carcinoma (kidney cancer).

18. ALT - The alanine aminotransferase (ALT) test is a blood test that checks for liver damage. The normal range of values for ALT (SGPT) is about 7 to 56 units per liter of serum.
19. AST - The aspartate aminotransferase (AST) test is a blood test that checks for liver damage. The normal range of values for AST (SGOT) is about 5 to 40 units per liter of serum (the liquid part of the blood).
20. ALP - An alkaline phosphatase level test (ALP test) measures the amount of alkaline phosphatase enzyme in your bloodstream. The normal range of ALP varies from person to person and depends on your age, blood type, gender, and whether you're pregnant. The normal range for serum ALP level is 20 to 140 IU/Trusted Source, but this can vary from laboratory to laboratory. ALP level is higher in children and it decreases with the age.

An example of a test result with test value rage is provided at [A]

3. EXPLORERING DATASET

UNDERSTANDING THE DATASET

```
#Prints the first 5 entries from the csv file
hcc.head()
```

	gender	symptoms	alcohol	hepatitis_b_surface_antigen	hepatitis_b_e_antigen	hepatitis_b_core_antibody	hepatitis_c_virus_antibody	cirrhosis	endemic_cour
0	1	1.0	1	0.0	0.0	0.0	0.0	1	
1	1	0.0	1	0.0	0.0	0.0	0.0	1	
2	1	0.0	0	0.0	0.0	1.0	1.0	1	
3	1	1.0	1	0.0	0.0	0.0	0.0	1	
4	1	0.0	1	0.0	0.0	0.0	1.0	1	

5 rows × 51 columns

```
#prints the number of rows and number of columns
hcc.shape
```

```
(165, 51)
```

Fig 3.1 Shape of the dataset

The CSV file contains 165 rows and 51 columns.

```
dtypes: float64(45), int64(6)
memory usage: 65.9 KB
None
```

Fig 3.2 Data types

```
hcc["class_attribute"].value_counts()
#gives each count of the status type
```

```
1    102
0     63
Name: class_attribute, dtype: int64
```

Fig 3.3 Class value count

5 columns are of the integer data type and 45 are Float data type

The datatype of the Column “Class attribute” is an integer, it can be converted to a categorical datatype Yes or No

In the Class column, the value 1 means the patient has survived 1 years and the value 0 means the patient died within 1 year.

The value counts () function tells how many data points for each class are present. Here, it tells how many patients survived and how many did not survive.

Out of 165 patients, 102 patients survived and 63 did not.

MISSING DATA

In order not to influence on the analysis, it is necessary to remove or fill the missing values. Please refer to below table for missing values status.

	Total	Percent
oxygen_saturation_%	80	0.484848
ferritin	80	0.484848
iron	79	0.478788
packs_of_cigarets_per_year	53	0.321212
esophageal_varices	52	0.315152
grams_of_alcohol_per_day	48	0.290909
direct_bilirubin_mg/dL	44	0.266667
smoking	41	0.248485
hepatitis_b_e_antigen	39	0.236364
endemic_countries	39	0.236364
hepatitis_b_core_antibody	24	0.145455
hemochromatosis	23	0.139394
nonalcoholic_steatohepatitis	22	0.133333
major_dimension_of_nodule_cm	20	0.121212
symptoms	18	0.109091

Fig 3.4 Percentage of null values

We filled the missing values with the most frequent values in each class(death-0, lived-1). Because we have only 165 patients' data. It is needed to keep all rows as possible as we can.

Nan values in continuous columns are replaced using knn imputation(k=3).

EXAMINE PATIENTS

SUMMARY OF DATASET

data['age']. isnull (). sum ()

The oldest patient: 93 years.

The youngest patient: 20 years.

Average age: 64.69090909090909 years.

Median age: 66.0 years.

AGE RANGE

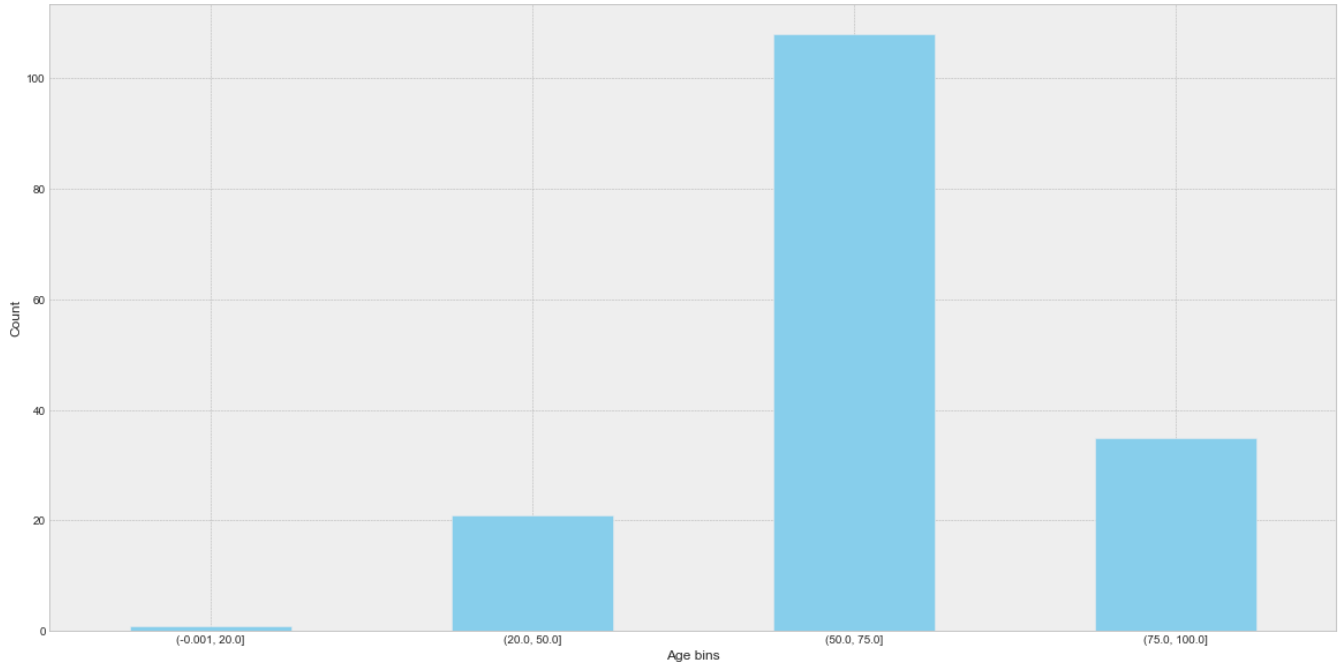

Fig 3.5 Age range

After binning and plots analysis we can see that the largest proportion of survivors/died patients were in age range of 50-75.

DEATH RANGE

The youngest and the oldest patient is 20 years old and 93 years old respectively. Both portions are quite similar. However, red portion is skewed to the right more. The death rate of the older people is somewhat higher than younger.

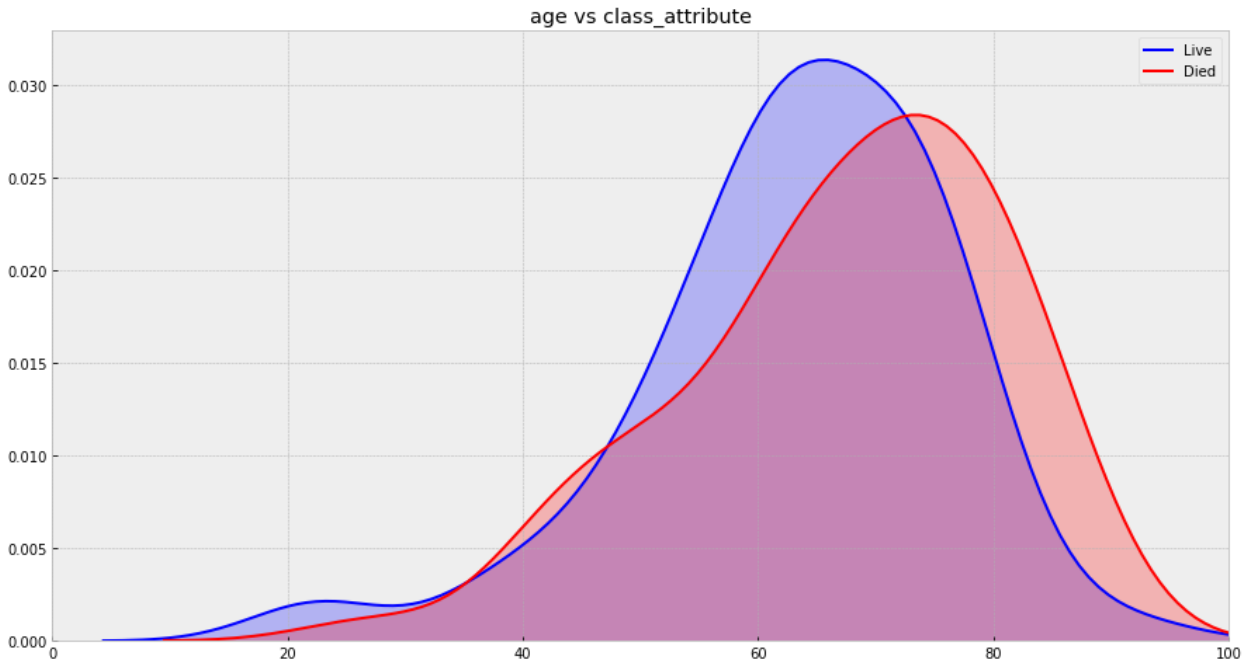

Fig 3.6 Death rate

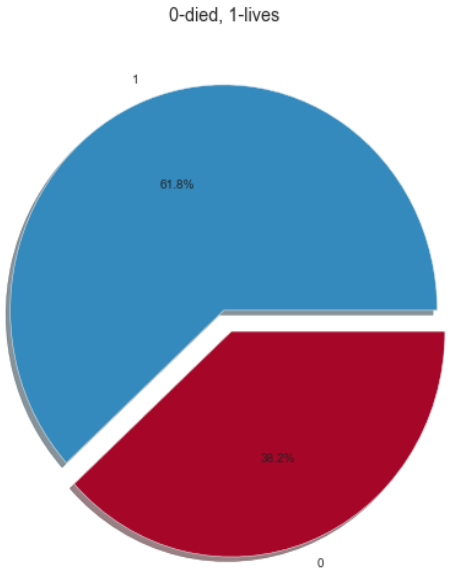

Fig 3.7 Percentage of class attribute

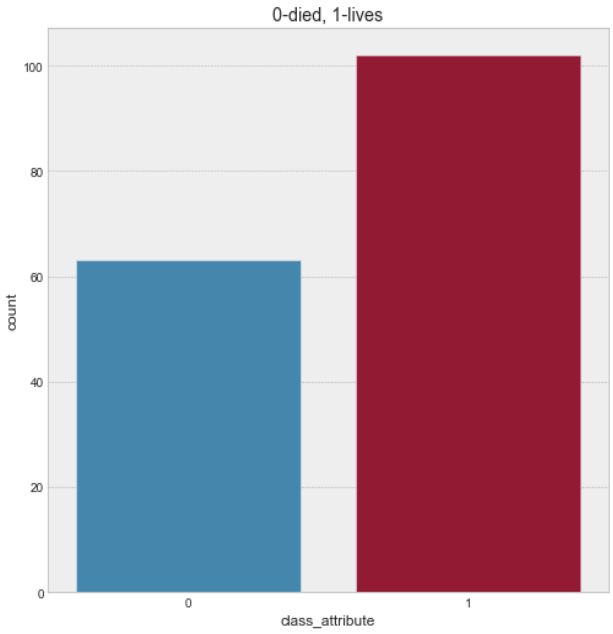

Fig 3.8 Count of class attributes

GRAMS OF ALCOHOL PER DAY VS CLASS ATTRIBUTE

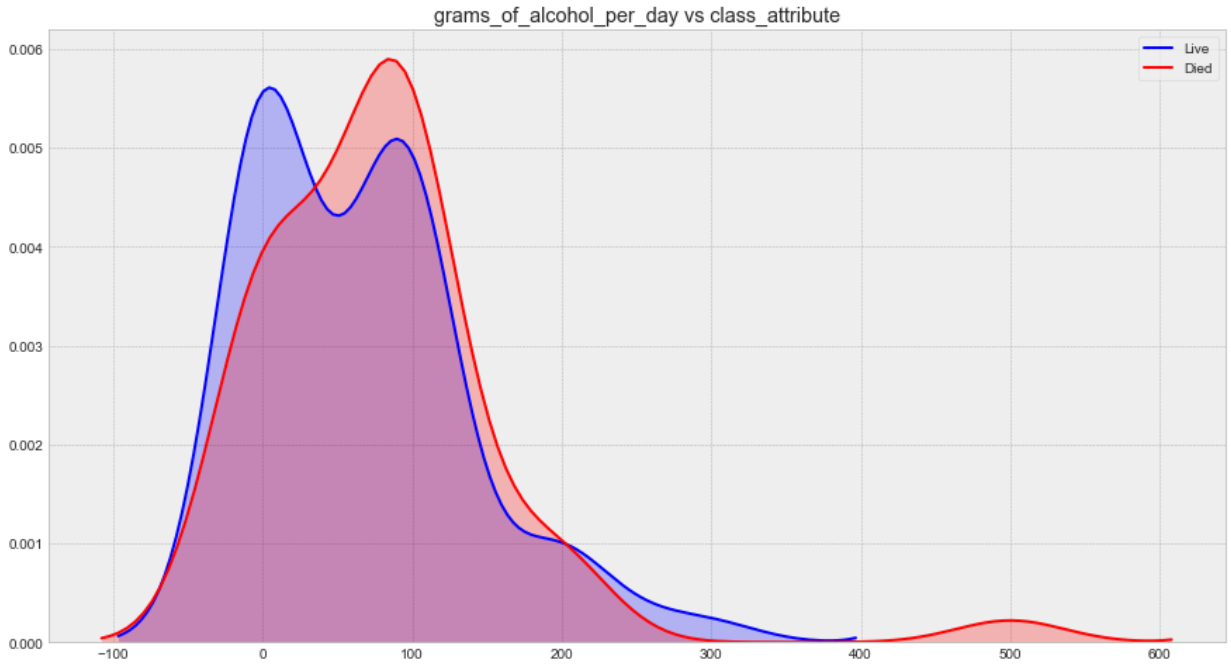

Fig3.9 Alcohol per day with class attributes

PACKS OF CIGARETTE VS CLASS ATTRIBUTE

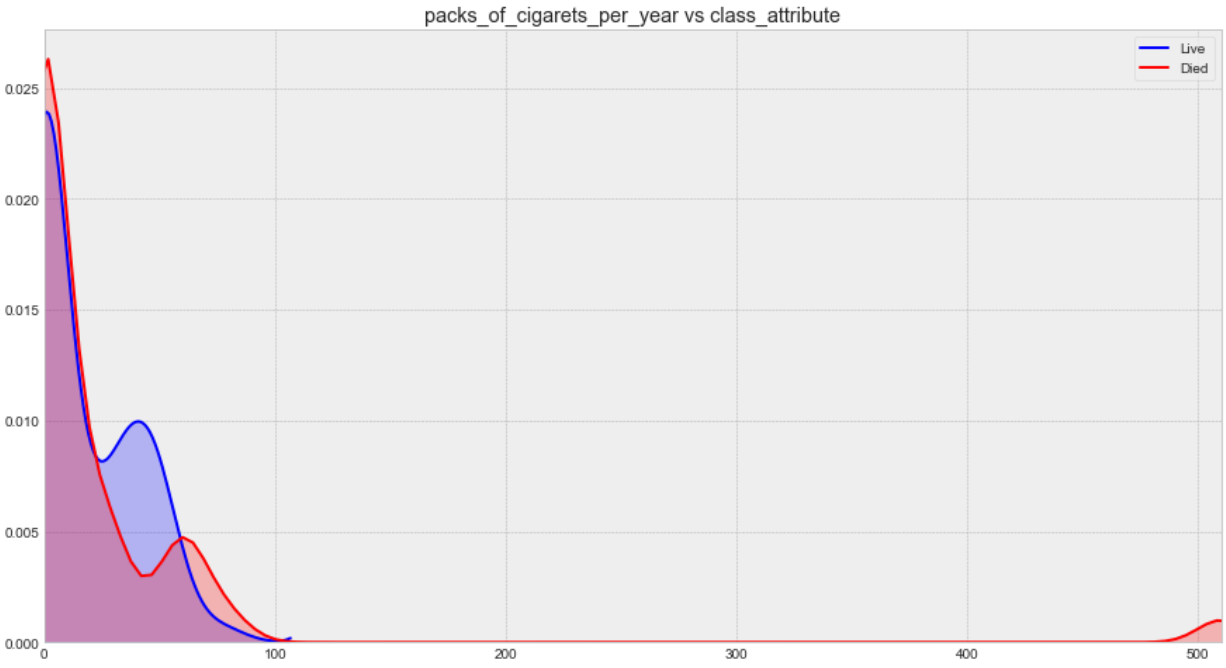

Fig3.10 Cigarettes per day with class attributes

As you can know from the above graph, there is no relationship between packs of cigarette and the death.

AFP VS CLASS ATTRIBUTE

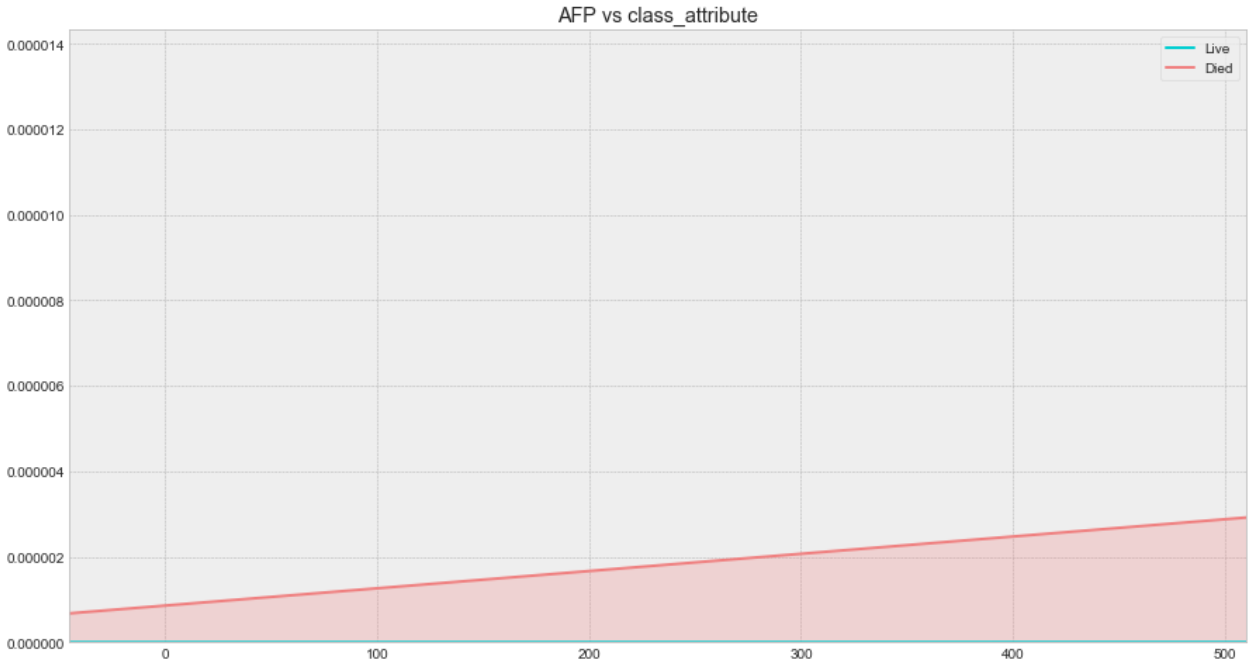

Fig3.11 AFP with class attributes

- AFP normal range is 0 ~ 8.1.
- Most patients who have values over threshold in AFP were deceased in a year.

IRON VS CLASS ATTRIBUTE

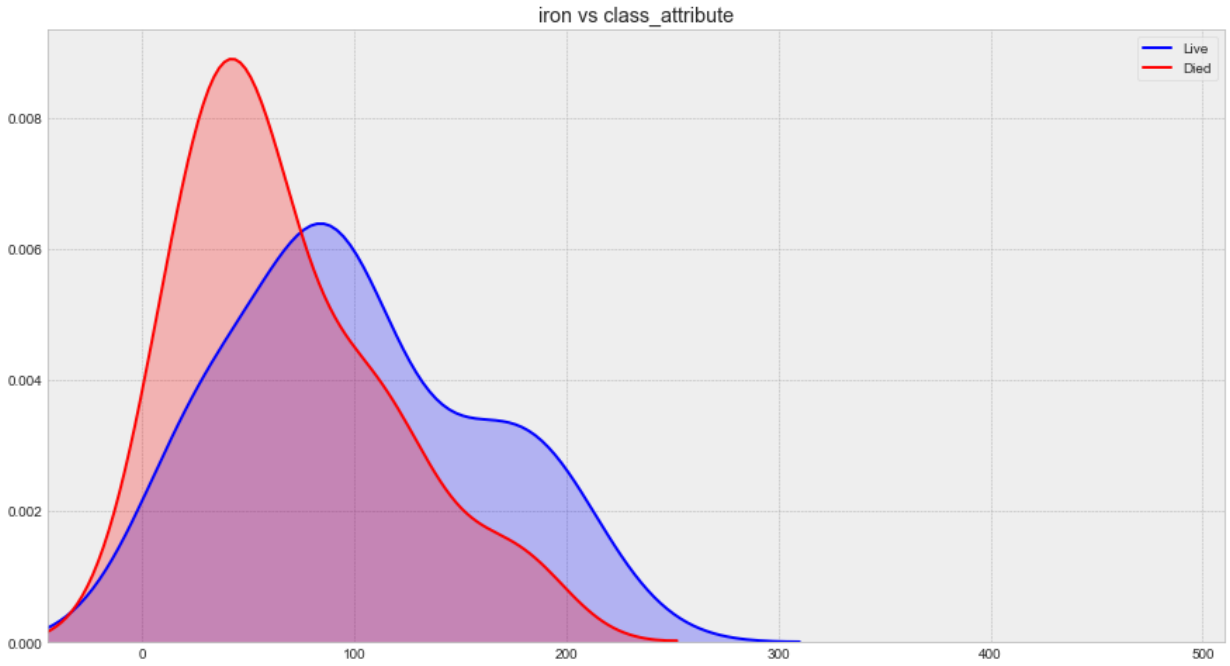

Fig3.12 Iron with class attributes

This graph explains that the patients experienced lack of iron seem to be deceased in 1 year.

COMPARING MEN VS WOMEN

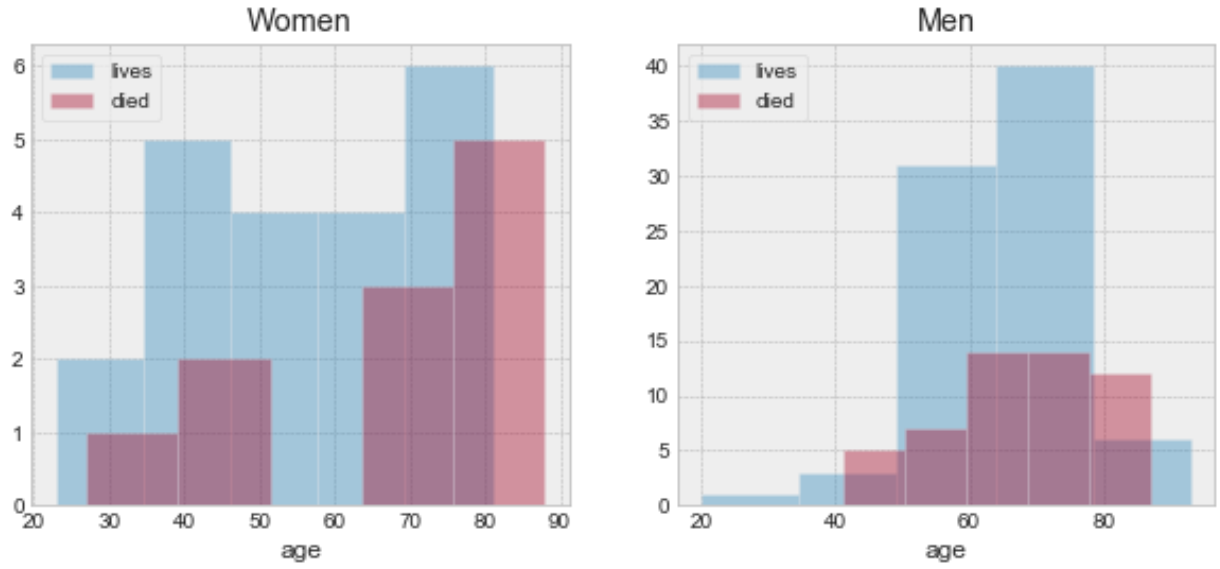

Fig3.13 Men and Women with class attributes

The left bar graph shows that female lives in various ages. However, the death rate in an year was increased after around age 65. Besides, survival rate looks like normal distribution.

The right one shows a slight left skewed bar chart. As we mentioned above, most of patients are between 50 and 75.

CORRELATION (ALL FEATURES) – HEATMAP

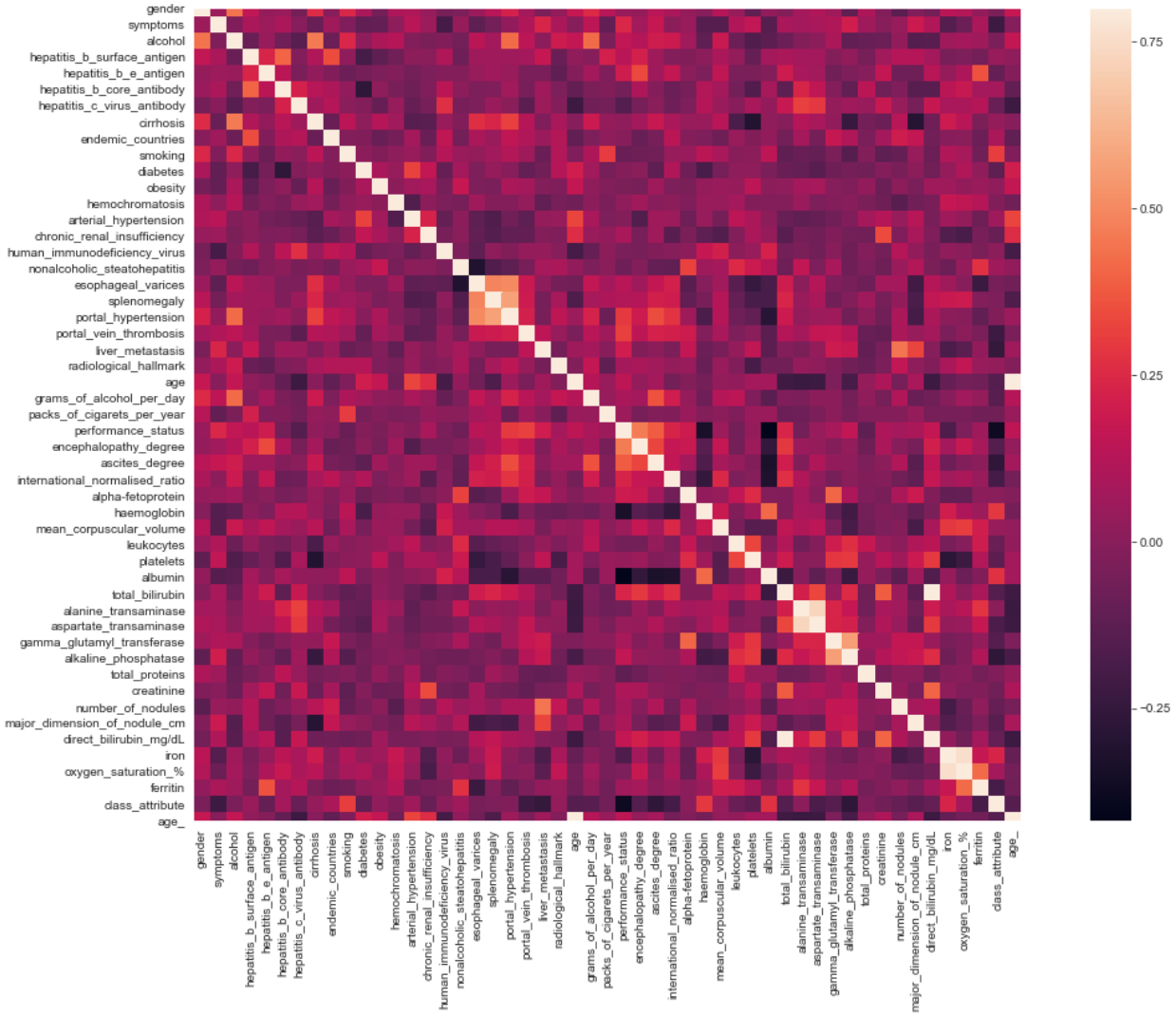

Fig3.14 All feature heatmap

Using the heatmap, it illustrates the correlation between two features at a glance. The Brighter cell shows the strong relationship. The darker means the weak relation between the two elements.

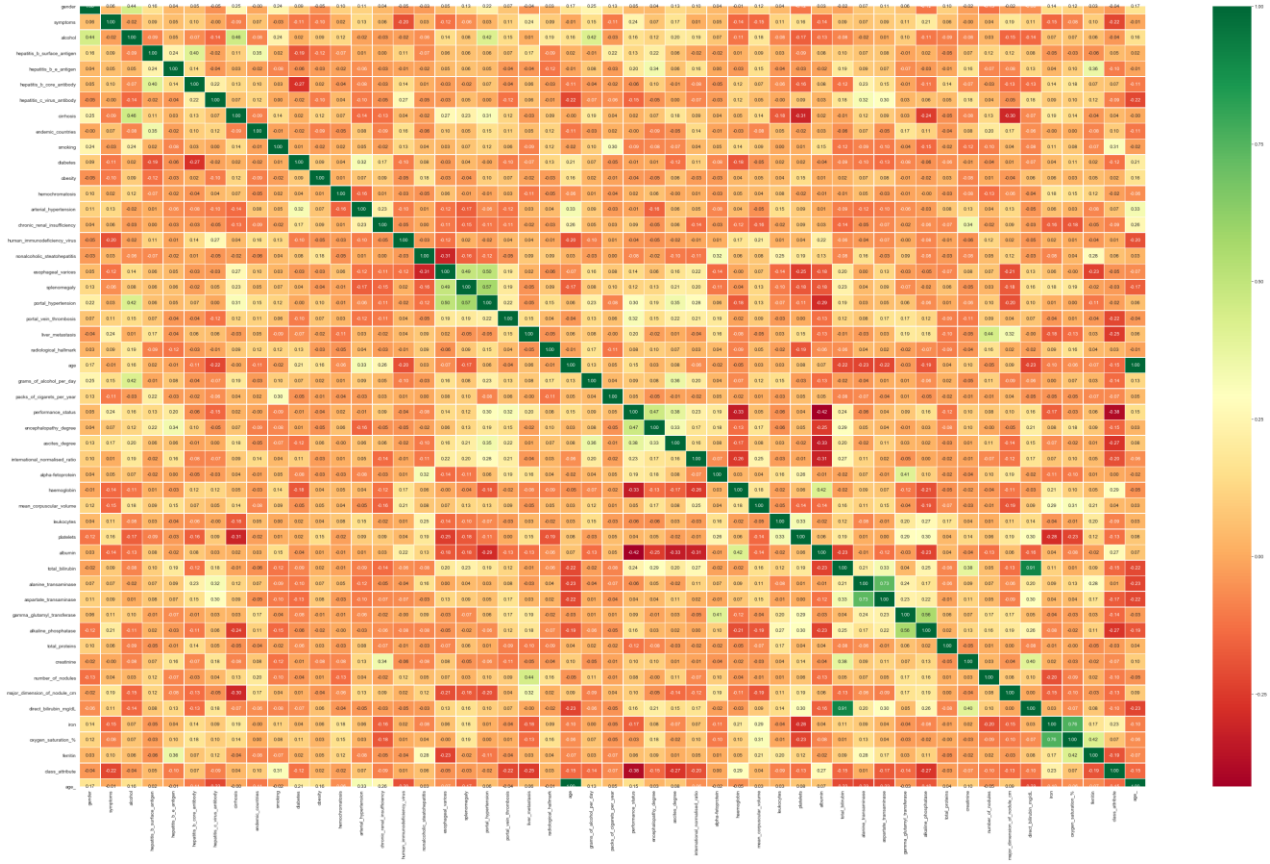

Fig3.15 feature correlation heatmap

In this case, we put the correlation values in the cell with the gradation color. The closer the number is to 1, the higher the correlation.

According to the heatmap, portal Hypertension and Hepatosplenomegaly has strong relationship. Also, (AST) AST(aspartate aminotransferase) and ALT(alanine aminotransferase) have a high correlation.

In fact, these factors are used for diagnosing liver function by doctors. If those values are out of normal range, we should suspect that there might be problems in our liver.

To sum up, using this model, we can classify the class using only blood test result. In the aspect of statistics, we can apply this model to the patients who comes to the hospital to take the blood testing or the ministry of health can provide individual blood testing periodically so that we can diagnose how citizens' current health status and recommend doing further test, such as CT, MRI.

4. OBSERVATION

Liver cancer is one of the leading causes of death today worldwide. Hepatocellular carcinoma (HCC) is one of the major liver cancers. Common factors include alcohol, obesity, diabetics, hepatitis-b, c etc are some of the common factors know to most people. There are some other factors affecting HCC like endemic countries, NASH, Homochromatic etc are some of the least known factors.

COMMON FACTORS

ALCOHOL

Alcohol is one of the major causes for HCC. Alcohol is directly toxic to liver cells. It causes cirrhosis a form of liver damage. The risk of cancer increases with the consumption of alcohol.

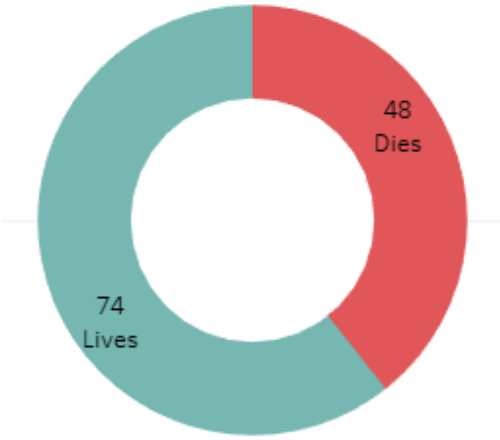

Fig4.1 Shows Alcoholics patients

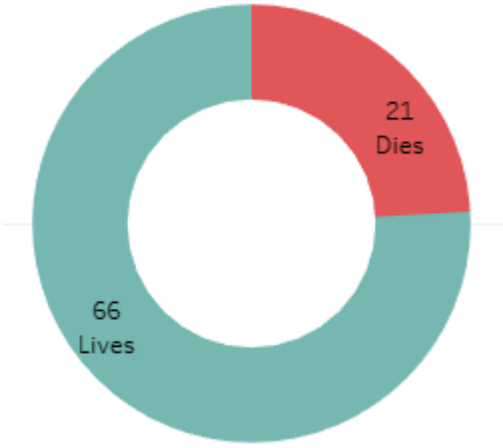

Fig4.2 Shows Smoking patients

SMOKING

Smoking increases your risk of many different types of cancers. Tobacco use leads to many diseases affecting the heart, liver and lungs. The effects increase as how much and number of years a person smoke.

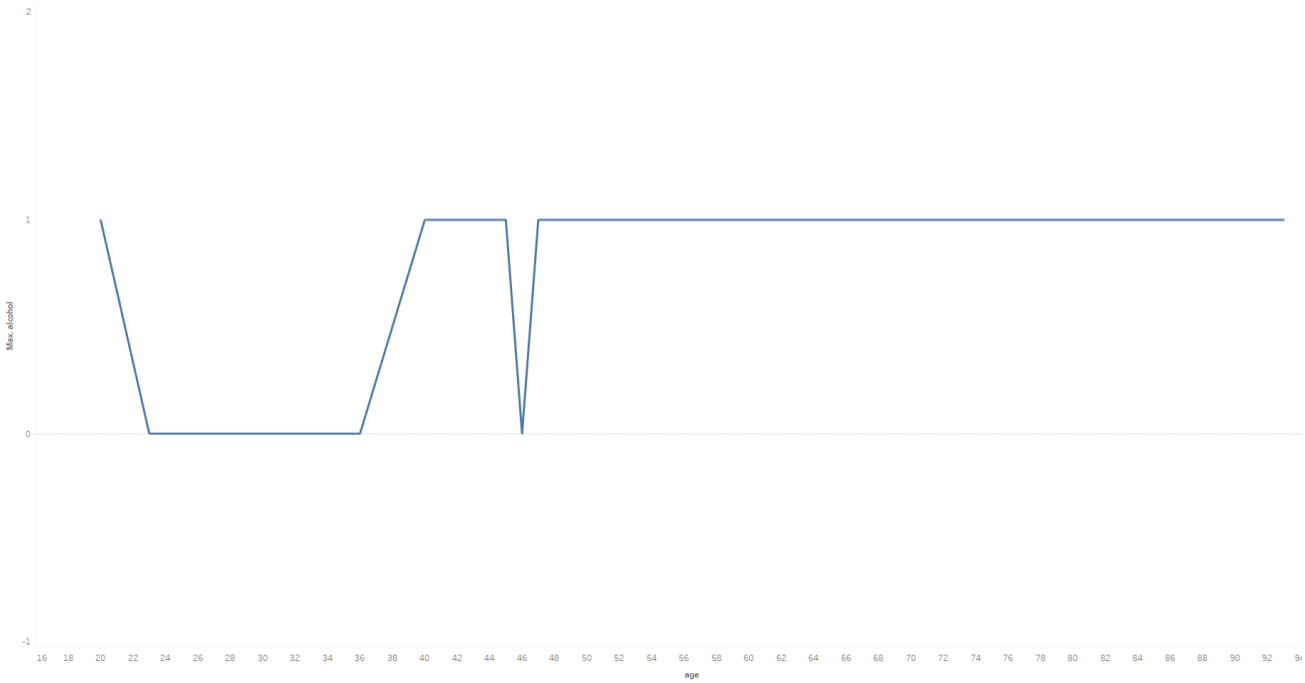

Fig4.3 Shows Smoking distribution with age

DIABETIC

People with type 2 diabetes have a high risk of cancer. The risk is due to high insulin in blood which cause liver damage if it is not controlled properly.

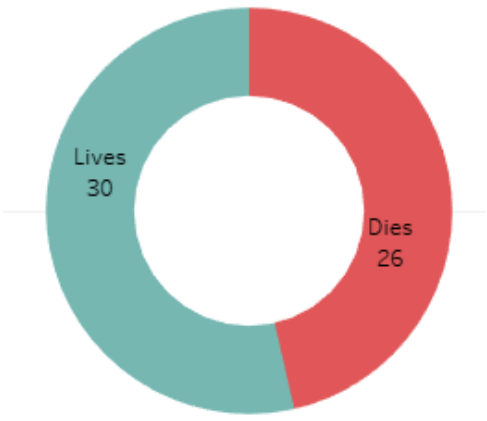

Fig4.4 Shows Diabetic patient

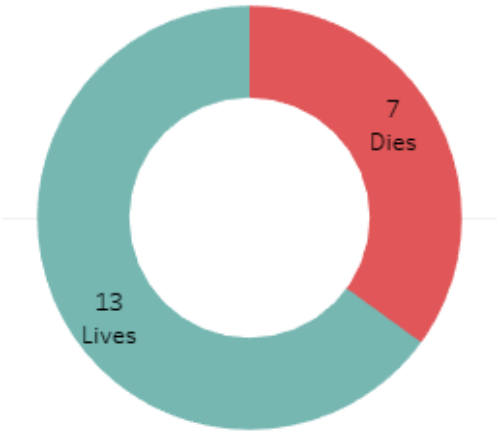

Fig4.6 Shows Obese patients

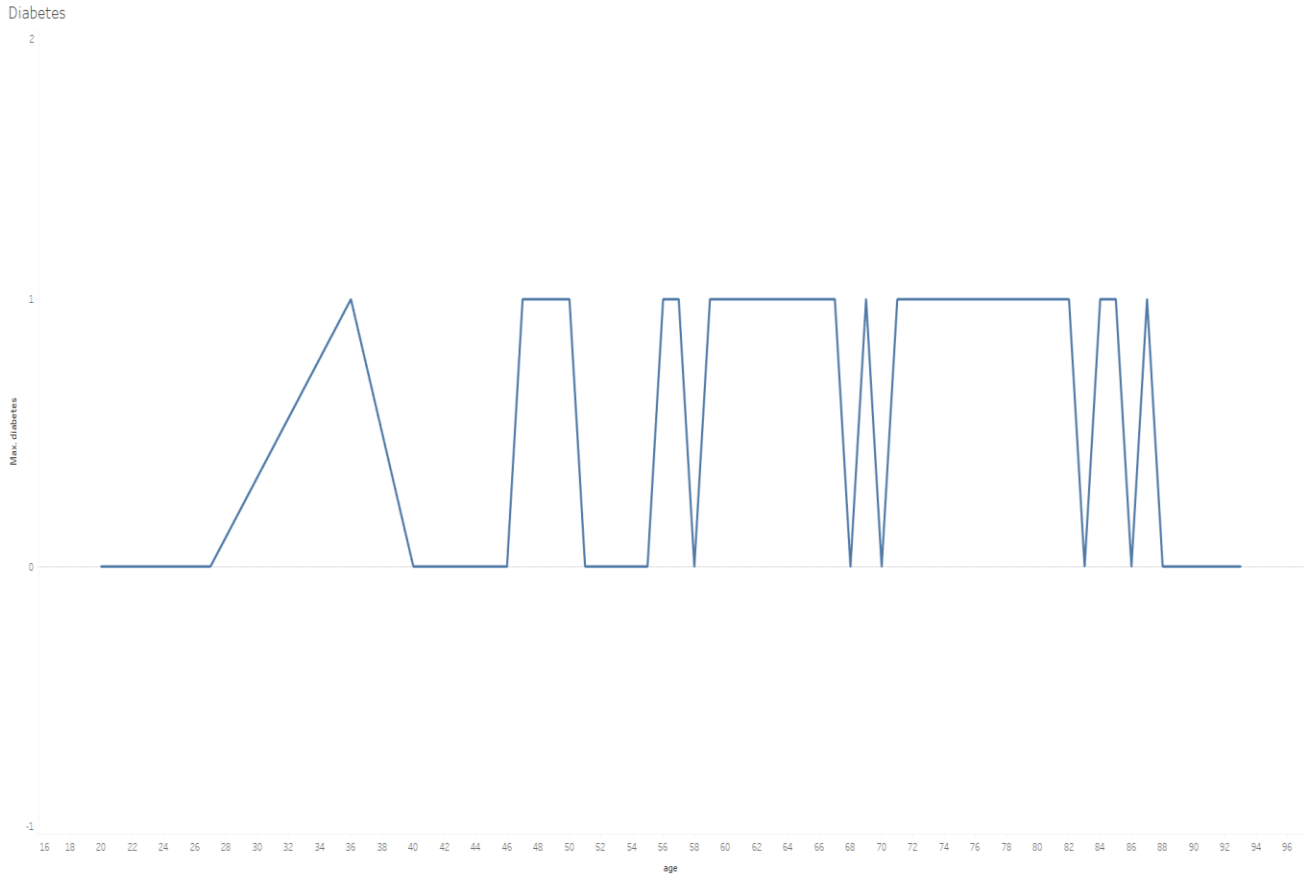

Fig4.5 Shows Diabetic distribution with age

OBESITY

This will lead to an excess fat build up in the liver which is NAFLD. This cause the risk of cirrhosis, type2 diabetes and this will intern leads to cancer.

ALPHA FETOPROTEIN (AFP)

It is a protein made in the liver. It various across different person. In general, it is high in a new born baby and reduce to low when the age of 1. High level of AFP can be a sign of liver cancer or cancer of the ovaries or testicles well as noncancerous liver diseases such as cirrhosis and hepatitis.

AFPCOMPARISON

Status	Normal	Vulnerable	Critical
Dies	14	22	27
Lives	52	34	16

Fig4.6 Shows AFP comparison

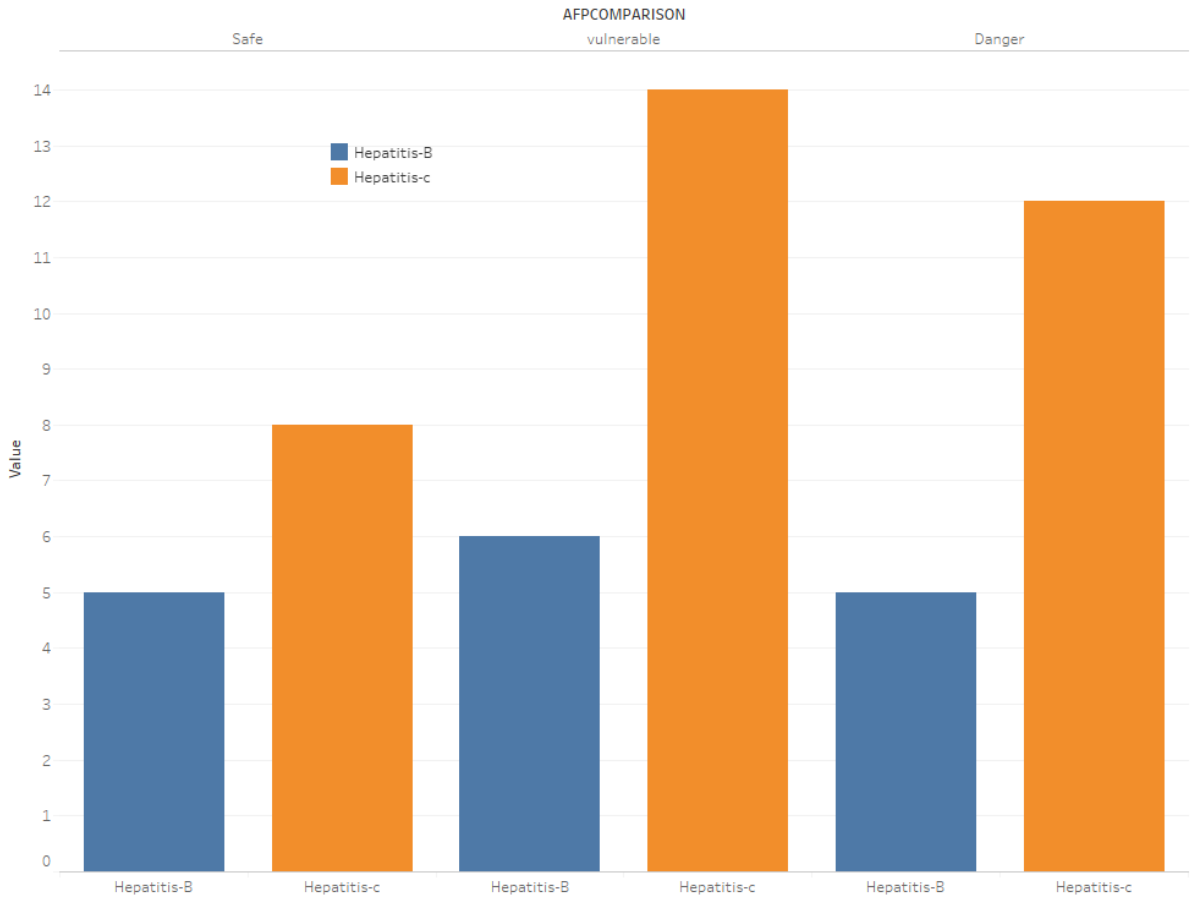

Fig4.7 shows Hepatitis B, C comparison of living and died patient

ENDEMIC

Through our study, we understood that there is a relationship between endemic regions and liver disease. Hemochromatosis is a condition caused by the overconsumption of Iron through foods and this absorption can lead to excessive concentration of that element in our body.

In our dataset, we found that people in some regions are more prone to the liver disease, especially people from China (Hong Kong SAR, Qidong, Shanghai, Tianjin), Gambia, India (Barshi, Karunagapally), Singapore, Republic of Korea (Busan, Incheon, Seoul), Thailand, Uganda. It is mainly because these areas have more Iron content in their soil than the rest of the world. So, their chance of having more iron in their food is high and therefore more prone to liver-related diseases.

Fig4.8 shows the most endemic regions

NON-ALCOHOLIC STEATOHEPATITIS (NASH)

Over time, the fat in the liver can lead to liver inflammation, the end stage of which we call NASH. If it leads to liver cirrhosis, then there is an increased risk of hepatocellular carcinoma. NASH is a diagnosis of exclusion. Which means it can only be determined by process of elimination, excluding other diagnostics. steatohepatitis means a fatty liver with inflammation. It tends to be chronic and requires long-term care.

NASH occurs in those who does not drink alcohol or drink in moderation. Its more seen from fourth decades of life. There are cases in young patients. NASH is associated with liver damage and can cause to cirrhosis, liver cancers.

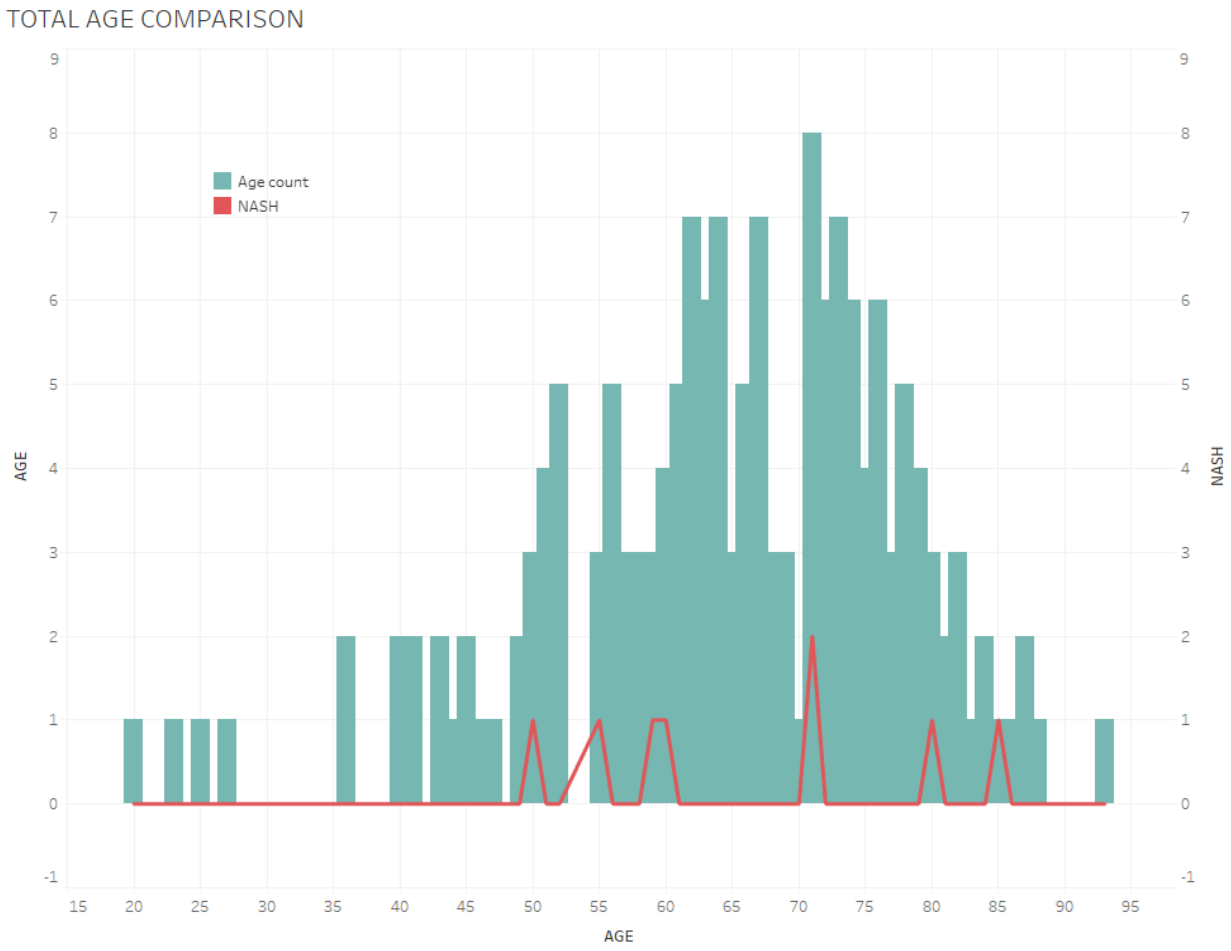

Fig4.9 shows the age comparison of patients with total data

Fig 3.1 shows the age count of patients in the dataset. The patients with Nash are indicated in red line which show they are above fifth decade of life. The bar is the total age count of patients in our dataset.

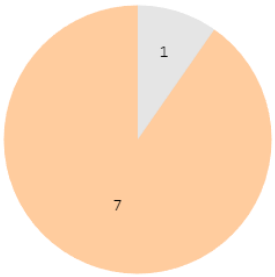

Fig3.2 show the number of cirrhosis patients in the data set. From the total Nash positive patients expect for one everyone else has cirrhosis.

Fig4.10shows the total cirrhosis patient

Symptoms

NASH have no symptoms. It is widely reported that abdomen discomfort/pain, fatigue and an unwell feeling are some of the symptoms.

Treatment

Currently, there is no proven treatment with medication for NASH. Treatment for NASH has not yet been established. Doctors recommended that people avoid drinking alcohol since it is another cause of NAFLD (Non-Alcohol Fatty Liver Disease), improve their glucose if they are diabetic and weight loss if they are obese.

Prevention

- healthy plant-based diet that's rich in fruits, vegetables, whole grains and healthy fats.
- If you are obese, reduce the calorie intake each day and get more exercise.
- If you have a good weight, maintain it and exercise daily.

HEMOCHROMATOSIS

Despite being the most prevalent monoallelic genetic disease in Caucasians, hereditary hemochromatosis is under-diagnosed. There are several reasons for this, including lack of awareness, a long latency period, and nonspecific symptoms.

Normal iron absorption occurs in the proximal small intestine at a rate of 1-2 mg per day. In people with hereditary hemochromatosis, this absorption rate can reach 4–5 mg per day with progressive accumulation to 15–40 grams of iron in the body. Humans have no physiologic pathway to excrete iron. Therefore, iron can accumulate in any body tissue, although depositions are most common in the liver, thyroid, heart, pancreas, gonads, hypothalamus and joints. Hemochromatosis causes or exacerbates arthritis, diabetes, impotence, heart failure, cirrhosis of the liver and liver cancer.

Up to half of people who have hemochromatosis don't get any symptoms. In men, symptoms tend to show up between ages 30 and 50. Women often don't show signs of this condition until they're over 50 or past menopause. That may be because they lose iron when they get their periods and give birth.

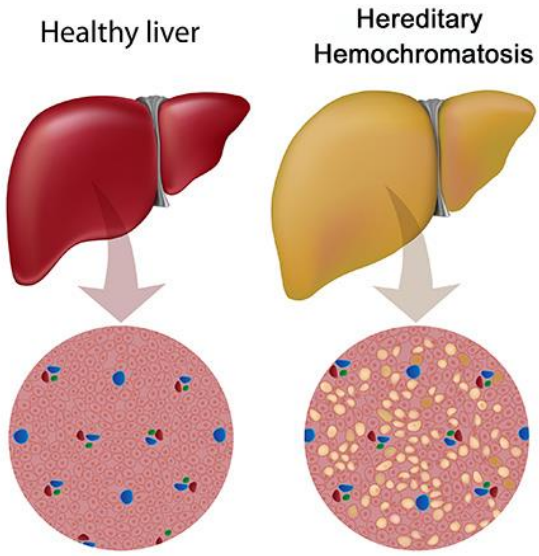

Figure 4.11 Healthy Liver and liver with hemochromatosis

Symptoms

- Joint pain
- Abdominal pain

- Fatigue
- Diabetes
- Loss of sexual drive
- Heart failure
- Liver failure

Treatments

Currently, the available medicines that work to decrease excess iron include:

- Deferoxamine (Desferal)
- Exjade (Deferasirox)
- Ferriprox (Deferiprone)

In addition to therapeutic blood removal, you may further reduce your risk of complications from hemochromatosis if you:

- **Avoid iron supplements and multivitamins containing iron.** These can increase your iron levels even more.
- **Avoid vitamin C supplements.** Vitamin C increases absorption of iron. There's usually no need to restrict vitamin C in your diet, however.
- **Avoid alcohol.** Alcohol greatly increases the risk of liver damage in people with hereditary hemochromatosis. If you have hereditary hemochromatosis and you already have liver disease, avoid alcohol completely.
- **Avoid eating raw fish and shellfish.** People with hereditary hemochromatosis are susceptible to infections, particularly those caused by certain bacteria in raw fish and shellfish.

In the dataset, we observed all the patients with hemochromatosis are men. Out of 7 men with hemochromatosis, 6 showed abnormal values in their ferritin test which is a blood cell protein that contains iron. A ferritin test helps your doctor understand how much iron your body is storing.

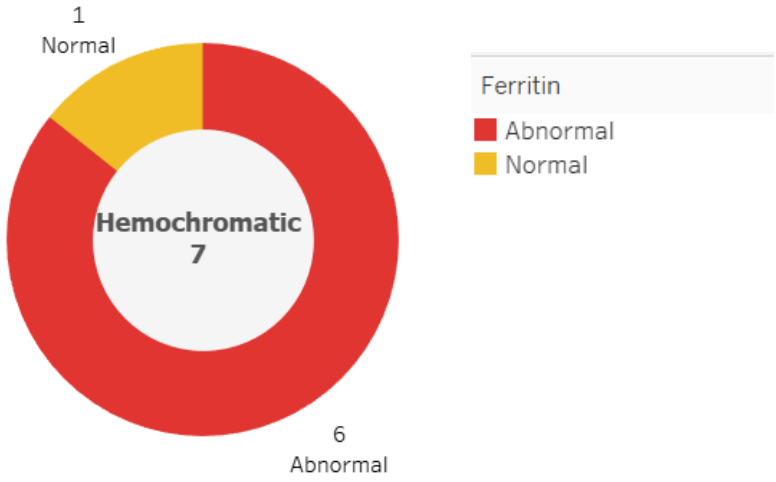

Figure 4.12. Hemochromatosis and Ferritin test result

The survival rate is observed as 3 out of 7 patients did not HCC survived the initial year while 4 of them were alive.

Figure4.13 HCC Survival Rate in hemochromatosis patients

5.CONCLUSION

HCC is one of the most prevalent cancers incident in the world. We analysed the data and discovered how different factors influence the cancer death rate. These are the some of the most influential factors: -

- Non-alcoholic steatohepatitis (NAS)
- Influence of endemic regions
- Homochromatic

The dataset contains strong evidence that these can alter the patient conditions depending upon their levels in the patient. They are not general observations and can be diagnosed upon specific tests.

Chances of patient with these features surviving cancer is less as their levels are not in the range. AFP values are checked to ensure the liver conditions. Hight AFP doesnt mean cancer but some abnormality with liver.

Other factors are seen in the dataset which influence the patient conditions. In general, they are not doing that much damage but are causing some interline conditions which intern affect the functions of liver. Commonly known factors of caner: -

- Alcohol
- Obesity
- Diabetics
- Smoking

Using regression, we were able to find the critical features that affecting patients diagnosed with HCC. We can predict the survival of HCC patients

6.DISCUSSION

We selected the topic **emergency department waiting time and solutions.** The dataset was limited and was not accessible to public. They will provide the dataset only with a formal request. We contacted the authorities to get the complete dataset. After a week we received the response from the authority. Access to data set will take about 2 months for approval. Even then we need to pay for using data. Discussion about this is attached[C].

We decided to drop the topic and progress with our back up project HCC. This dataset is relatively small with 164 observations. There are about 42 features in with most of them are related to blood tests and different health condition. A lot of research is needed to understand and complete this project successfully. We consulted doctors and other professionals to validate the data and provide a meaningful analysis[D]

7.REFERENCES

https://www.kaggle.com/mrsantos/hcc-dataset1

https://www.mayoclinic.org/diseases-conditions/hemochromatosis/symptoms-causes/syc-20351443

https://www.cancerresearch.org/immunotherapy/cancer-types/liver-cancer

https://www.liver.ca/patients-caregivers/liver-diseases/nash/

https://www.mayoclinic.org/diseases-conditions/esophageal-varices/symptoms-causes/syc-20351538

https://www.healthline.com/health/hepatosplenomegaly

https://www.healthline.com/health/portal-vein-thrombosis

https://phassociation.org/patients/aboutph/diseases-and-conditions-associated-with-ph/liver-disease/

https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=alpha_fetoprotein_tumor_marker

https://www.webmd.com/a-to-z-guides/alanine-aminotransferase-test#1

https://www.medicinenet.com/liver_blood_tests/article.htm#what_are_normal_levels_of_ast_and_alt

https://www.healthline.com/health/alp

https://medlineplus.gov/lab-tests/alpha-fetoprotein-afp-tumor-marker-test/

https://www.diabetes.co.uk/diabetes-complications/liver-cancer.html

https://www.mayoclinic.org/diseases-conditions/hepatocellular-carcinoma/cdc-20354552

https://www.cancer.org/cancer/liver-cancer/about/what-is-key-statistics.html

https://www.wcrf.org/dietandcancer/cancer-trends/liver-cancer-statistics

https://www.intechopen.com/books/hepatocellular-carcinoma-advances-in-diagnosis-and-treatment/diagnostic-algorithm-of-hepatocellular-carcinoma-classics-and-innovations-in-radiology-and-pathology

https://aasldpubs.onlinelibrary.wiley.com/doi/full/10.1002/hep.27969#support-information-section

https://www.clearvuehealth.com/info/nash-symptoms/

https://www.mayoclinic.org/diseases-conditions/nonalcoholic-fatty-liver-disease/symptoms-causes/syc-20354567

https://www.liver.ca/patients-caregivers/liver-diseases/nash/

https://www.cancertherapyadvisor.com/home/decision-support-in-medicine/hospital-medicine/nonalcoholic-steatohepatitis/

https://www.webmd.com/digestive-disorders/picture-of-the-liver#1

https://gastroinflorida.com/blog/hepatocellular-carcinoma-hcc/

https://www.medicalnewstoday.com/articles/305075.php#structure

https://en.wikipedia.org/wiki/Logistic_regression

https://www.youtube.com/watch?v=5gea0tneL34

APPENDIX

A. NORMAL RANGE IN AN ORIGINAL TEST RESULT

검사코드	검 사 명	한 글 명	단위	정상치
BL2011	WBC Count, Blood	백혈구	$x10^3$ / μL	3.8~10.58
BL2012	RBC Count, Blood	적혈구	x10^6/ μL	4.23~5.59
			$x100^3$ / μL	4.23~5.59
BL2013	Hemoglobin, Blood	혈색소	g/dL	13.6~17.4
BL2014	Hematocrit, Blood	헤마토크리트	%	40.4~51.3
BL201401	MCV (Mean Corpuscular Volume)	평균 적혈구 용적	fL	85.8~98.1
BL201402	MCH (Mean Corpuscular Hemoglob	평균 적혈구 혈색소량	pg	28.8~33.5
BL201403	MCHC (Mean Corpuscular Hemoglo	평균 적혈구 혈색소 농	g/dL	32.3~34.9
BL2016	Platelet Count, Blood	혈소판	$x10^3$ / μL	141~316
BL201801	Blast		%	0~0
BL201802	Promyelocyte		%	0~0
BL201803	Myelocyte		%	0~0
BL201804	Metamyelocyte		%	0~0
BL201805	Band neutrophil		%	0~5
BL201806	Segmented neutrophil	호중구	%	41.5~73.5
BL201807	Eosinophil	호산구	%	0~9.3
BL201808	Basophil	호염구	%	0~1.6
BL201809	Lymphocyte	림프구	%	19.9~49.2
BL201810	Monocyte	단핵구	%	2.2~8.2
BL201811	Atypical Lymphocyte		%	0~0
BL201812	Immature cell		%	0~0
BL201813	Plasma cell		%	0~0
BL201814	Nucleated RBC		/100WBC	0~0
BL201815	ANC (Absolute Neutrophil Count		$x10^3$ / μL	1.57~8.30
BL201816	ALC (Absolute Lymphocyte Count		$x10^3$ / μL	1.00~4.80
BL201818	Abnormal Lymphoid cell		%	0~0
BL211101	PT(sec)	프로트롬빈 타임	sec	12.6~14.9
BL211102	PT(%)	프로트롬빈 시간	%	82~113

BL211103	PT(INR)	프로트롬빈 시간(국제표	INR	0.90~1.10
BL3111	Protein, Total	총단백	g/dℓ	6.4~8.3
BL3112	Albumin	알부민	g/dℓ	3.5~5.2
BL311201	Globulin	글로블린	g/dℓ	2~3.5
BL311202	A/G ratio	알부민/글로블린 비		1.3~2.2
BL3113	Cholesterol	총 콜레스테롤	mg/dℓ	0~200
BL3114	Bilirubin, Total	총 빌리루빈	mg/dℓ	0~1.2
BL3115	AST	에이에스티	U/ℓ	0~40
BL3116	ALT	에이엘티	U/ℓ	0~41
BL3117	ALP	알카라인 포스파타제	U/ℓ	40~129
BL3118	Glucose, Fasting	포도당	mg/dℓ	74~109
BL3120	Creatinine	크레아티닌	mg/dℓ	0.70~1.20
BL312002	Estimated GFR		mL/min/1.73	60~150
BL3122	Ca	칼슘	mg/dℓ	8.6~10.2
BL3123	P	인	mg/dℓ	2.5~4.5
BL3125	GGT (Gamma-Glutamyl Transferas	감마 지티피	U/ℓ	10~71
BL3711	AFP	알파피토 프로테인	ng/mℓ	0~8.1
BL3732	PIVKA -II Test	PIVKA Ⅱ(정량)	mAU/mL	0~40
BL5111	HBsAg	B형 간염항원	COI	~
BL5115	HBeAg	B형 간염 e항원	COI	~
BL5116	Anti-HBe Antibody	B형 간염 e항체	COI	~

B. MACHINE LEARNING USING PYTHON

STANDARDIZATION

```
to_drop_columns = [
   'age',
   'encephalopathy_degree',
   'ascites_degree',
   'performance_status',
   'number_of_nodules'
]

columns_set = set(columns)

_columns = list(columns_set.difference(to_drop_columns))

X = data2[_columns].as_matrix()
y = data2.class_attribute.values

std_scaler = StandardScaler() #StandardScaler() # RobustScaler
X_new = std_scaler.fit_transform(X_new)
```

- We removed above 5 features which can lead to wrong result.

SPLITTING DATA INTO TRAIN AND TEST DATASET

```
X_train, X_test, y_train, y_test = train_test_split(
    X_new,
    y,
    random_state=42,
    test_size=0.20
)
```

LOGISTIC REGRESSION

```
log_reg = LogisticRegression(
    solver='lbfgs',
    random_state=42,
    C=0.1,
    multi_class='ovr',
    penalty='l2',
)

log_reg.fit(X_train, y_train)
log_reg_predict = log_reg.predict(X_test)
log_reg.score(X_test, y_test)

0.9696969696969697
```

Logistic Regression Accuracy: 96.97%
Logistic Regression AUC: 97.37%
Logistic Regression Classification report:

	precision	recall	f1-score	support
0	0.93	1.00	0.97	14
1	1.00	0.95	0.97	19
accuracy			0.97	33
macro avg	0.97	0.97	0.97	33
weighted avg	0.97	0.97	0.97	33

Accuracy of Logistic Regression is 96.97 %.

After using 5-fold cross validation for the model validation, accuracy of this model records 0.98.

Cross-validated scores: [0.95238095 1. 1. 0.97435897 1.]

	precision	recall	f1-score	support
0	0.97	0.98	0.98	63
1	0.99	0.98	0.99	102
accuracy			0.98	165
macro avg	0.98	0.98	0.98	165
weighted avg	0.98	0.98	0.98	165

LogisticRegression: F1 after 5-fold cross-validation: 98.53% (+/- 0.04%)

C. CIHI DATA REPLY

Client Support Representative, Decision Support Services (DSS)
Canadian Institute for Health Information (CIHI)
4110 Yonge Street, Suite 300
Toronto, Ontario M2P 2B7
T: 416-549-5237
F: 416-481-2950
smagee@cihi.ca

Better data. Better decisions. Healthier Canadians.

Twitter | Facebook | LinkedIn | Instagram | YouTube

From: BV5@myscc.ca <BV5@myscc.ca>
Sent: Friday, September 20, 2019 10:16 PM
To: snap <SNAP@cihi.ca>
Subject: Form Submission - Data Inquiry Form 201909 - 1261

Submitted on: Friday, September 20, 2019 - 22:16
Submitted by user: **Bhavya Vinod**
Name: **Bhavya Vinod**
Position: **Student**
Organization: **St.Clair college, Windsor, Ontario**
Email: BV5@myscc.ca
Telephone: **226-759-3397**
Extension:
Type of Data: **To be determined**
Topic: **Emergency department**
Detailed description of data requested:
1, Injury and Trauma Emergency Department and Hospitalization Statistics, 2016–2017

2, NACRS Emergency Department Visits
and Length of Stay by Province/Territory, 2016–2017

Detailed information on the above two datasets required.
Research question or purpose of request:
As a part of Health informatics course in my bachelor's degree, I am doing a project on "emergency department waiting time and solutions".

For further information, you can contact my instructor :
Mahmoud Artima <MARTIMA@stclaircollege.ca>

"*emergency department waiting time and solutions*". If applicable please also include a copy of your protocol which includes the finer details of your research project.

In the meantime, we would like to bring to your attention the following initial **key points** regarding the data:

- CIHI does not collect data from private clinics/practices. Data in DAD and NACRS is based on hospital visits/stays and is a reflection of what is recorded in the patient's hospitalization medical record.
- Typically we release data that is based on a specific cohort which is defined by the requestor, in support of their research project at hand.
  - Once you are able to provide is with greater for what it is you are interested in, we will be able to better advise you on what data is available.
- In NACRS we do not have 100% ED data coverage, rather we have 100% ED data coverage from Alberta, Ontario and the Yukon **only**.
- FY2018-2019 is the most recent year of complete data we have available for release (which includes ED visits taking place from April 1st 2018 to March 31st, 2019.

Please note: Data request submissions that are received within 5-7 business days from the date of this communication, will be assigned to an analyst sometime in late **November 2019**. For all requests that are received after this timeframe, assignment to an analyst will be based on DSS's workload and resources at the time of submission.

Data Request Process:
All requests for custom reports must be initiated by an **Aggregate Level Data Request Form**, which I have attached to this email. The form can be completed, signed and emailed to snap@cihi.ca. Once we have received the form, your request will be placed in the queue for specifications development.

- When an analyst becomes available to work on your request, his/her first task will be to review the data request form and prepare the detailed specifications for the report (inclusion/exclusion criteria, aggregation required, format of the output, etc).
- This process will be assisted greatly by you providing as much detail as possible about what you would like to see in the report/record layout in your request form.
- Once the analyst has a sound understanding of your requirements, he/she will also prepare a cost estimate.
- Both the detailed specifications and the cost estimate will be forwarded to you for your sign off prior to the request being forwarded to the programming queue for processing.
- The turnaround time for data requests is dependent on a number of factors including the complexity of that data request, completeness of the data request form, consultation with other department/program areas, participation of data requestor, etc.
- The turnaround time, once the request is assigned and the specifications are approved, can range from two weeks to three months+ depending on its complexity.
- The cost for a data request is calculated based on the production time which includes the number of hours to review the data request form, consult and develop specifications, manipulate and/or analyse data, seek advice from CIHI support areas (e.g. Classifications), perform data quality assurance, and transmit data. From that point, you will be invoiced for all worked performed including if the data request is cancelled after that poin

Should you have any questions and/ or concerns based on the information provided, please be sure to let me know.
Thank-you again,

Sarah Magee, ***CHIM***

From: snap <SNAP@cihi.ca>
Sent: Friday, September 27, 2019 9:05 AM
To: Bhavya Vinod <BV5@myscc.ca>
Subject: RE: Form Submission - Data Inquiry Form 201909 - 1261

Good day Bhavya,

We would like to thank you for your expressed interest in CIHI data.

We are the Decision Support Services team here at CIHI and our program area produces customized reports and data sets derived from the Discharge Abstract Database (DAD), the National Ambulatory Care Reporting System (NACRS) and the Hospital Morbidity Database (HMDB).

Customized data would be obtainable through the submission of a formal data request. All data requests are handled through our data request process, which is a cost-recovery process, billable at $320.00/hrly.

Please note that we are unclear what are you referring to when you say "Detailed information on the above two datasets required".

Based on the limited information in your email, we ask that if you do decide to move forward with submitting a formal data request, please be sure to include sufficient details in *PART C – Details of the Project (or Research Study)*. It will be vital that we have a solid understanding of your project on

D.DOCTOR REFERENCE

Hepatocellular Carcinoma:

I need to know top 15 - 20 of following features that are important about this disease.
We will be doing an analysis based on these features and come out with patterns that shows and predicts the population/Patient who likely to survive.

1. Gender
2. Alcohol
3. Hepatitis B Surface Antigen
4. Hepatitis B e Antigen
5. Hepatitis B core Antibody
6. Hepatitis C Virus Antibody
7. Cirrhosis
8. Endemic Countries
9. Smoking Diabetes
10. Obesity
11. Hemochromatosis
12. Arterial Hypertension
13. Chronic Renal Insufficiency
14. Human Immunodeficiency Virus
15. Nonalcoholic Steatohepatitis
16. Esophageal Varices
17. Splenomegaly
18. Portal Hypertension
19. Portal Vein Thrombosis
20. Liver Metastasis
21. Radiological Hallmark
22. Age
23. Grams of Alcohol per day
24. Packs of Cigarettes per year
25. Performance Status
26. Encephalopathy Degree
27. Ascites Degree
28. International Normalised Ratio
29. Alpha Fetoprotein (ng/mL)
30. Haemoglobin (g/dL)
31. Mean Corpuscular Volume (fl)
32. Leukocytes(G/L)
33. Platelets (G/L)
34. Albumin (mg/dL)
35. Total Bilirubin(mg/dL)
36. Alanine transaminase (U/L)
37. Aspartate transaminase (U/L)
38. Gamma glutamyl transferase (U/L)
39. Alkaline phosphatase (U/L)
40. Total Proteins (g/dL)
41. Creatinine (mg/dL)
42. Number of Nodules
43. Major dimension of nodule (cm)
44. Direct bilirubin (mg/dL)
45. Iron (mcg/dL)
46. Oxygen Saturation (%)
47. Ferritin (ng/mL)