HCC Survival

Wisam Barkho

Contents

Introduction

Survival data was collected on patients of liver cancer (Hepatocarcinoma, or HCC) from a University Hospital in Portugal. The response variable is survival at 1 year of initial diagnosis and is classified as lives = 1 and dies = 0. The dataset contains several demographic, risk factors, and laboratory data of 165 patients that have been diagnosed with HCC. The dataset is heterogeneous with 23 quantitative predictor variables and 26 qualitative predictor variables. Missing values account for 10.22% of the whole dataset with only 8 patients having complete data in all fields.

The problem to answer here is what demographic or clinical data contribute to a patient's survival of HCC beyond 1 year. To solve the problem, exploratory analysis consisting of finding correlated variables, imputation of missing values, and characterizing the distribution via histograms and boxplots. The entire dataset is then analyzed using several models, including logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Gaussian finite mixture models using the `mclust` package, random forest, and support vector machines (SVM). Each model was run using the validation set approach (VSA) which splits the data into 50% training set and 50% test set, leave-one-out cross validation (LOOCV), and 5-fold cross validation. From these results, the best performing models (determined by less than 30% test error rate) are run again on a subset of predictors which are chosen using forward and backward stepwise selection. These results are also reported and the best model is chosen.

Notes on the analysis models: Since this is survival data, special consideration is required for analysis; namely, that survival data is generally not normally distributed. By breaking the normality assumption, this dataset is not ideal for LDA and QDA; however these models are still run for comparison. Instead, I anticipate this dataset is ideal for either logistic regression, SVM, or non-parametric models, such as kNN. Which of these two models will perform better depends on the shape of the decision boundary. If the decision boundary is linear, then logistic regression can be used. If it is not, kNN would be the model of choice. However, since there are only 165 rows, training data is limited which is not optimal for a kNN model. In that case and if logistic regression cannot be used, then more data collection is required.

The HCC dataset can be found here.

Programming Languages/Software: R, RStudio

Skills Used: Machine Learning, Predictive Modeling, Imputation of Missing Values

Exploratory Analysis

Imputation of Missing Values

The below plot illustrates how many attributes contain `NA` values and what percentage of `NA` make up those attributes. Three attributes in particular contain greater than 40% missing values. This percentage is relatively low compared to other datasets, and therefore none of the attributes are excluded based on missing values alone.

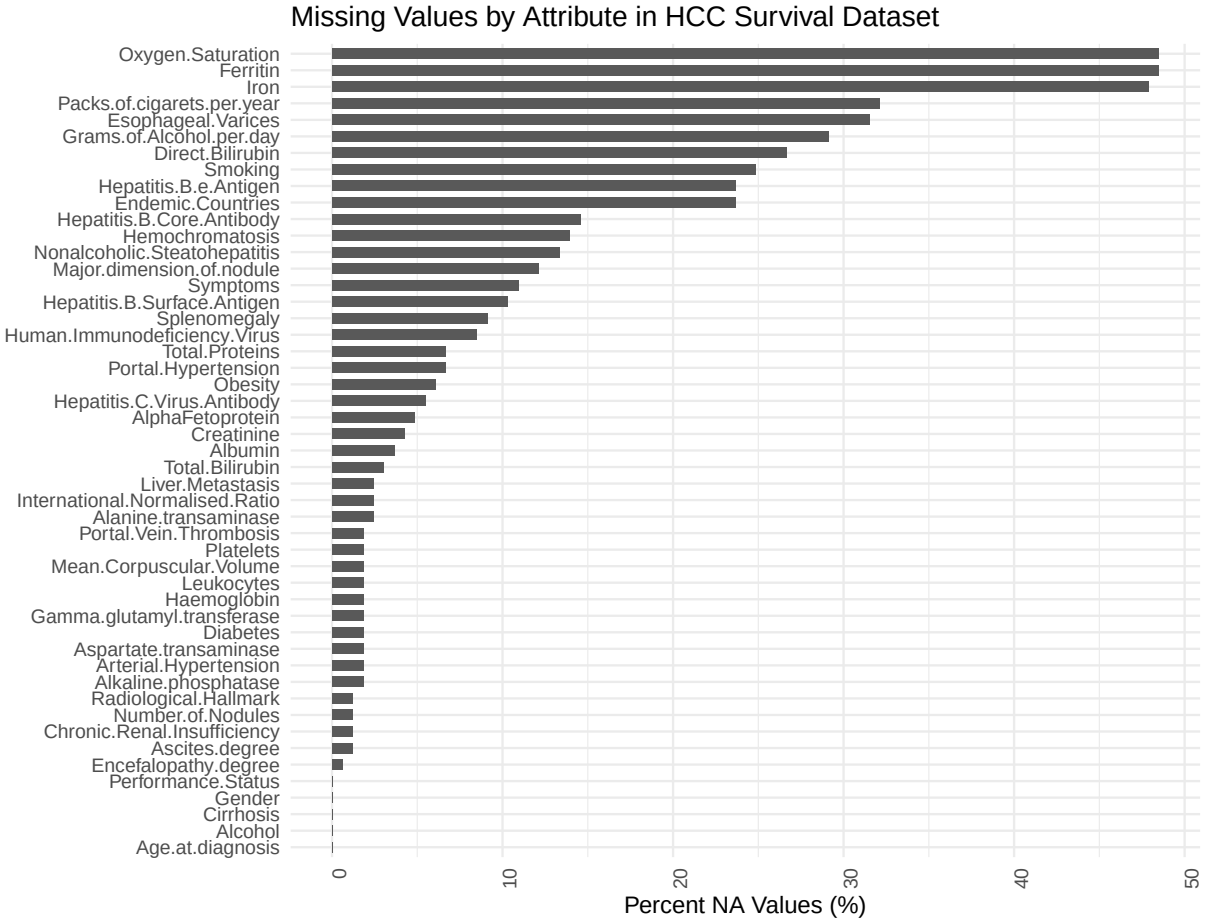

Figure 1: Percent Missing Values in HCC Survival Dataset

I also verify there are no missing values in the response variable, as these will be meaningless.

```
## [1] "NA values in response variable (Survival): "
```

```
## NULL
```

Imputation of missing values is done using the `mice` package. `Nominal` attributes are imputated with the `logreg` method, `ordinal` attributes are imputated with the `polyreg` method, and `continuous` variables are imputated with the `norm` method.

Correlation Table

Correlated attributes are reported in the table below using a custom function which reports the highest correlated values (Pearson Correlation Coefficient of greater than absolute value of 0.7). `Direct.Bilirubin`, `Oxygen.Saturation`, `Aspartate.transaminase` and `Grams.of.Alcohol.per.day` are excluded from our analysis as these have more missing values than their counterparts. Surprisingly, the Pearson Correlation Coefficient for `Smoking` and `Packs.of.cigarets.per.year` is only 0.436. Nevertheless, `Packs.of.cigarets.per.year` is exclude as well since it makes sense this attribute is related to `Smoking`.

Table 1: Correlated Variables for HCC Survival Dataset

	row	column	cor	p
1027	Total.Bilirubin	Direct.Bilirubin	0.978	0
1128	Iron	Oxygen.Saturation	0.783	0
741	Alanine.transaminase	Aspartate.transaminase	0.728	0
279	Alcohol	Grams.of.Alcohol.per.day	0.713	0

Histograms and BoxPlots

The following histograms and boxplots illustrate the distribution of each continuous and categorical predictor variable. Interestingly, at first glance of the boxplots for the variable `Number of Nodules`, survival does not seem to be affected by the number of nodules, which is counterintuitive. However, there might be differences in survival based on the variables `Leukocytes`, `Albumin`, `Gamma Glutamyl Transferase`, and `Alkaline Phosphatase`.

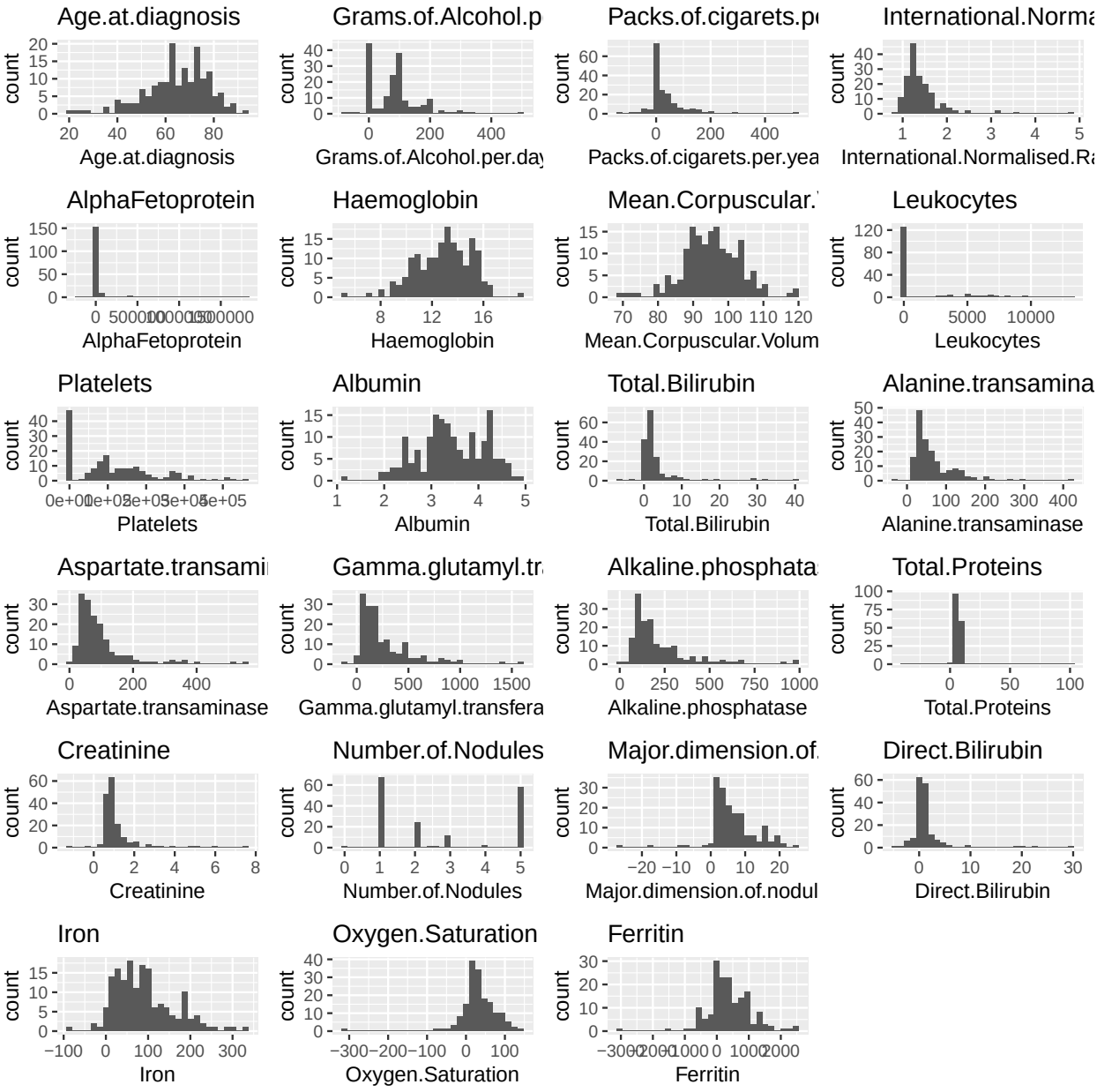

Figure 2: Histograms of Continuous Variables in HCC Survival Dataset

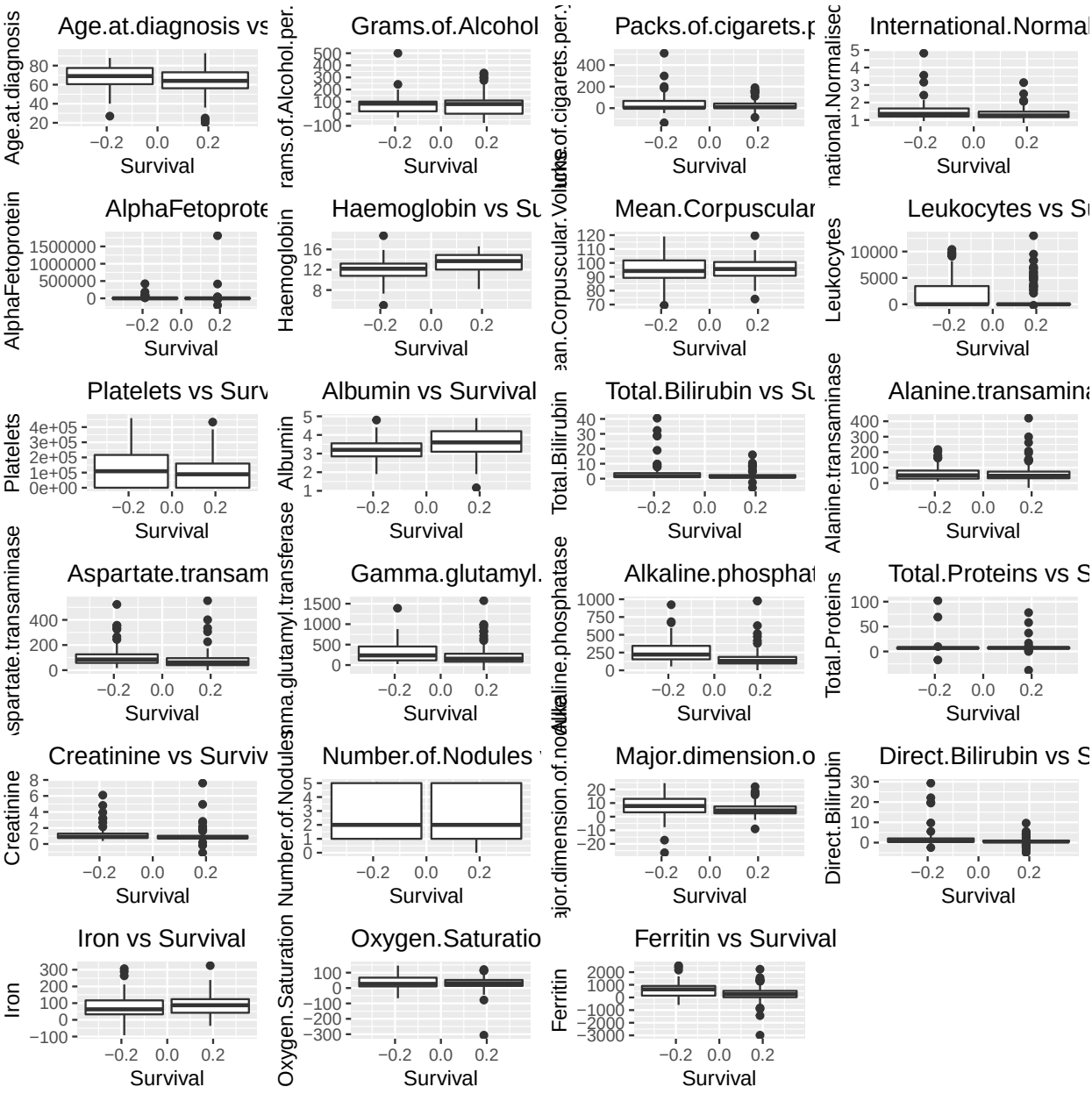

Figure 3: Boxplots of Continuous Variables in HCC Survival Dataset

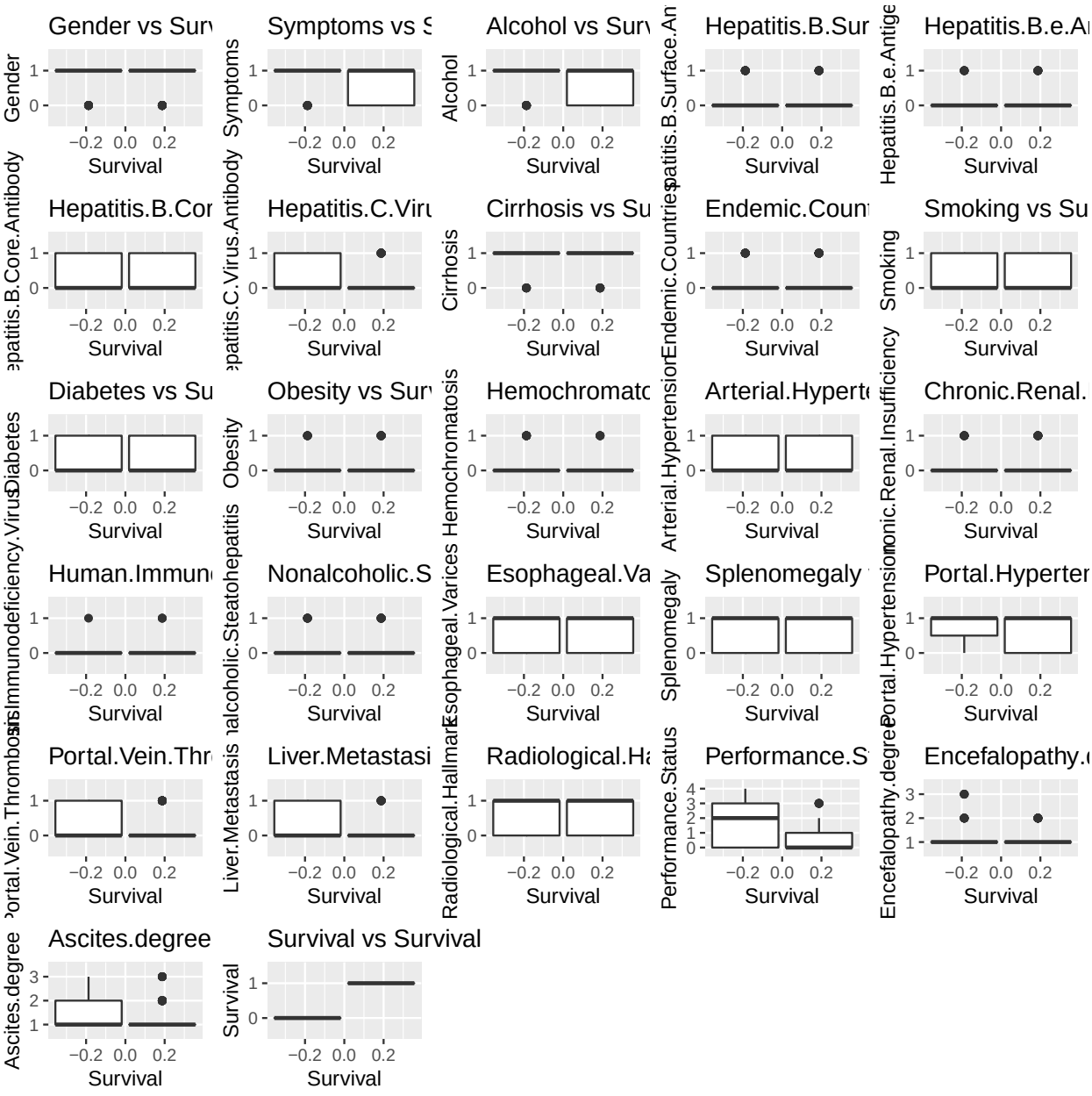

Figure 4: Boxplots of Categorical Variables in HCC Survival Dataset

Analysis using Entire Dataset

Preliminary Findings: Results for all models are reported in table 2. SVM performed the best with test error rates less than 25%. Logistic regression, LDA, and MclustDA with modelType=EDDA also performed well with test error rates between 25% and 30%. However, since survival data usually breaks the normality assumption, LDA will no longer be considered.

Table 2: Model Comparison of Test Error Rates (as percent)

Method	VSA	LOOCV	Five.fold.CV
Logistic Regression	31.3	27.3	27.3
kNN	32.5	36.4	37.0
LDA	28.9	27.9	25.5
QDA	39.8	40.0	41.2
MclustDA	49.4	49.7	40.6
MclustDA, Model Type = EDDA	43.0	27.9	29.1
Random Forest	34.9	41.3	40.9
SVM Linear	30.0	26.7	23.0
SVM Radial	40.0	38.2	38.2
SVM Polynomial	30.0	36.4	23.6

Stepwise Selection

Since this dataset has many features, prediction accuracy might be improved by selecting for the most relevant features. A subset of predictors is chosen using forward and backward stepwise selection, and then the best performing models (test error rates below 30%) are run again.

Forward Stepwise Selection

Forward stepwise selection reduces the original 44 predictors to only 23. The new formula to becomes:

$$
\begin{aligned}
Survival = {} & Alcohol+ \\
& Hepatitis.B.Surface.Antigen+ \\
& Hepatitis.C.Virus.Antibody+ \\
& Smoking+ \\
& Diabetes+ \\
& Hemochromatosis+ \\
& Arterial.Hypertension+ \\
& Nonalcoholic.Steatohepatitis+ \\
& Splenomegaly+ \\
& Portal.Hypertension+ \\
& Portal.Vein.Thrombosis+ \\
& Age.at.diagnosis+ \\
& Performance.Status+ \\
& Encefalopathy.degree+ \\
& Ascites.degree+ \\
& AlphaFetoprotein+ \\
& Haemoglobin+ \\
& Total.Bilirubin+ \\
& Alanine.transaminase+ \\
& Alkaline.phosphatase+ \\
& Major.dimension.of.nodule+ \\
& Iron+ \\
& Ferritin
\end{aligned}
$$

Backward Stepwise Selection

Backward stepwise selection reduces the original 44 predictors to 22. The new formula to becomes:

$$
\begin{aligned}
Survival = {} & Alcohol+ \\
& Hepatitis.B.Surface.Antigen+ \\
& Hepatitis.C.Virus.Antibody+ \\
& Smoking+ \\
& Diabetes+ \\
& Hemochromatosis+ \\
& Arterial.Hypertension+ \\
& Nonalcoholic.Steatohepatitis+ \\
& Splenomegaly+ \\
& Portal.Hypertension+ \\
& Portal.Vein.Thrombosis+ \\
& Age.at.diagnosis+ \\
& Performance.Status+ \\
& Encefalopathy.degree+ \\
& Ascites.degree+ \\
& AlphaFetoprotein+ \\
& Haemoglobin+ \\
& Total.Bilirubin+ \\
& Alanine.transaminase+ \\
& Alkaline.phosphatase+ \\
& Major.dimension.of.nodule+ \\
& Ferritin
\end{aligned}
$$

Analysis using Subset of Predictors

Results for all models using a subset of predictors are reported in table 3. Overall, we find a significant reduction in test error rate for all models, with forward stepwise selection performing better than backward stepwise selection, with two exceptions. In general, logistic regression using and SVM performed better than other models and four of those models had test error rates below 20%. SVM using a polynomial kernel and backward step selection performed the best with a test error rate of 18.2%. However, the polynomial kernel uses degree = 1.

Table 3: Top Performing Models with Subset of Predictors

Models	Forward_Selection	Backward_Selection
Logistic Regression LOOCV	19.4	20.6
Logistic Regression 5-fold CV	24.8	24.8
MclustDA, Model Type = EDDA LOOCV	28.5	27.3
MclustDA, Model Type = EDDA 5-fold CV	26.7	26.7
SVM Linear LOOCV	20.0	20.6
SVM Linear 5-fold CV	19.4	20.0
SVM Polynomial 5-fold CV	19.4	18.2

Final Conclusion

This dataset attempts to find a relationship between several predictor variables in order to be able to predict patients' survival of HCC beyond 1 year. In our analysis, we have narrowed down the list of 44 predictor variables to just 22 using backward stepwise selection. The proposed model is:

$$\begin{aligned} Survival = {} & Alcohol+ \\ & Hepatitis.B.Surface.Antigen+ \\ & Hepatitis.C.Virus.Antibody+ \\ & Smoking+ \\ & Diabetes+ \\ & Hemochromatosis+ \\ & Arterial.Hypertension+ \\ & Nonalcoholic.Steatohepatitis+ \\ & Splenomegaly+ \\ & Portal.Hypertension+ \\ & Portal.Vein.Thrombosis+ \\ & Age.at.diagnosis+ \\ & Performance.Status+ \\ & Encefalopathy.degree+ \\ & Ascites.degree+ \\ & AlphaFetoprotein+ \\ & Haemoglobin+ \\ & Total.Bilirubin+ \\ & Alanine.transaminase+ \\ & Alkaline.phosphatase+ \\ & Major.dimension.of.nodule+ \\ & Ferritin \end{aligned}$$

There is indication that the shape of the decision boundary is in fact linear since the best performing models are SVM with a polynomial kernel adn degree = 1, SVM with linear kernel, and logistic regression. Additional data can potentially vastly improve the approximately 20% test error rate, and all three models should be reevaluated to determine the best model. In doing so, this data and prediction model will help doctors determine a particular patient's stage of HCC, and therefore determine best course of treatment.