Predicting Survival in AML Patients from RNASeq Data using SVM

```
#| echo: false
#| message: false

library(tidyverse)
```

Introduction

Acute myeloid leukemia (AML) is a cancer of the blood and bone marrow that affects the myeloid cells, which are responsible for producing red blood cells, platelets, and white blood cells. While AML is the most common kind of leukemia, it is usually very aggressive with limited therapeutic options and only around 20% patients can achieve durable remission.[1]

Dataset and Objectives

As part of the Beat AML 2.0 program, a dataset consisting of clinical outcomes, genomic and transcriptomic data from a cohort of 805 AML patients were collected.[2] Among which 571 patients have both RNASeq data and either survived for at least one year after diagnosis (n = 311) or have died (n = 260). We propose a methodology to predict the one-year survival of AML patients using a support vector machine (SVM) model based on transcriptomic (RNASeq) data of these patients. We will use the (pre-) normalized RNASeq z-score data (571x22843) as the input to the model. Thus no additional normalization is required.

Methodology

Boruta Feature Selection

Due to the data is already high-dimensional, we will use a feature selection method to reduce the dimensionality of the data. We will use the Boruta algorithm[3], which is a wrapper method that uses a random forest classifier to identify the most important features. It is a recursive algorithm that iteratively proves or disproves the importance of each feature.

Support Vector Machine (SVM)

Then we apply a support vector machine (SVM) to the selected features. The SVM model is a supervised learning model that can be used for classification. We will build SVM-based models to predict the one-year survival of AML patients. If we are able to reasonably control the number of features, we may consider using a non-linear kernel which may result in improved prediction.

Model Evaluation

Additionally, since the number of features is large, we need to ensure overfitting is minimized. Since we have somewhat limited (571 rows) number of observations we may use a 10-fold cross validation to evaluate the model. We will use the Matthews correlation coefficient (MCC) as the primary metric to evaluate the model.[4,5] The MCC is a measure of the quality of binary classifications, and is defined as:

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP}) * (\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TN} + \text{FN})}}$$

MCC is a very interpretable metric, as it ranges from −1 to 1, where 1 is a perfect prediction, 0 is a random prediction, and −1 is a perfectly wrong prediction.

Hyperparameter Tuning

We will use a grid search to find the optimal hyperparameters for the SVM.

Final Model Production

After an optimal model + hyperparameter set is selected, we will train the model on the entire dataset and present the final model.

Preliminary Results

The boruta algorithm was able to isolate 62 features deemed important after 1000 iterations. Of which 57 were confirmed important and 5 were tentative.

The AOC curve of a linear model fitted with or without feature selection is shown below.

```
#| echo: false
#| message: false
svm_boruta_roc <- read_csv("stat5353_project/svm_boruta_roc.csv")

svm_boruta_roc %>%
  mutate(fs = "boruta") %>%
  filter(c_pos == 10 & c_neg == 10) %>%
  bind_rows(read_csv("stat5353_project/svm_roc.csv") %>% filter(dropped == "NONE") %>% mutate(fs =
"none")) %>%
  ggplot(aes(x = x, y = y, color = fs)) +
  geom_line() +
  coord_fixed() +
  labs(color = "Feature Selection")
```

Bibliography

(1) Pulte, D. ; Jansen, L. ; Castro, F. A. ; Krilaviciute, A. ; Katalinic, A. ; Barnes, B. ; Ressing, M. ; Holleczek, B. ; Luttmann, S. ; Brenner, H. ; Group, f. t. G. C. S. W. Survival in Patients with Acute Myeloblastic Leukemia in Germany and the United States: Major Differences in Survival in Young Adults. *International Journal of Cancer* **2016**, *139* (6), 1289–1296. https://doi.org/10.1002/ijc.30186

(2) Bottomly, D. ; Long, N. ; Schultz, A. R. ; Kurtz, S. E. ; Tognon, C. E. ; Johnson, K. ; Abel, M. ; Agarwal, A. ; Avaylon, S. ; Benton, E. ; Blucher, A. ; Borate, U. ; Braun, T. P. ; Brown, J. ; Bryant, J. ; Burke, R. ; Carlos, A. ; Chang, B. H. ; Cho, H. J. ; Christy, S. ; Coblentz, C. ; Cohen, A. M. ; d'Almeida, A. ; Cook, R. ; Danilov, A. ; Dao, K.-H. T. ; Degnin, M. ; Dibb, J. ; Eide, C. A. ; English, I. ; Hagler, S. ; Harrelson, H. ; Henson, R. ; Ho, H. ; Joshi, S. K. ; Junio, B. ; Kaempf, A. ; Kosaka, Y. ; Laderas, T. ; Lawhead, M. ; Lee, H. ; Leonard, J. T. ; Lin, C. ; Lind, E. F. ; Liu, S. Q. ; Lo, P. ; Loriaux, M. M. ; Luty, S. ; Maxson, J. E. ; Macey, T. ; Martinez, J. ; Minnier, J. ; Monteblanco, A. ; Mori, M. ; Morrow, Q. ; Nelson, D. ; Ramsdill, J. ; Rofelty, A. ; Rogers, A. ; Romine, K. A. ; Ryabinin, P. ; Saultz, J. N. ; Sampson, D. A. ; Savage, S. L. ; Schuff, R. ; Searles, R. ; Smith, R. L. ; Spurgeon, S. E. ; Sweeney, T. ; Swords, R. T. ; Thapa, A. ; Thiel-Klare, K. ; Traer, E. ; Wagner, J. ; Wilmot, B. ; Wolf, J. ; Wu, G. ; Yates, A. ; Zhang, H. ; Cogle, C. R. ; Collins, R. H. ; Deininger, M. W. ; Hourigan, C. S. ; Jordan, C. T. ; Lin, T. L. ; Martinez, M. E. ; Pallapati, R. R. ; Pollyea, D. A. ; Pomicter, A. D. ; Watts, J. M. ; Weir, S. J. ; Druker, B. J. ; McWeeney, S. K. ; Tyner, J. W. Integrative Analysis of Drug Response and Clinical Outcome in Acute Myeloid Leukemia. *Cancer Cell* **2022**, *40* (8), 850–864. https://doi.org/10.1016/j.ccell.2022.07.002

(3) Kursa, M. B. ; Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal of Statistical Software* **2010**, *36*, 1–13. https://doi.org/10.18637/jss.v036.i11

(4) Chicco, D. ; Tötsch, N. ; Jurman, G. The Matthews Correlation Coefficient (MCC) Is More Reliable Than Balanced Accuracy, Bookmaker Informedness, And Markedness in Two-Class Confusion Matrix Evaluation. *BioData Mining* **2021**, *14* (1), 1–22. https://doi.org/10.1186/s13040-021-00244-z

(5) Chicco, D. ; Jurman, G. The Matthews Correlation Coefficient (MCC) Should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification. *BioData Mining* **2023**, *16* (1), 1–23. https://doi.org/10.1186/s13040-023-00322-4