ML class project

HCC data

**Introduction and statement of project goals**

Liver cancer is the sixth most frequently diagnosed cancer and, especially, Hepatocellular Carcinoma (HCC) represents more than 90% of primary liver malignancy. Clinicians assess a patient's treatment on the basis of evidence, which may not always apply to a specific case, given the biological variability among individuals. Over the years, and for the particular case of HCC, studies have developed strategies for assisting clinicians in decision making, using data mining methods to extract knowledge from the clinical data. However, these studies have shortcomings: some did not focus entirely on HCC patients, others used strict application boundaries, and none considered the heterogeneity between patients nor the presence of missing data. Using the HCC dataset, we undertook the data mining approach to evaluate the patient level factors to identify those who are at risk of one year mortality.

**Descriptive summary statistics for the dataset**

As for the data source, we usd Hepatocellular Carcinoma dataset (HCC dataset) collected at a University Hospital in Portugal, which contains real clinical data of 165 patients diagnosed with HCC. The HCC data contains several demographic, risk factors, laboratory and overall survival features of 165 real patients. The dataset contains 49 features selected according to the EASL-EORTC (European Association for the Study of the Liver - European Organization for Research and Treatment of Cancer) Clinical Practice Guidelines, which are the current state-of-the-art on the management of HCC.

This is an heterogeneous dataset, with 23 quantitative variables, and 26 qualitative variables. Overall, missing data represents 10.22% of the whole dataset and only eight patients have complete information in all fields (4.85%). The target variables is the survival at 1 year, and was encoded as a binary variable: 0 (dies) and 1 (lives). A certain degree of class-imbalance is also present (63 cases labeled as dead and 102 as alive).

Table 1 shows the demographic characteristics stratified by status.

**Table 1. Demographic data by status**

0 1 p- test

n 63 102

gender (mean (SD)) 0.83 (0.38) 0.79 (0.41) 0.624

symptom (mean (SD)) 0.81 (0.40) 0.52 (0.50) <0.001

alc (mean (SD)) 0.76 (0.43) 0.73 (0.45) 0.607

hepBsurfAnti (mean (SD)) 0.09 (0.30) 0.12 (0.32) 0.689

hepBeAnti (mean (SD)) 0.02 (0.15) 0.00 (0.00) 0.188

hepBcorAnti (mean (SD)) 0.24 (0.43) 0.29 (0.45) 0.562

hepCvirAnti (mean (SD)) 0.28 (0.45) 0.18 (0.39) 0.152

cirr (mean (SD)) 0.89 (0.32) 0.91 (0.29) 0.632

endemicCountries (mean (SD)) 0.04 (0.20) 0.10 (0.31) 0.223

smoke (mean (SD)) 0.46 (0.50) 0.54 (0.50) 0.382

diabetes (mean (SD)) 0.41 (0.50) 0.30 (0.46) 0.154

obese (mean (SD)) 0.12 (0.33) 0.14 (0.34) 0.764

hemochro (mean (SD)) 0.06 (0.25) 0.04 (0.20) 0.577

artHyper (mean (SD)) 0.32 (0.47) 0.39 (0.49) 0.389

chronRenal (mean (SD)) 0.16 (0.37) 0.10 (0.30) 0.269

hiv (mean (SD)) 0.02 (0.13) 0.02 (0.14) 0.911

Nasteato (mean (SD)) 0.04 (0.19) 0.07 (0.25) 0.471

esophVarices (mean (SD)) 0.57 (0.50) 0.63 (0.49) 0.571

spleno (mean (SD)) 0.58 (0.50) 0.55 (0.50) 0.717

portalHyper (mean (SD)) 0.74 (0.44) 0.70 (0.46) 0.605

portalVeinThromb (mean (SD)) 0.33 (0.48) 0.15 (0.36) 0.006

LiverMeta (mean (SD)) 0.35 (0.48) 0.14 (0.35) 0.001

RadioHallmark (mean (SD)) 0.67 (0.48) 0.69 (0.46) 0.757

Age (mean (SD)) 67.16 (13.51) 63.17 (13.04) 0.061

aclPerDay (mean (SD)) 78.62 (83.55) 65.33 (70.47) 0.353

cigPerYr (mean (SD)) 26.65 (81.73) 17.03 (21.31) 0.346

performance (mean (SD)) 1.59 (1.29) 0.67 (0.96) <0.001

encefalopathy (mean (SD)) 1.24 (0.56) 1.11 (0.31) 0.060

ascites (mean (SD)) 1.67 (0.78) 1.30 (0.58) 0.001

international (mean (SD)) 1.54 (0.63) 1.34 (0.32) 0.010

alphaFreto (mean (SD)) 18927.37 (61891.01) 19530.42 (183766.05) 0.980

hemaglobin (mean (SD)) 12.10 (2.13) 13.38 (2.01) <0.001

meanCorpusVol (mean (SD)) 94.68 (9.94) 95.40 (7.30) 0.595

leukocyte (mean (SD)) 1845.48 (3265.03) 1237.54 (2648.36) 0.196

platelet (mean (SD)) 133408.14 (122392.77) 100350.82 (94574.61) 0.055

albumin (mean (SD)) 3.20 (0.58) 3.60 (0.70) <0.001

bilirubin (mean (SD)) 4.61 (8.03) 2.10 (2.44) 0.004

alanineTransaminase (mean (SD)) 67.53 (49.49) 66.82 (62.29) 0.939

aspartateTransaminase (mean (SD)) 117.22 (96.92) 83.12 (78.57) 0.015

gammaGlutamyl (mean (SD)) 316.10 (262.67) 237.44 (252.81) 0.059

alkalinePhosphatase (mean (SD)) 273.89 (173.92) 172.96 (152.31) <0.001

totalProteins (mean (SD)) 9.38 (14.47) 8.68 (9.59) 0.719

creatine (mean (SD)) 1.26 (1.04) 1.04 (0.90) 0.174

numNodules (mean (SD)) 2.94 (1.82) 2.61 (1.78) 0.260

majorDimNode (mean (SD)) 8.13 (5.73) 6.09 (4.54) 0.020

directBilirubin (mean (SD)) 3.30 (6.24) 1.03 (1.43) 0.003

iron (mean (SD)) 67.88 (47.20) 100.30 (58.36) 0.006

oxygen (mean (SD)) 35.68 (31.95) 38.17 (26.54) 0.695

ferritin (mean (SD)) 593.95 (489.75) 301.26 (380.90) 0.003

status (mean (SD)) 0.00 (0.00) 1.00 (0.00) <0.001

Reference.

A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients, Journal of biomedical informatics, 58, 49-59, 2015.