Effect of CD34+ Concentration on Hematopoietic Stem Cell Transplants in Children

Rudraharsh Tewary
University of Toronto

December 15, 2023

1 Introduction

In recent times, hematological diseases have increased in prevalence. Comprising of two major categories, Malignant(Cancerous) and Non-Malignant disorders. Owing to the painful nature of these diseases and high mortality rates, it has been imperative to find a treatment that can effectively combat and cure these conditions. One promising avenue for this is Hematopoietic stem cell transplants, which have been shown to be effective in treating malignant disorders and prolonging the lifespans of those suffering from non-malignant cases. In our project, we consider the case of pediatric patients suffering from such disorders, with main motivating factors being the absence of more data due to low prevalence rates, while at the same time, more emphasis has to be placed on treatment owing to the fragile nature of our patients considering the age. Our goal is to analyze the effect of the concentration of a certain cell 'CD34+', mainly, ongoing research supports[1],[6],[3],[5] that increasing the concentration of this specifc cell during transplants has improvements in survival metrics for our patients. So, we will attempt to verify whether our analysis also predicts the same trend.

2 Data

Our Dataset [7]has been collected from the UC Irvine Machine Learning Repository, Titled 'Bone Marrow Transplants: Children'. It contains pediatric patients with a variety of Malignant Disorders (Acute Lymphoblastic Leukemia (ALL), , Chronic Myelogenous Leukemia (CML), Myelodysplastic Syndrome (MDS)) and Non-Malignant Disorders (Severe Aplastic Anemia (SAA), , with X-linked Adrenoleukodystrophy (X-ALD)). All patients in the dataset were treated with Unrelated Allogeneic Hematopoietic Stem Cell Transplant (UHSCT). We have a total of 187 observations with 37 features for each observation. There are 12 continuous features and 25 discrete features in the dataset. A table containing the name, type and description of each feature has been included in the report and is a part of the Appendix.

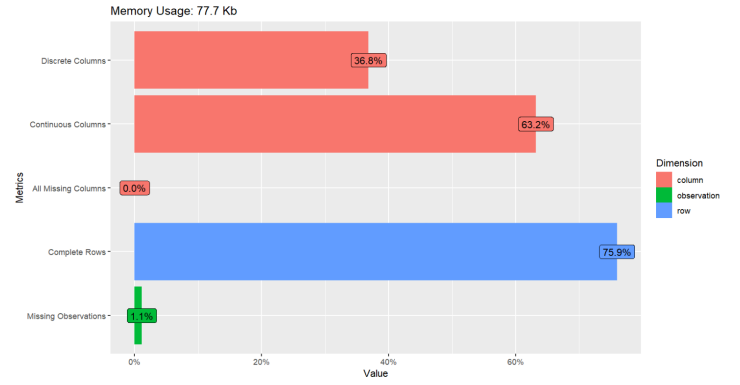
2.1 Data Collection

The dataset *Bone Marrow Transplant: Children* was collected to analyze the effect of multiple factors present during a bone marrow transplant, specfic to the case of pediatric patients. It is not described how the curators collected the records comprising the dataset.

2.2 Data Description

The Dataset has 187 observations with 37 features, with 12 of the features being continuous and 25 of them being discrete.

Initial analysis of the dataset showed that multiple categorical features were listed as a 'character' data types and missing values were represented as '?' instead of 'NA', so cleaning was done to fix those issues. After which, we get this initial plot



As we can see, around 76% of our observations are fully complete, giving us around 142 workable observations. However, it must be noted that the amount is still not ideal compared to the number of distinct features present in the dataset, as the ideal ratio usually is 10 observations per feature.

Now, we will see the features containing missing values

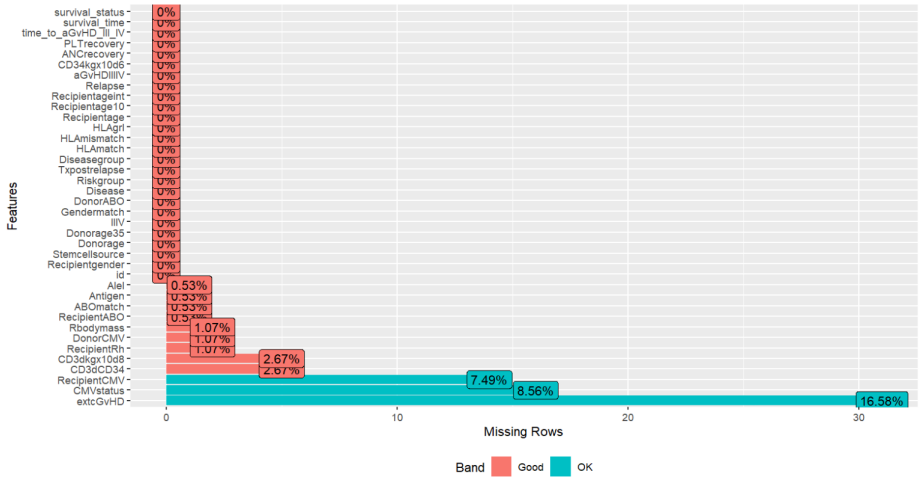
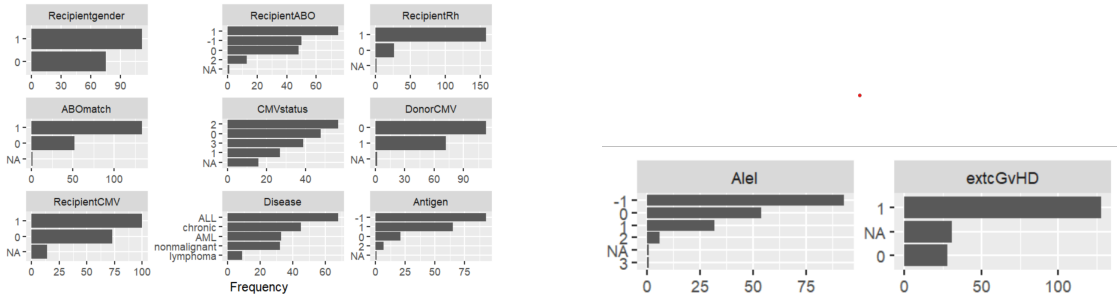


Figure 1: Features by missing value %

Here, we see that the features with the highest amount of missing values are '*RecipientCMV*','*CMVstatus*' and '*extcGvHD*' which tell us about the cytomegalovirus (CMV) status of the patient, donor and development of chronic Graft vs Host disease respectively. As these features are important to our response '*survival_status*' from a theoretical perspective, and it is not feasible to accurately impute these values, we will focus our analysis on the complete subset of our dataset.

Now, we look at the distribution of some of our categorical variables:



(a) Figure 1

(b) Figure 2

Figure 2: Frequency distribution of Categorical variables

We can see a very clear trend that a majority of the patients in our dataset have malignant disorders. That is, a majority of them are suffering from a form of leukemia, while the remaining patients suffer from non-malignant disorders like anemia, adrenoleukodystrophy, etc. Thus, due to the low number of non-malignant cases, we cannot make any strong judgements on the applicability of our results on this specific subset of the data.

Now, we will proceed to look at the distribution of the continuous variables:

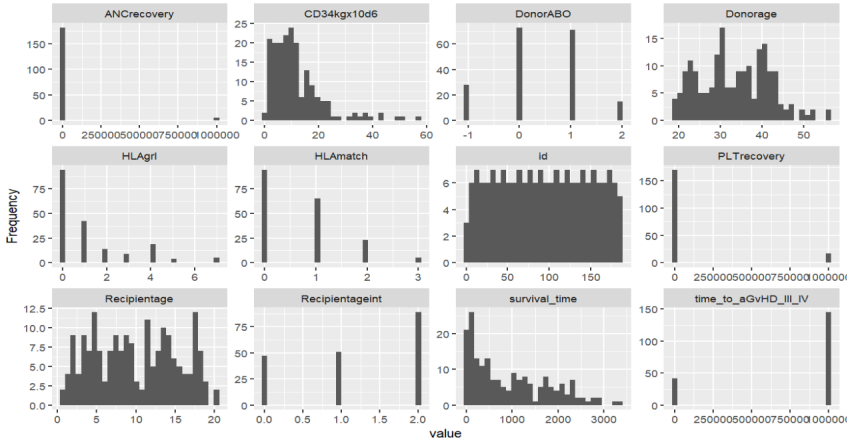


Figure 3: Histogram of Continuous variables

Now, although the distribution is mostly as expected. We can clearly see there are certain issues in some features. For example, features like '*ANCrecovery*','*PLTrecovery*','*time_to_aGvHD_III_IV*' have a skewed disribution of most values at 0 and then some at 1000000. This is occuring as patients who never recovered their s/s etc. to baseline levels or patients who never developed acute graft vs host disease, have the time coded in as 1000000. We solve this problem by removing the entries with the value of '1000000' and introducing three categorical features known as '*PLTrecovered*','*ANCrecovered*' and '*aGvHD_Developed*' with '1' for recovered and '0' for not recovered.

We then proceed to visualize the distribution without the '1000000' entries:

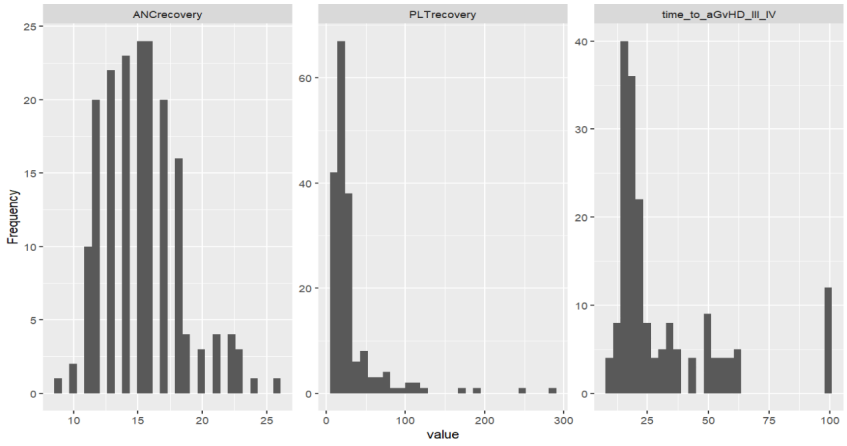


Figure 4: Adjusted plot of continuous variables

Finally, we proceed to create a correlation heatmap of our features:

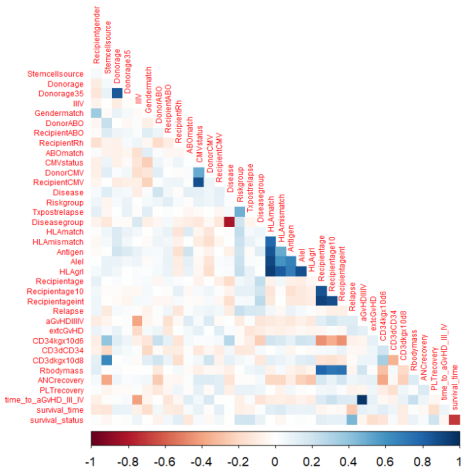


Figure 5: Correlation Heatmap of all our features

As we can see from the heatmap, most features are sparsely correlated. Which indicates that we might require most of them for our models.

Finally, for the purpose of model fitting, the feature 'Disease', was One-Hot-Encoded and decomposed into multiple features like '*DiseaseAML*', '*DiseaseALL*' etc. for easier model fitting. This has no impact on the information of the dataset.

3 Methods and Analysis

Now, our primary feature of interest in the dataset is '*survival_status*', which indicates whether a patient survived or not after being treated with a Hematopoietic stem cell transplant. As it is a binary categorical variable, we would require a classification model. Also, owing to our research question, all our models have been chosen with interpretability in mind.

Now, as we can see, due to the sparse nature of our correlation plot, we can see that its redundant to perform variable selection on our dataset. Therefore, we will move to analyzing which models we can use for classification.

Our First choice will be a logistic regression model, expressed using the formula

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

Not only is it efficient and usable for the task of binary classification, the model and the coefficients produced from it will allow us to perform inference and evaluate impact of a feature and its importance.

Our second choice would be a decision tree model, represented by the figure here :

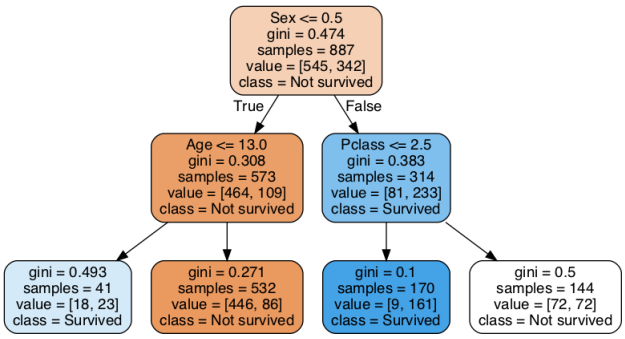


Figure 6: Example figure of Decision Tree Model[2]

the advantages of a decision tree model are offered by its ability to handle non-linear relationships in the data, as well as the ability to better model mixed data, which is important to us due to the heavy presence of categorical variables in the dataset. Also, another advantage is the interpretability of a decision tree due to its simple structure.

Our final choice will be the random forest model, the key operating idea of a random forest model is to combine and average multiple decision trees, its represented by the following formula and figure

$$y = \frac{1}{T} \sum_{t=1}^{T} y_t(x)$$

Here, y_t is the prediction for input x, T is the number of trees and y is the true label.

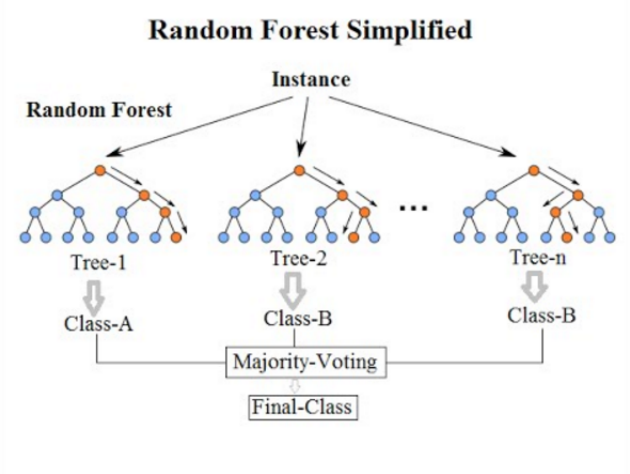


Figure 7: Representation of Random Forest[4]

The random forest model captures all the advantages of a decision tree model, along with the fact that due to the averaging of multiple trees, it gives better accuracy and reduces overfitting, while at the same time, adds further robustness to noise and outliers.

Due to the relatively small number of observations in our dataset, we will be evaluating the models under 10-fold cross validation to get an idea of their performance.

4 Application and Results

So, we first perform 10-fold cross validation on all 3 models fitted on all covariates and get the following results for their accuracy :

Method	Accuracy
Logistic Regression	69.76%
Decision Tree	70.48%
Random Forest	72%

Table 1: Accuracy scores for our tried models

Now, one key thing that we notice is that all models have relatively poor accuracy, less than the desired accuracy of 85% or 90% or more. One explanation for this could be the small size of the dataset. As we would want around 10 observations per feature, instead we are operating with half of the desired amount. So, in the future, we might be able to improve the accuracy through a bigger dataset. So, as the random forest model has the highest accuracy in our tested models, we will proceed forward with it for our analysis.

Now, we can rank how important a feature is to our random forest model, by computing a metric known as 'Mean decrease in impurity', which measures the impact on the accuracy of our model when it begins to randomly permute values of a given feature. If the impact is significant,

then the feature has higher importance. If not, then that feature is less important. So, on the basis of that we obtain the following feature importance plot:

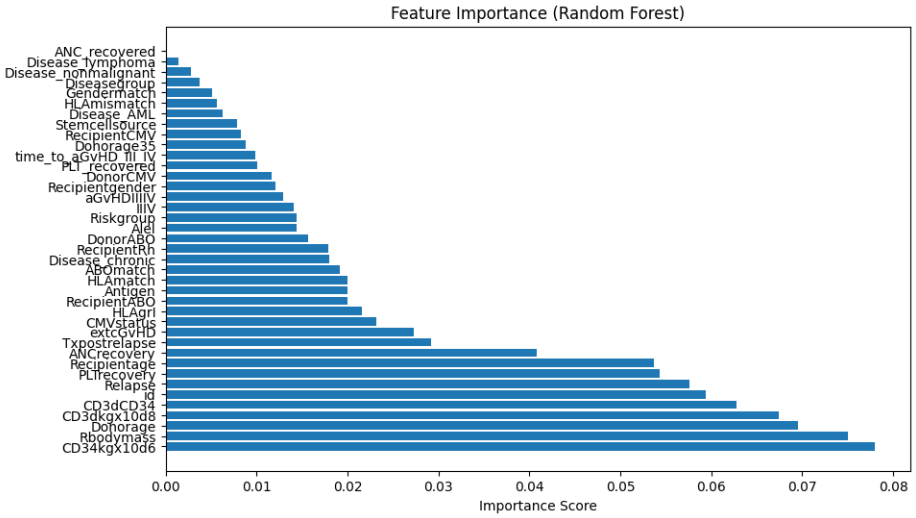


Figure 8: Feature importances determined by fitted random forest model

We see that the main feature of interest in our hypothesis '*CD34kgx10d6*', is recognized as the most important feature of our analysis. So, we focus on how change in its levels impact our response variable.

To check how an individual feature impacts our response variable, we use a partial dependence plot:

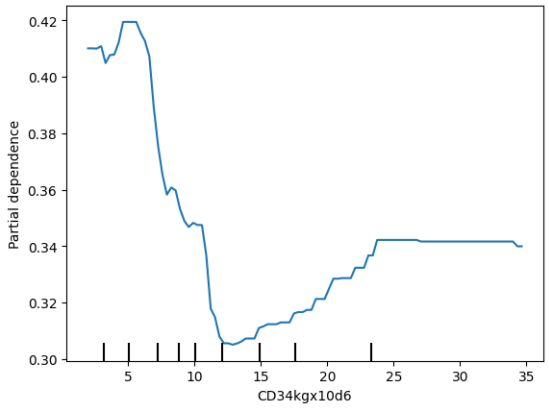


Figure 9: Partial Dependence Plot of *CD34kgx10d6*

A partial dependence plot allows us to determine the impact of a covariate on the value of the response, in this case as our response is binary, the y-axis would then represent the partial expectation of response. with higher values closer to 1 indicating the response will be '1' and lower values

closer to 0 indicating '0'. In our scenario, we want it to be closer to '0' as 0 indicates survival. As we can see, there is a trend observed that as the concentration of CD34+ increases, our patient's odds of survival increase as well.

5 Summary and Conclusions

We use the *Bone Marrow Transplant : Children* Dataset, and inspect whether there are any specific features impacting transplant success for pediatric patients suffering from hematological diseases. In order to make our analysis more robust, we only focus on the set of complete records. We attempt to fit classification models like logistic regression, decision trees and random forest. Using the best performing one, we inspect whether our feature of interest, '*CD34kgx10d6*' has a noticeable impact on the response. We discover that our model suggests that concentrations of CD34+ do have an impact on survival, with higher concentrations being positively correlated with survival. However, our conclusions are not that strong due to the accuracy of the model being relatively low, and a possible direction in future to mend this problem would be access to more data to make a robust conclusion

After analyzing the *Bone Marrow Transplant : Children* dataset, performing slight data engineering and using traditional statistical classification models. We discover that our model suggests there exists a correlation between the concentration of 'CD34+' known as '*CD34kgx10d6*' in the dataset. Which is also supported by literature. However, due to relatively low accuracy of the classification model, the link may be taken to be weak.

Glossary

Acute Lymphoblastic Leukemia (ALL) A type of cancer that starts in the bone marrow and affects white blood cells called lymphocytes.. 1

Chronic Myelogenous Leukemia (CML) A type of cancer that starts in the bone marrow and causes abnormal growth of white blood cells called granulocytes.. 1

cytomegalovirus (CMV) A common virus that can cause infection in people with weakened immune systems.. 2

Myelodysplastic Syndrome (MDS) A condition in which the bone marrow doesn't produce enough healthy blood cells.. 1

Severe Aplastic Anemia (SAA) A life-threatening condition in which the bone marrow stops producing enough blood cells.. 1

Unrelated Allogeneic Hematopoietic Stem Cell Transplant (UHSCT) A medical procedure that replaces diseased stem cells with healthy stem cells from a donor who is not a family member.. 1

X-linked Adrenoleukodystrophy (X-ALD) A rare genetic disorder that affects the adrenal glands and nervous system, primarily in males.. 1

A Appendix A

Data Dictionary of Dataset Bone Marrow Transplant : Children

Index	Variable Name	Role	Type	Description
1	Recipientgender	Feature	Binary	1 for Male, 0 for Female
2	Stemcellsource	Feature	Binary	Source of Stem cells, 1 for Peripheral Blood, 0 for Bone Marrow)
3	DonorAge	Feature	Integer	Age of Donor at time of Stem Cell Apheresis
4	DonorAge35	Feature	Binary	Donor age $<$ 35-0, Donor age $\geq$ 35-1
5	IIIV	Feature	Binary	Development of Acute Graft vs Host Disease, Stage II,III or IV yes-1,no-0
6	GenderMatch	Feature	Binary	Compatibility of Donor and Recipient Based on Gender yes-1,no-0
7	DonorABO	Feature	Categorical	ABO blood group of the Donor of Hematopoietic Stem Cells (B = -1,O=0,A=1,AB=2)
8	RecipientABO	Feature	Categorical	ABO blood group of the Recipient of Hematopoietic Stem Cells (B = -1,O=0,A=1,AB=2)
9	RecipientRH	Feature	Binary	Presence of the RH factor on Recipient's Red blood cells ('+'- 1,'-'-0)
10	ABOmatch	Feature	Binary	Compatibility of the Donor and Recipient of the Hematopoietic Stem Cells according to ABO blood group (matched - 1, mismatched - 0)
11	CMVstatus	Feature	Categorical	Serological Compatibility of the Donor and the Recipient of the hematopoietic stem cell according to Cytomegalovirus
12	DonorCMV	Feature	Binary	Presence of Cytomegalovirus infection in the donor of the hematopoietic stem cell transplant (present-1,absent-0)
13	RecipientCMV	Feature	Binary	Presence of Cytomegalovirus infection in the donor of the hematopoietic stem cell transplant (present-1,absent-0)
14	Disease	Feature	Categorical	Type of Disease (ALL,AML,chronic,nonmalignant,lymphoma)
15	Riskgroup	Feature	Binary	High Risk-1,Low Risk-0
16	Txpostrelapse	Feature	Binary	Second Bone Marrow transplant post relapse (yes-1,no-0)
17	DiseaseGroup	Feature	Binary	Type of Disease (malignant-1,non-malignant-0)
18	HLAmatch	Feature	Categorical	Compatibility of antigens of the main Histocompatiblity complex of the donor and the recipient of hematopoietic stem cells according to all international BFM SCT 2008 Criteria (10/10 -0, 9/10-1,8/10-2,7/10-3(allele/antigens))
19	HLAmismatch	Feature	Binary	HLA matched-0, HLA mismatched -1
20	Antigen	Feature	Categorical	In how many antigens there is a difference between the donor and the recipient (-1 no difference, 0- one difference,1 - two differences, 2- three differences, 3- four differences)
21	Allele	Feature	Categorical	In how many alleles there is a difference between the donor and the recipient (-1 no difference, 0- one difference,1 - two differences, 2- three differences, 3- four differences)
22	HLAgrl	Feature	Categorical	The difference type between the donor and the recipient (HLA matched -0,Difference in one antigen -1, Difference in only one allele -2, Difference in only one DRB1 cell -3, two differences(two alleles or two antigens) -4, mismatched-5)

23	RecipientAge	Feature	Integer	Age of Recipient of Hematopoietic Stem cells at the time of transplantation
24	RecipientAge10	Feature	Binary	Recipient Age ¡10 = 0, Recipient Age ¿= 10 -1
25	RecipientAgeint	Feature	Categorical	Recipient Age (0,5],(5-10]-1,[10,20)-2
26	Relapse	Feature	Binary	Reoccurence of the disease (yes-1,no-0)
27	aGvHDIIIV	Feature	Binary	Development of Acute Graft vs Host Disease Stage III or IV (Yes - 0, no-1)
28	extcGvHD	Feature	Binary	Development of Extensive Chronic Graft vs Host Disease (Yes - 0, no-1)
29	CD34kgx10d6	Feature	Integer	CD34+ cell dose per kg of recipient Body weight
30	CD3dCD34	Feature	Integer	CD3+ to CD34 cell ratio
31	CD3dkgx10d8	Feature	Integer	CD3+ cell dose per kg of recipient weight
32	Rbodymass	Feature	Integer	Body mass of Recipient at time of transplant
33	ANCrecovery	Feature	Integer	Time to neutrophil engraftment defined as $0.5 * 10^9/L$
34	PLTrecovery	Feature	Integer	Time to platelet engraftment defined as $> 50000/mm^3$
35	time-to-aGvHD-III-IV	Feature	Integer	Time to development of acute Graft vs Host Disease Stage III or IV
36	survival-time	Feature	Integer	Time of Observation(If alive) Time till event (If dead)
37	survival-status	Target	Integer	(0-Alive,1-Dead)

Table 3: Table containing information of all features

The following link contains the .Rmd File used to do the Exploratory Data Analysis, the .ipynb file used for model fitting and analysis and the copies of the Full and complete subset of the Dataset, as well as an image of the complete Random Forest model :
`https://utoronto-my.sharepoint.com/:f:/r/personal/rudraharsh_tewary_mail_utoronto_ca/Documents/Applied%20Stat%20Appendix?csf=1&web=1&e=zhc58T`

References

[1] Sara Bowman et al. "CD34 Stem Cell Boost in Pediatric Allogeneic Stem Cell Transplant Recipients: A Case Series and Review of Literature". In: *Clinical Hematology International* 5 (2023), pp. 155–164. URL: `https://api.semanticscholar.org/CorpusID:258008865`.

[2] Mikkel Duif. *An Introduction to Decision Trees with Python and scikit-learn — towardsdatascience.com*. `https://towardsdatascience.com/an-introduction-to-decision-trees-with-python-and-scikit-learn-1a5ba6fc204f`. [Accessed 16-12-2023].

[3] Tristan E. Knight et al. "Effect of Autograft CD34 + Dose on Outcome in Pediatric Patients Undergoing Autologous Hematopoietic Stem Cell Transplant for Central Nervous System Tumors". In: *Transplantation and Cellular Therapy* (2022). URL: `https://api.semanticscholar.org/CorpusID:263423178`.

[4] Will Koehrsen. *Random Forest Simple Explanation — williamkoehrsen.medium.com*. https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d. [Accessed 16-12-2023].

[5] Alexandra Pedraza et al. "Effect of CD34+ cell dose on the outcomes of allogeneic stem cell transplantation with post-transplant cyclophosphamide." In: *Transplantation and cellular therapy* (2022). URL: https://api.semanticscholar.org/CorpusID:254717130.

[6] Mats Remberger et al. "The CD34+ Cell Dose Matters in Hematopoietic Stem Cell Transplantation with Peripheral Blood Stem Cells from Sibling Donors". In: *Clinical Hematology International* 2 (2020), pp. 74–81. URL: https://api.semanticscholar.org/CorpusID:215957183.

[7] Sikora,Marek, Wróbel,Łukasz, and Gudyś,Adam. *Bone marrow transplant: children*. https://archive.ics.uci.edu/dataset/565/bone+marrow+transplant+children. DOI: https://doi.org/10.24432/C5NP6Z. 2020.