SBE304 Project Proposal

First Author
mr.adel98@hotmail.com

Second Author
engmohamedyasser8@gmail.com

Third Author
iomar9606@gmail.com

Fourth Author
MuhammedMohsen1111@gmail.com

1. Introduction

Project objective: the aim of this project is to estimate the survival rate characteristics of 165 hepatocellular carcinoma (HCC) patients and identify useful prognostic factors to help in the clinical management of patients with (HCC). this project is important because according to "Hebei Medical University in China" the median survival time is 1, 2, 3, and 5 year survival rates equal to 49.3%, 35.3%, 26.6%, and 19.5% respectively. The results showed that the Barcelona clinic liver cancer (BCLC) stage and Tumor size were independent prognostic factors to HCC patients. The survival rate of HCC patients has increased in the recent years, at the same time the overall survival rate and the prognosis were poor[**?**].

Our work: we will do prepossessing on the data-set using:(feature selection, feature normalization, and data imputation) for 165 patient with HCC before starting, so our result will be accurate, and get a good overall view about the survival rate of HCC patients, we have selected three appropriate methods for our project:(KNN model, NB classifier, and Logistic regression) and we will compare the result to get the most accurate method.

2. Motivation

with more than 600000 deaths a year,HCC is one of the most common cancers in the world.
in Egypt it represents the second most common cancer in men and the 6th most common cancers in women [**?**].
Geographical distribution of HCC varies throughout the world with an incidence rate ranging from 2.1 in Central America to 35.5 in Eastern Asia.
The burden of HCC has been increasing in Egypt with a doubling in the incidence rate in the past 10 years,mostly due to high prevalence of viral hepatitis and its complications, Also There is a geographic correlation between the incidence of HCC and the prevalence of chronic hepatitis B and C, suggesting that these two viral infections are the most important risk factors of HCC worldwide.
And as we know Egypt is the 1^{st} country in the world regarding HCV prevalence , so we want to take a closer look at HCC and the factors that affect it, to provide a useful information that helps the doctors for better understanding of the HCC and what affect it's survival rate,and also we hope this information help the government to put a well thought out plan for raise awareness about HCC.; and there lies our motivation.

3. Evaluation

in our point of view a successful outcome of this project is to be able to predict whether the patient dies or not,i.e to be able to interpret the data to useful information after applying the chosen methods to it.

4. Resources

What resources are you going to use (datasets, computer hardware, computational tools, etc.)?

5. Preprocessing

- Feature selection:
  we will remove any redundant feature (if there is any) by calculating the correlation matrix and removing feature with correlation -ideally- of 0.75 or higher.
  this will be done using Caret R package.

- Feature normalization:
  we will rescale the features so they will have a distribution with $\sigma = 1$ and $\mu = 0$
  this is a general requirement for many machine learning algorithms .
  this will be done using Z-Score Standardization technique using the built-in function in R (scale()).

- Data imputation:
  we will use PMM imputation,to over come the prob-

lem of the missing values in our data.

6. EDA

- The Heat Map procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient color scale is used to represent values of the quantitative variable and there is an example for it.

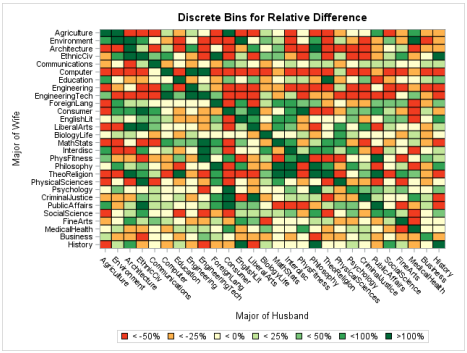

Figure 1. Heat Map

- The Violin Plot Statlet displays data for a single quantitative sample using a combination of a box-and-whisker plot and a nonparametric density estimator. It is very useful for visualizing the shape of the probability density function for the population from which the data came. A separate procedure is available for creating violin plots for multiple samples and there is an example for it.

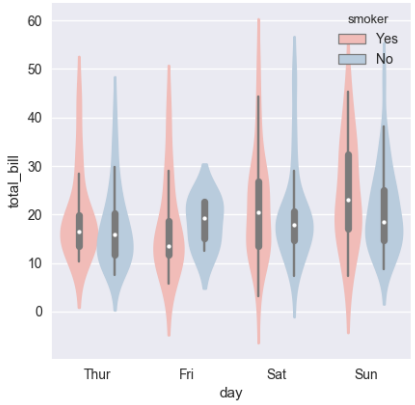

Figure 2. Violin Plot

7. Contributions & Timetable

table of tasks to be done and its date.

TASK	Date
Apply pre-processing and EDA to the data	31\10 to 2 \11
Learn and apply NB classifier to the data	3\11 to 6\11
Learn and apply KNN model to the data	7\11 to 10\11
Learn and apply logistic regression to the data	11\11 to 14\11
revision and editing	22\11
submitting the prototype	25\11

8. Chosen methods

- NB classifier.
- KNN model.
- Logistic regression.

9. Websites

- Adel Moustafa
- Omar Ibrahim
- Mohammed Yasser
- Muhammed Mohsen

References

- Gastroenterology & Hepatology-HCC Burden in Egypt.