IME672A: DATA MINING AND KNOWLEDGE DISCOVERY

Survival Prediction of Children Undergoing Hematopoietic Stem Cell Transplantation Using Different Machine Learning Classifiers

Project Report

1. Introduction

Bone Marrow Transplant (BMT), a gradational rescue for a wide range of neoplastic, chronic, and allied disorders emanating from the bone marrow, is an efficacious surgical treatment. Several risk factors, such as post-transplant illnesses, new malignancies, and even organ damage, can impair long-term survival after BMT. Therefore, technologies like Machine Learning (ML) are being linked to hematology for investigating the survival prediction of BMT receivers along with the influences that limit their resilience. In this study, a comprehensive approach is undertaken where an efficient survival classification model is presented, incorporating SMOTE to balance data and ensembling methods like Bagging and Adaboost to improve accuracy. The motivation of the study was to identify the most important factors influencing the success or failure of the transplantation procedure. In particular, the aim was to verify the hypothesis that increased dosage of CD34+ cells / kg extends overall survival time without simultaneous occurrence of undesirable events affecting patients' quality of life (Kawłak et al., 2010).

2. Dataset Description

The dataset used in this study covers medical information for children who have been diagnosed with a variety of hematologic diseases and who underwent unmodified allogeneic unrelated donor HSCT. Hence, this dataset comprises 187 occurrences and 37 attributes that contain information about individuals who have been diagnosed with a range of hematologic malignant or benign diseases. Most of the properties contain categorical data, while others contain Boolean and numerical values. The dataset's attributes are listed below. Following data extraction, it was subjected to exploratory data analysis using Jupyter Notebook and Python to find out the dataset's properties. The dataset has many categorical attributes and missing values. The statistical data for the dataset's numerical variables are summarized in Table 2.

Dataset attributes

There are 10 numeric, 6 boolean, 21 categorical attributes.

Numeric attributes:

1. **donor_age** - Age of the donor at the time of hematopoietic stem cells apheresis
2. **recipient_age** - Age of the recipient of hematopoietic stem cells at the time of transplantation
3. **recipient_body_mass** - Body mass of the recipient of hematopoietic stem cells at the time of the transplantation
4. **CD34_x1e6_per_kg** - CD34kgx10d6 - CD34+ cell dose per kg of recipient body weight (10^6/kg)
5. **CD3_x1e8_per_kg** - CD3+ cell dose per kg of recipient body weight (10^8/kg)
6. **CD3_to_CD34_ratio** - CD3+ cell to CD34+ cell ratio
7. **ANC_recovery** - Time required for neutrophil recovery is defined as a neutrophil count greater than 0.5 x 10^9/L
8. **PLT_recovery** - Platelet reproducing period is defined as count >50000/mm3
9. **time_to_acute_GvHD_III_IV** - Time required for the onset of stage III or IV acute graft against host disease

10. **survival_time** -I n days, the time of observation or time to event

Boolean attributes:

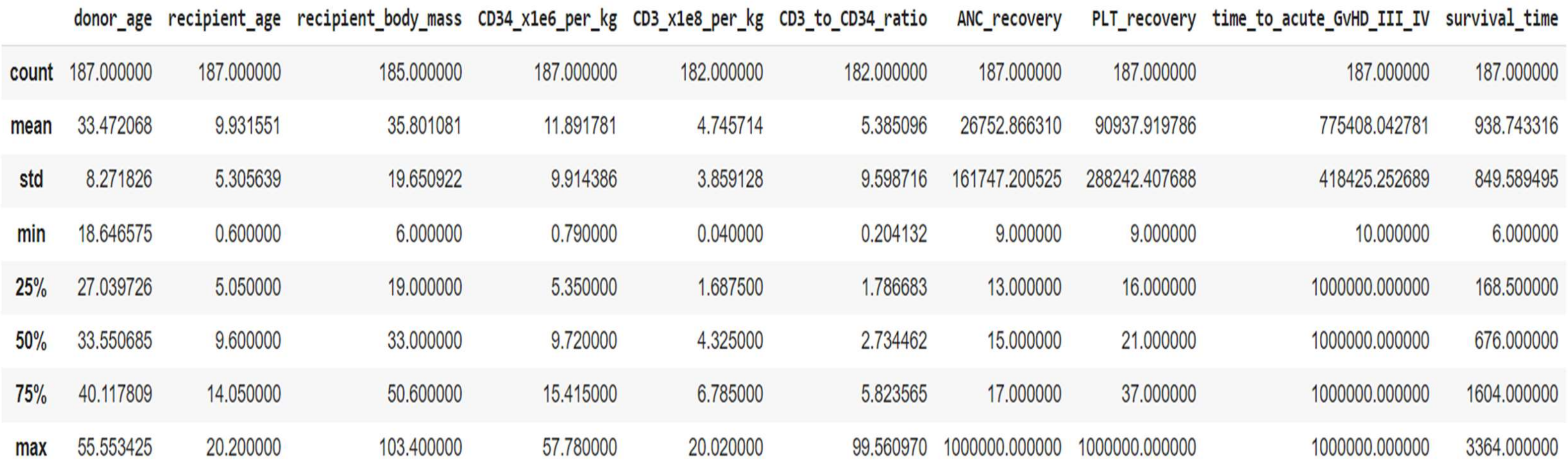
1. **donor_age_below_35** - Is the donor under the age of 35?
2. **recipient_age_below_10** - Is the recipient's age under ten?
3. **tx_post_relapse** - The second bone marrow transplant following recurrence
4. **acute_GvHD_II_III_IV** - Development of stage II, III, or IV acute graft versus host disease
5. **acute_GvHD_III_IV** - Stage III or IV growth of acute graft versus host disease
6. **relapse** - Disease relapse

Categorical attributes:

1. **disease** - Disease classification
2. **disease_group** - Malignant or nonmalignant
3. **gender_match** - Gender compatibility between donor and recipient
4. **ABO_match** - HSC donor-recipient blood group compatibility
5. **CMV_status** - Serological compatibility of hematopoietic stem cell donors and recipients based on CMV infection prior to transplantation
6. **HLA_match** - Antigen compatibility between the donor and receiver of hematopoietic stem cells
7. **HLA_mismatch** - HLA mismatches or matches
8. **antigen** - How many antigens differ between the donor and receiver
9. **allel** - How many alleles differ between the donor and receiver
10. **HLA_group_1** - The donor-recipient difference
11. **risk_group** - Group at risk
12. **stem_cell_source** - Hematopoietic stem cell source
13. **extensive_chronic_GvHD** - Chronic graft versus host disease develops to a large extent
14. **survival_status** - Status of survival

15. **donor_ABO** - The hematopoietic stem cell donor's ABO blood group
16. **donor_CMV** - Cytomegalovirus infection prior to transplantation in the donor of hematopoietic stem cells
17. **recipient_age_int** - Distinct intervals of the recipient's age
18. **recipient_gender** - The recipient's gender
19. **recipient_ABO** - The recipient's ABO blood group
20. **recipient_rh** - The Rh factor is present on the recipient's red blood cells
21. **recipient_CMV** - Cytomegalovirus infection prior to transplantation in the donor of hematopoietic stem cells

Following data extraction, it was subjected to exploratory data analysis using Jupyter Notebook and Python to find out the dataset's properties. The dataset has many categorical attributes and missing values. The statistical data for the dataset's numerical variables are summarized below.



	donor_age	recipient_age	recipient_body_mass	CD34_x1e6_per_kg	CD3_x1e8_per_kg	CD3_to_CD34_ratio	ANC_recovery	PLT_recovery	time_to_acute_GvHD_III_IV	survival_time
count	187.000000	187.000000	185.000000	187.000000	182.000000	182.000000	187.000000	187.000000	187.000000	187.000000
mean	33.472068	9.931551	35.801081	11.891781	4.745714	5.385096	26752.866310	90937.919786	775408.042781	938.743316
std	8.271826	5.305639	19.650922	9.914386	3.859128	9.598716	161747.200525	288242.407688	418425.252689	849.589495
min	18.646575	0.600000	6.000000	0.790000	0.040000	0.204132	9.000000	9.000000	10.000000	6.000000
25%	27.039726	5.050000	19.000000	5.350000	1.687500	1.786683	13.000000	16.000000	1000000.000000	168.500000
50%	33.550685	9.600000	33.000000	9.720000	4.325000	2.734462	15.000000	21.000000	1000000.000000	676.000000
75%	40.117809	14.050000	50.600000	15.415000	6.785000	5.823565	17.000000	37.000000	1000000.000000	1604.000000
max	55.553425	20.200000	103.400000	57.780000	20.020000	99.560970	1000000.000000	1000000.000000	1000000.000000	3364.000000

3. Workflow

3.1 Data Preprocessing

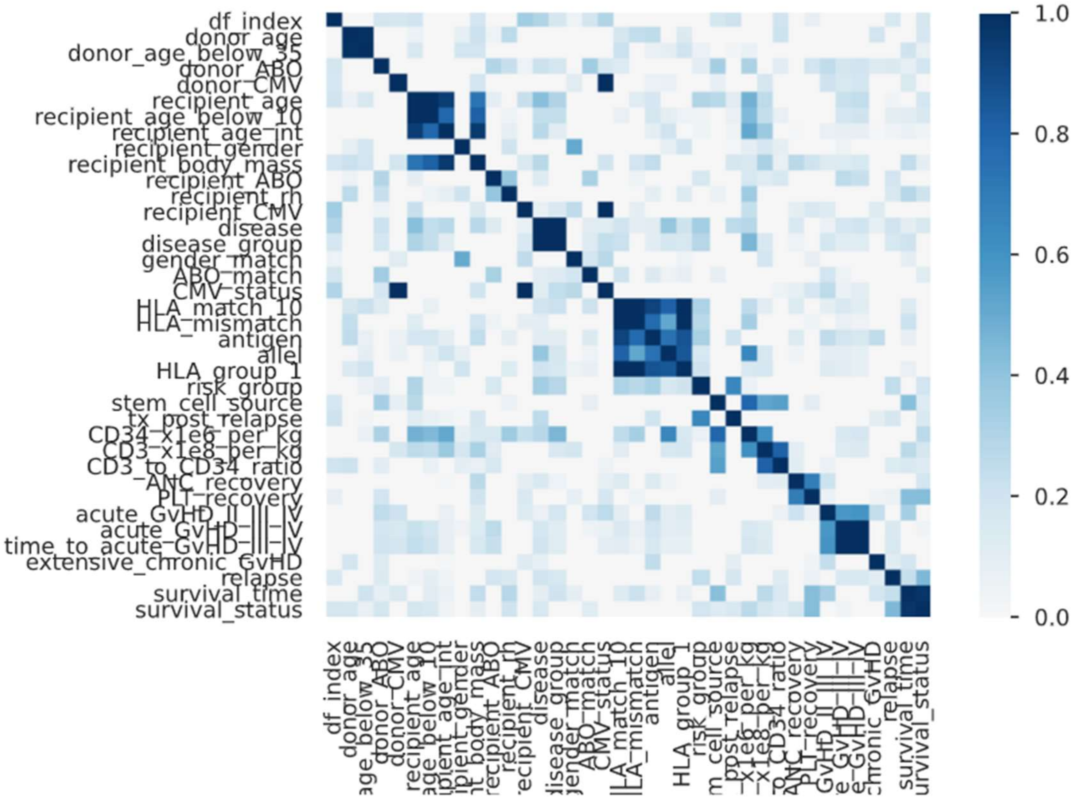
The dataset underwent multiple preprocessing stages before being used in ML models.

1. The missing values of the dataset were filled with median values for numerical ones and most frequent values for categorical ones.
2. The column ‘id’ is dropped as it is not useful.
3. Outliers in the data were replaced by median values.
4. Since categorical data cannot be handled by ML models, the categorical variables were encoded into numerical form. LabelEncoder from sklearn was employed for this purpose.
5. Data was normalized using the min-max scaling method.

6. Class imbalance problem is dealt by balancing the data by employing the SMOTE algorithm.

3.2 EDA

1. For aiding the outlier analysis box plots were plotted for numerical attributes.
2. To gain insights into data, pair plots for numerical attributes and bar plots for categorical attributes were plotted.
3. To discover the correlation between attributes, the correlation heatmap is essential and is generated in the pandas profiling report.



Results of EDA:

- If the donor age is less than 20 then the recipient has survived.
- Most of the donors are less than 35 years of age.

- If the donor blood group is A then it is more vulnerable for death of the recipient, If donor BG is AB then more chances of the recipient for survival.
- If donor CMV is present, the recipient has some more chance for survival than if donor_CMV is absent.
- recipient age lies between 0 to 20 , max of people are of age 4 to 5 or 13 to 14.
- Most recipients have body mass between 10-30.
- Lower age recipients have more chances for survival.Especially which are below 10, and if age is below 2 of recipients then he has a high percentage of Survival.
- No outlier in recipient age all values lie between 0 to 20.
- If recipient age is below 10 then there are good chances of its survival, otherwise if recipient age is more than 10 then there are approx 50percent chances of survival.
- If the recipient age is below 10 then there are good chances of its survival, otherwise if recipient age is more than 10 then there are approx 50 percent chances of survival.
- Males Survive more than that of females (in percentages).
- the one with body mass more than 75 has not survived.
- when CD34_x1e6_per_kg is above 10, the survival chance is high.
- Survival rate for the recipients is high with CD3 cell dose per kg above 7 and low if it is below 2.
- Most recipients have body mass between 10-30.
- the one with body mass more than 75 has not survived.
- Higher the Compatibility of antigens (HLA_match) causes higher probability of alive
- Lower the antigens difference between the doner and the recipient causes higher probability of being alive.
- If the survival time is below 500, that person most probably won’t survive.
- The chance of survival is higher when relapse is yes.
- More chances of survival when acute_GvHD_II_III_IV IS NO.
- This KDE is negatively skewed, so the mean of HLA_matched is less than median. Most of the HLA_matched matches range between 8.5-10.5.

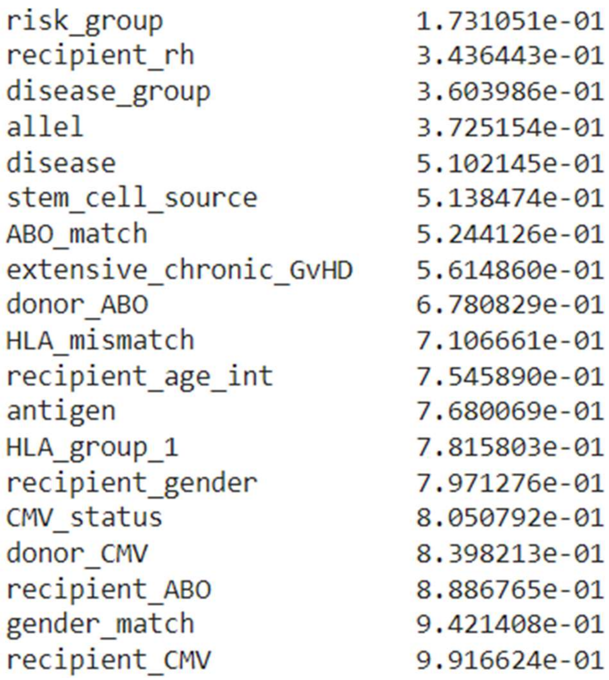
3.3 ML algorithms used to build models

1. The dataset was split into train and test sets in proportions of 80% and 20%, respectively.
2. 6 ML algorithms: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Classifier (SVC) were fed and trained on this dataset, and performance metrics were obtained.
3. Two ensembling methods were used to improve the accuracy of some models. Bagging was used with DT as base classifier and Adaboost was used to improve the accuracy of DT classifier and SVC.

4. Results

4.1 Chi-squared test

Chi-squared feature ranking is performed on categorical attributes, and the resulting scores are summarized. Based on this “Risk_group” is the most important column.

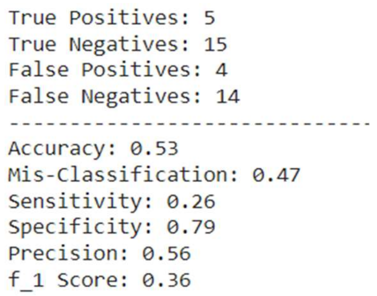


```
risk_group                 1.731051e-01
recipient_rh               3.436443e-01
disease_group              3.603986e-01
allel                      3.725154e-01
disease                    5.102145e-01
stem_cell_source           5.138474e-01
ABO_match                  5.244126e-01
extensive_chronic_GvHD     5.614860e-01
donor_ABO                  6.780829e-01
HLA_mismatch               7.106661e-01
recipient_age_int          7.545890e-01
antigen                    7.680069e-01
HLA_group_1                7.815803e-01
recipient_gender           7.971276e-01
CMV_status                 8.050792e-01
donor_CMV                  8.398213e-01
recipient_ABO              8.886765e-01
gender_match               9.421408e-01
recipient_CMV              9.916624e-01
```

4.2 Performance metrics of models

1. ANN

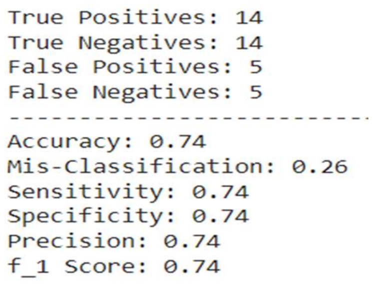
AUROC = 0.526



```
True Positives: 5
True Negatives: 15
False Positives: 4
False Negatives: 14
------------------------------
Accuracy: 0.53
Mis-Classification: 0.47
Sensitivity: 0.26
Specificity: 0.79
Precision: 0.56
f_1 Score: 0.36
```

2. SVC

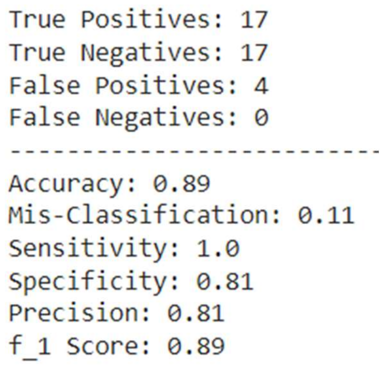
AUROC = 0.737



```
True Positives: 14
True Negatives: 14
False Positives: 5
False Negatives: 5
-------------------------
Accuracy: 0.74
Mis-Classification: 0.26
Sensitivity: 0.74
Specificity: 0.74
Precision: 0.74
f_1 Score: 0.74
```

3. Logistic Regression

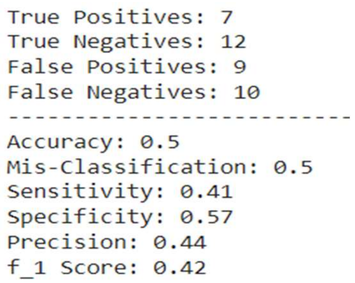
AUROC= 0.905



```
True Positives: 17
True Negatives: 17
False Positives: 4
False Negatives: 0
--------------------------
Accuracy: 0.89
Mis-Classification: 0.11
Sensitivity: 1.0
Specificity: 0.81
Precision: 0.81
f_1 Score: 0.89
```

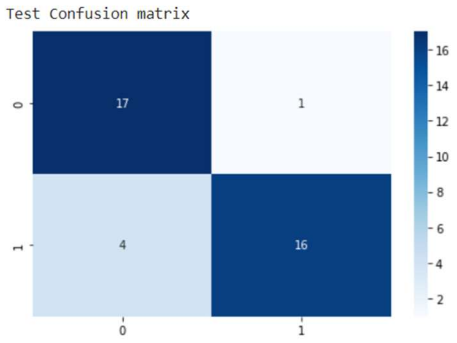
4. KNN

AUROC=0.396



```
True Positives: 7
True Negatives: 12
False Positives: 9
False Negatives: 10
--------------------------
Accuracy: 0.5
Mis-Classification: 0.5
Sensitivity: 0.41
Specificity: 0.57
Precision: 0.44
f_1 Score: 0.42
```

5. DT



Test Confusion matrix

Accuracy = 86.84

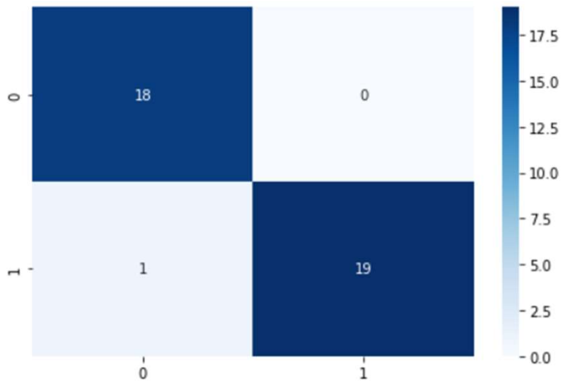
Precision = 0.8

Recall = 0.94

F1 score = 0.86

AUROC = 1

6. RF



Accuracy = 97.36

Precision = 0.8

Recall = 0.94

F1 score = 0.86

AUROC = 1

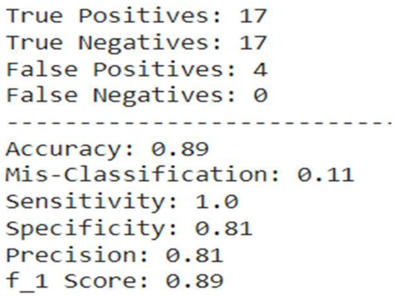
7. Bagging with DT

Accuracy of model before bagging = 91.05

Accuracy of model before bagging = 97.36

8. Adaboost

   AUROV = 0.905



```
True Positives: 17
True Negatives: 17
False Positives: 4
False Negatives: 0
--------------------------
Accuracy: 0.89
Mis-Classification: 0.11
Sensitivity: 1.0
Specificity: 0.81
Precision: 0.81
f_1 Score: 0.89
```

5. Conclusion

A Bone Marrow Transplant is a crucial life-saving treatment for a certain type of malignancy. For this reason, early detection of survivability after BMT can play a vital role in the patient's treatment process. Moreover, if healthcare providers have a prior prediction, they can make more informed decisions about treatment options. In this regard, technologies like ML can be useful, since they can be used in situations requiring prediction and can uncover hidden patterns in previous data in order to create an accurate prediction. Nowadays, it is increasingly being employed in every situation that requires prediction. All ML models yielded satisfactory predictions. DT, Bagging using DT, RF have very high accuracy compared to other models. The top five attributes that influence the survivability rate are "PLT recovery", "ANC recovery", "survival time", and "recipient body mass",

6. References

- “What is a Bone Marrow Transplant? | Be The Match.” https://bethematch.org/patients-and-families/about-transplant/what-is-abone-marrow-transplant-/
- “Bone Marrow Transplant: Types, Procedure & Risks.” https://www.healthline.com/health/bone-marrow-transplant#preparation
- “What is a Bone Marrow Transplant (Stem Cell Transplant)? | Cancer.Net.” https://www.cancer.net/navigating-cancer-care/howcancer-treated/bone-marrowstem-cell-transplantation/what-bone-marrow-transplant-stem-cell-transplant
- “Bone Marrow Transplant: Preparation, Procedure, Risks, and Recovery.” https://www.webmd.com/cancer/multiplemyeloma/bone-marrow-transplants
- “Blood and Marrow Transplant Statistics and Outcomes.” https://www.chp.edu/our-services/blood-marrow-transplant-cellulartherapies/statistics-outcomes