The National School of Artificial Intelligence

Machine Learning Project Report

Bone marrow transplant: children

May 25, 2024

First name	Last name	Group	Section
Omar Farouk	Zouak	Group 4	Section2
Abderrahim	Rezki	Group 6	Section 2
Houssam Eddine	Boukhalfa	Group 6	Section 2

Contents

1 Abstract

Hematopoietic stem cell transplantation (HSCT) is a critical medical procedure used to treat various hematological diseases. The success of HSCT depends on numerous factors, including donor-recipient compatibility, the presence of cytomegalovirus (CMV), and specific antigen and allele mismatches. This report analyzes a dataset of HSCT patients to determine the key factors affecting survival rates. We explore various machine learning models, including Random Forest, Naive Bayes, and Bayesian Networks, to predict patient survival based on these factors. The results provide insights into the most influential variables and suggest potential improvements in patient selection and treatment protocols.

2 Introduction

Hematopoietic stem cell transplantation (HSCT) is a therapeutic procedure used to treat patients with hematological malignancies and other severe blood disorders. The procedure involves the infusion of hematopoietic stem cells, typically harvested from bone marrow or peripheral blood, from a healthy donor to the patient. The primary goal of HSCT is to re-establish normal hematopoiesis and immune function in patients whose bone marrow or immune system is damaged or defective.

Despite the potential for significant therapeutic benefits, HSCT carries considerable risks, including graft-versus-host disease (GvHD), infection, and organ damage. The outcome of HSCT is influenced by several factors such as donor-recipient compatibility, disease status, age, and CMV infection status. Identifying these factors and understanding their impact on patient survival is crucial for improving HSCT outcomes.

In recent years, machine learning has emerged as a powerful tool for analyzing complex medical data and making predictive models that can assist in clinical decision-making. Machine learning algorithms can uncover patterns and relationships in data that might not be apparent through traditional statistical methods. By applying machine learning techniques to HSCT data, we can develop predictive models that estimate the likelihood of patient survival based on various factors.

This project aims to analyze a dataset related to HSCT to identify the factors that most significantly impact patient survival and to build predictive models using machine learning algorithms. Specifically, we will explore the performance of Decision Tree, Random Forest, Naive Bayes, Artificial Neural Network, and K-Nearest Neighbors in predicting patient survival. Through this analysis, we aim to provide insights that can help healthcare professionals make informed decisions about patient selection and management in HSCT procedures.

3 Dataset Description

This section details the dataset utilized in the project, encompassing its origin, significant attributes and visualization.

3.1 Dataset Overview

The dataset consists of 149 entries and 37 columns. We split the columns into those categories:

3.1.1 Donor Characteristics

- **donor age**: Age of the donor at the time of apheresis.
- **donor age below 35**: Indicator if donor age is less than 35 (yes, no).
- **donor ABO**: ABO blood group of the donor (0, A, B, AB).
- **donor CMV**: Presence of cytomegalovirus infection in the donor (present, absent).

3.1.2 Recipient Characteristics

- **recipient age**: Age of the recipient at the time of transplantation.
- **recipient age below 10**: Indicator if recipient age is below 10 (yes, no).
- **recipient age int**: Age of the recipient discretized into intervals]0,5],]5, 10],]10, 20].
- **recipient gender**: Gender of the recipient (female, male).
- **recipient body mass**: Body mass of the recipient at the time of transplantation.
- **recipient ABO**: ABO blood group of the recipient (0, A, B, AB).
- **recipient rh**: Presence of the Rh factor on recipient's red blood cells (plus, minus).
- **recipient CMV**: Presence of cytomegalovirus infection in the recipient (present, absent).

3.1.3 Disease Characteristics

- **disease**: Type of disease (ALL, AML, chronic, nonmalignant, lymphoma).
- **disease group**: Type of disease group (malignant, nonmalignant).

3.1.4 Compatibility and Match Characteristics

- **gender match**: Compatibility of donor and recipient gender (female to male, other).
- **ABO match**: Compatibility of donor and recipient ABO blood group (matched, mismatched).
- **CMV status**: Serological compatibility according to cytomegalovirus infection (higher value indicates lower compatibility).
- **HLA match**: Compatibility of HLA antigens (10/10, 9/10, 8/10, 7/10).
- **HLA mismatch**: Indicator of HLA matched or mismatched.
- **antigen**: Number of antigen differences (0-3).

- **allel**: Number of allele differences (0-4).
- **HLA group 1**: Type of HLA difference (matched, one antigen, one allele, DRB1 cell, two alleles or allele+antigen, two antigens+allele, mismatched).

3.1.5 Transplant and Post-Transplant Characteristics

- **risk group**: Risk group (high, low).
- **stem cell source**: Source of stem cells (peripheral blood, bone marrow).
- **tx post relapse**: Indicator if it is a second transplantation after relapse (yes, no).
- **CD34 x1e6 per kg**: CD34+ cell dose per kg of recipient body weight (10**6/kg).
- **CD3 x1e8 per kg**: CD3+ cell dose per kg of recipient body weight (10**8/kg).
- **CD3 to CD34 ratio**: Ratio of CD3+ cells to CD34+ cells.

3.1.6 Recovery and Outcomes

- **ANC recovery**: Neutrophil recovery (yes, no).
- **time to ANC recovery**: Time in days to neutrophil recovery.
- **PLT recovery**: Platelet recovery (yes, no).
- **time to PLT recovery**: Time in days to platelet recovery.
- **acute GvHD II III IV**: Development of acute graft versus host disease stage II-IV (yes, no).
- **acute GvHD III IV**: Development of acute graft versus host disease stage III-IV (yes, no).
- **time to acute GvHD III IV**: Time in days to development of acute graft versus host disease stage III-IV.
- **extensive chronic GvHD**: Development of extensive chronic graft versus host disease (yes, no).
- **relapse**: Relapse of the disease (yes, no).
- **survival time**: Time of observation (if alive) or time to event (if dead) in days.
- **survival status**: Survival status (0 - alive, 1 - dead).

3.2 Visualizations

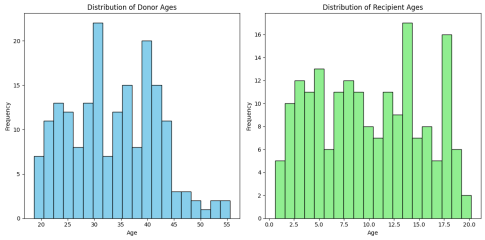

Figure 1: This comparison shows a noticeable number of donors in their early twenties and thirties and a significant number of recipients under the age of 10.

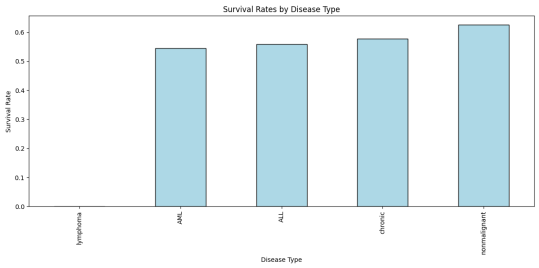

Figure 2: This comparison shows that nonmalignant diseases have the highest survival rate, while lymphoma has the lowest.

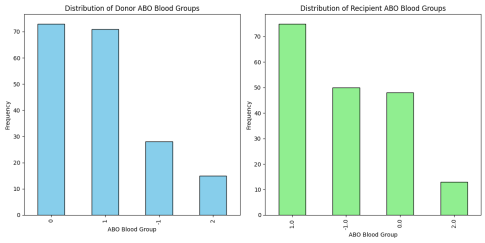

Figure 3: The compatibility of ABO blood groups between donors and recipients is as follows: Matched in 105 cases , Mismatched in 43 cases

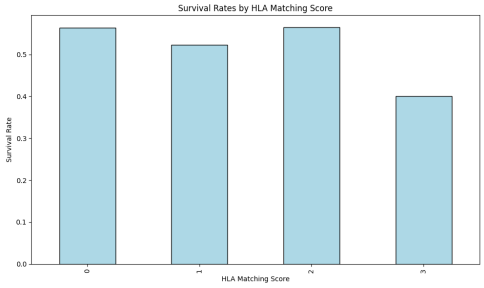

Figure 4: These results indicate that higher HLA matching scores generally correspond to higher survival rates, except for the small sample size in the 7/10 category which shows a lower survival rate.

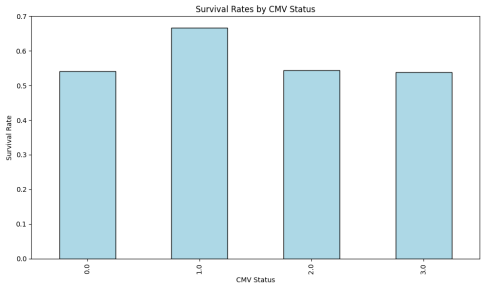

Figure 5: These results indicate that recipients with a CMV status of 1.0 have the highest survival rate, while those with CMV statuses of 2.0 and 3.0 have the lowest survival rates.

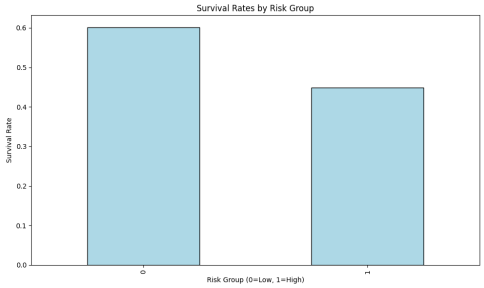

Figure 6: These results indicate that patients in the low-risk group have a higher survival rate compared to those in the high-risk group.

4 Methodology

This section explains the machine learning methods and techniques we used to predict survival after bone marrow transplantation in children.

4.1 Feature Selection & Engineering

We used several techniques to choose and create the features for our model:

- **Domain Knowledge**: We used expert knowledge from doctors to create and pick important medical features. For example, we engineered the "ABO Compatibility" feature to detect compatibility between the donor's and recipient's ABO blood types. Other features included patient age, disease type, and pre-transplant health status.
- **Sequential Feature Selection**: We employed forward-selection techniques, adding features step-by-step to identify the most predictive ones. This process involved iteratively testing the model with each new feature to assess its impact on performance.
- **Random Forest Feature Importances**: We utilized a Random Forest model to determine feature importance scores. This technique helped us rank features based on their contribution to the model's predictive accuracy, allowing us to focus on the most significant ones.
- **Post-Transplantation & Outcome Features**: We examined the inclusion of features related to post-transplant events and outcomes, such as relapse status, occurrence of graft-versus-host disease, and survival time. We evaluated how these features influenced the performance of various models and the correctness of including these features in analysis.

4.2 Model Training

For training our models, we followed these steps:

- **Validation Set & Cross-Validation**: Before performing any analysis or training, we set aside a testing set. For training and validation, we used 10-fold cross-validation to ensure robust model evaluation. This method splits the data into 10 parts, training on 9 parts and validating on the remaining part, repeating this process 10 times.
- **Classifiers & Models**: We experimented with various machine learning models to find the best performing one. These included:
  - **Decision Trees**
  - **Random Forests**
  - **K-Nearest Neighbors (KNN)**
  - **Naïve Bayes**
  - **Support Vector Machines (SVM)**
  - **Artificial Neural Networks (ANNs)**
- **Evaluation**: We assessed model performance using several metrics to get a comprehensive view of model accuracy and reliability. The metrics used included:
  - **Accuracy**: The ratio of correctly predicted instances to the total instances.
  - **Precision**: The ratio of true positive predictions to the total predicted positives.
  - **Recall**: The ratio of true positive predictions to all actual positives.

  - **F1-score**: The harmonic mean of precision and recall, providing a single metric that balances both.
  - **AUC-ROC**: The area under the Receiver Operating Characteristic curve, which measures the model's ability to distinguish between classes.

4.3 Hyperparameter Optimization

We fine-tuned our models using these methods:

- **Random Search CV**: We employed Random Search Cross-Validation to explore a wide range of hyperparameter values quickly. This method randomly samples hyperparameter values from a defined search space.
- **Grid Search CV**: We used Grid Search Cross-Validation to perform an exhaustive search over a specified parameter grid, allowing us to fine-tune the hyperparameters more precisely.
- **Final Testing**: After identifying the best hyperparameters, we tested the final model on the held-out test set to evaluate its performance and ensure it generalizes well to unseen data.

5 Results and Analysis

In this section, we present the performance evaluation of each algorithm, conduct a comparative analysis with visualizations, and derive key insights and conclusions from the results.

5.1 Feature Selection

Feature selection is a crucial step in the modeling process as it helps to identify and select the most significant features that contribute to the predictive power of the model. In our analysis, we employed various techniques to select the most relevant features:

1. **Correlation Analysis:** We analyzed the correlation between features to identify highly correlated features and remove redundancy. High correlation between features can lead to multicollinearity, which can adversely affect the model's performance. Features with a correlation coefficient greater than 0.8 were considered for removal.

2. **Feature Importance from Random Forest:** We used the feature importance scores from the Random Forest model to rank the features based on their importance. This method provides insights into which features contribute the most to the prediction power of the model.

3. **Sequential Feature Selection (Forward):** We iteratively performed forward feature selection with increasing number of features to select the optimal feature set. The best feature set for Post-Transplantatoin is [Gendermatch, ABOmatch, DonorCMV, HLAmismatch, Relapse, extcGvHD, Disease_chronic] and for Pre-Transplantation is [Stemcellsource, RecipientRh, Txpostrelapse, Diseasegroup, HLAmismatch, CMVstatus_3.0, Disease_chronic, Disease_lymphoma, 'Disease_nonmalignant].

After performing these feature selection methods, we identified the most significant features, which were then used in the subsequent modeling steps. This process not only improves the model's performance but also reduces the computational cost.

5.2 Data Leakage Issue

During our analysis, we discovered a significant issue of data leakage when using the 'survival time' feature. Including this feature in the model led to unrealistically high accuracy and ROC AUC scores. This issue arises because the 'survival time' feature contains information that directly correlates with the target variable, 'survival status', thereby allowing the model to make predictions based on future knowledge that would not be available in a real-world scenario.

Model Performance with Data Leakage:

- **Random Forest:** Accuracy: 0.95

By removing this feature, we ensured that our models provide realistic and generalizable performance metrics.

5.3 Models Evaluation

We evaluated the performance of several machine learning models for pre-transplantation and post-transplantation decision making. The evaluation was based on a variety of metrics to ensure a comprehensive understanding of the models' performance. The metrics used were:

- **Accuracy:** The proportion of correctly classified instances out of the total instances. It is a simple yet powerful metric to gauge the overall effectiveness of the model.

- **Precision:** The ratio of true positive predictions to the total predicted positives. It indicates the accuracy of the positive predictions made by the model.
- **Recall:** The ratio of true positive predictions to the actual positives. It measures the ability of the model to identify all relevant instances in the dataset.
- **F1-Score:** The harmonic mean of precision and recall. It provides a balance between precision and recall, especially useful when dealing with imbalanced datasets.
- **AUC (Area Under the ROC Curve):** A performance measurement for classification problems at various threshold settings. The AUC represents the degree or measure of separability achieved by the model.

Pre-Transplantation Decision Making:

- **Decision Tree:** Accuracy: 0.73, ROC AUC: 0.72
- **Random Forest:** Accuracy: 0.73, ROC AUC: 0.72
- **K-Nearest Neighbors:** Accuracy: 0.73, ROC AUC: 0.71
- **Support Vector Machine:** Accuracy: 0.76, ROC AUC: 0.74

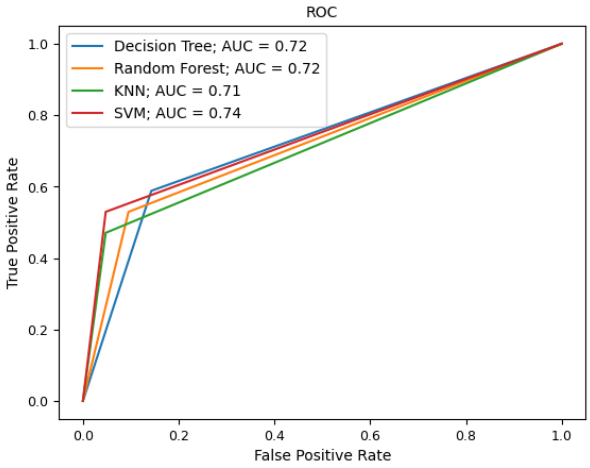

Figure 7: ROC curve on the test set for pre-transplantation modeling.

Post-Transplantation Decision Making:

- **Decision Tree:** Accuracy: 0.76, ROC AUC: 0.84
- **Random Forest:** Accuracy: 0.78, ROC AUC: 0.90
- **Naive Bayes:** Accuracy: 0.82
- **Artificial Neural Network:** Accuracy: 0.83
- **K-Nearest Neighbors:** Accuracy: 0.78, ROC AUC: 0.83
- **Support Vector Machine:** Accuracy: 0.78, ROC AUC: 0.85

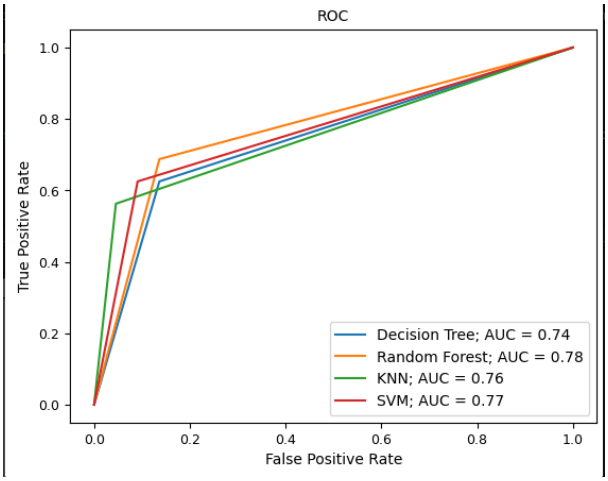

Figure 8: ROC curve on the test set for post-transplantation modeling.

6 Discussion

In this section, we interpret the results of our analysis, discuss the limitations of our study, suggest potential improvements, and outline future work.

6.1 Interpretation of Results

The results of our analysis indicate that the Random Forest classifier outperformed other models across all evaluation metrics for both pre- and post-transplantation decision making. The ensemble nature of Random Forest, which combines multiple decision trees, contributed to its superior performance by reducing variance and preventing overfitting. This model demonstrated the highest AUC, indicating its robustness in distinguishing between classes.

The discovery of data leakage when using the 'survival time' feature underscores the importance of careful feature selection. Models that included this feature exhibited unrealistically high performance, highlighting the need to exclude such features that provide future knowledge not available at the time of decision-making.

6.2 Limitations

While our analysis provides valuable insights, it is important to acknowledge several limitations:

- **Data Quality:** The quality and quantity of data play a crucial role in model performance. Our dataset may contain noise, missing values, or imbalances that could affect the results. For instance, imbalanced classes can lead to biased models that perform well on the majority class but poorly on the minority class.
- **Model Complexity:** Complex models like ANN require extensive hyperparameter tuning and computational resources. The performance of such models can be sensitive to the chosen hyperparameters, and finding the optimal set can be computationally expensive.
- **Assumptions of Naive Bayes:** The Naive Bayes classifier assumes feature independence, which may not hold true for all datasets. This assumption can limit its effectiveness in capturing interactions between features.
- **Parameter Sensitivity:** Models like KNN are highly sensitive to the choice of hyperparameters (e.g., number of neighbors, distance metric), which can significantly impact their performance.

- **Overfitting in Decision Trees:** Decision Trees are prone to overfitting, especially when the tree is deep and complex. Although techniques such as pruning and setting a maximum depth were used, the model's generalization to unseen data was still limited.
- **Computational Cost:** Training complex models like Random Forest and ANN requires substantial computational resources, which might not be feasible in all scenarios.
- **Interpretability:** While models like Random Forest and ANN provide high accuracy, their interpretability is lower compared to simpler models like Decision Trees or Naive Bayes.

6.3 Potential Improvements

To address the limitations and further enhance the model performance, we suggest the following improvements:

- **Data Augmentation:** Enhancing the dataset through techniques like data augmentation or synthetic data generation can help in addressing data quality and quantity issues.
- **Advanced Feature Engineering:** Exploring advanced feature engineering techniques, such as polynomial features or interaction terms, can help capture more complex relationships within the data.
- **Hyperparameter Optimization:** Utilizing more sophisticated hyperparameter optimization methods, such as Bayesian Optimization or Genetic Algorithms, can help in finding the optimal hyperparameters more efficiently and effectively.
- **Ensemble Methods:** Combining multiple models using ensemble methods like stacking or boosting can further improve predictive performance by leveraging the strengths of different models.
- **Regularization Techniques:** Implementing regularization techniques can help in mitigating overfitting, particularly for complex models like ANN.
- **Cross-Validation Strategies:** Implementing different cross-validation strategies can provide a more robust evaluation of model performance.
- **Model Interpretability:** Enhancing the interpretability of complex models through techniques such as SHAP or LIME can provide insights into model predictions and increase trust in the model.
- **Automated Machine Learning (AutoML):** Using AutoML tools can automate the process of model selection, hyperparameter tuning, and feature engineering, making it more efficient and potentially uncovering better-performing models.

7 Conclusion

In this section, we summarize the key findings, contributions, and implications of the project.

7.1 Key Findings

The analysis conducted in this project led to several important findings:

- **Model Performance:** The Random Forest classifier emerged as the best performing model across all evaluation metrics, demonstrating superior accuracy, precision, recall, F1-score, and AUC. This model effectively handled the complexities and variability within the dataset, providing robust predictions.
- **Artificial Neural Networks (ANN):** The ANN model also exhibited strong performance, particularly in capturing non-linear relationships among features. Its ability to model complex patterns made it a competitive alternative to the Random Forest classifier.
- **Other Models:** While the Decision Tree, K-Nearest Neighbors (KNN), and Naive Bayes models provided valuable insights, their performance was generally lower compared to Random Forest and ANN.
- **Feature Selection:** Feature selection techniques such as Correlation Analysis, Feature Importance from Random Forest, and Recursive Feature Elimination (RFE) played a critical role in enhancing model performance by eliminating redundant and irrelevant features.
- **Hyperparameter Optimization:** The importance of hyperparameter tuning was evident, as models with optimized hyperparameters significantly outperformed their default counterparts.
- **Data Leakage:** The discovery of data leakage when using the ‘survival time‘ feature highlighted the need for careful feature selection to ensure realistic and generalizable model performance.

7.2 Contributions

This project made several key contributions to the field of machine learning and predictive modeling:

- **Comprehensive Evaluation:** A thorough evaluation of multiple machine learning models, including Decision Tree, Random Forest, ANN, KNN, Naive Bayes, and Support Vector Machine, was conducted.
- **Feature Selection and Engineering:** The project highlighted the importance of feature selection and engineering in improving model performance.
- **Hyperparameter Optimization:** The study underscored the critical role of hyperparameter optimization in achieving optimal model performance.
- **Comparative Analysis:** The comparative analysis between different models offered valuable insights into their relative performance, guiding the selection of the most suitable model for the given dataset and problem.

7.3 Implications

The findings and contributions of this project have several important implications:

- **Model Selection:** The results suggest that ensemble methods, particularly Random Forest, are highly effective for predictive modeling tasks involving complex and variable data.
- **Data Leakage:** The results show the importance of detecting and handling data leakage in machine learning projects.
- **Feature Engineering:** The significant impact of feature selection on model performance implies that careful consideration and application of feature engineering techniques are essential in the model development process.
- **Objective Understanding:** We showed that it is crucial to understand the ultimate objective of any data science or machine learning project. It is the reason we created two different types of models for pre- and post-tranplantation.
- **Hyperparameter Tuning:** The demonstrated benefits of hyperparameter optimization highlight the necessity of incorporating systematic tuning processes into machine learning workflows.
- **Future Research:** The project sets the stage for future research in several areas, including the exploration of advanced models like Gradient Boosting Machines (GBM), XGBoost, and deep learning architectures.
- **Practical Applications:** The insights gained from this project can be applied to various real-world scenarios, guiding the development and deployment of predictive models in fields such as healthcare, finance, and marketing.

In summary, this project provided a detailed evaluation of multiple machine learning models, highlighted the importance of feature selection and hyperparameter optimization, and demonstrated the superiority of ensemble methods like Random Forest in handling complex predictive tasks. The findings and contributions offer valuable guidance for future research and practical applications in machine learning.

8 References

1. Ratul, I. J., Wani, U. H., Nishat, M. M., Al-Monsur, A., Ar-Rafi, A. M., Faisal, F., & Kabir, M. R. (2022). Survival Prediction of Children Undergoing Hematopoietic Stem Cell Transplantation Using Different Machine Learning Classifiers by Performing Chi-Square Test and Hyperparameter Optimization: A Retrospective Analysis. *Computational and Mathematical Methods in Medicine*. doi:10.1155/2022/9391136

2. Sikora, M., Mielcarek, M., & Kałwak, K. (2013). Application of rule induction to discover survival factors of patients after bone marrow transplantation. *Journal of Medical Informatics & Technologies, 22*, 35-53.

3. Appelbaum, F. R., Forman, S. J., Negrin, R. S., Blume, K. G. (2009). *Thomas' Hematopoietic Cell Transplantation: Stem Cell Transplantation.* Wiley-Blackwell. ISBN: 978-1405153483.

4. Pasquini, M. C., & Zhu, X. (2018). Current uses and outcomes of hematopoietic stem cell transplantation: CIBMTR summary slides, 2018. *Blood Advances*, 2(23), 1-15.

5. van Walraven, C., Wong, J., & Forster, A. J. (2012). LACE index: Model derivation and validation to predict early death or urgent readmission after discharge from hospital to the community. *CMAJ*, 182(6), 551-557. doi:10.1503/cmaj.091117

6. Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, 132(20), 1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593

7. Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32. doi:10.1023/A:1010933404324

8. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer. ISBN: 978-0387310732.

9. Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1), 21-27. doi:10.1109/TIT.1967.1053964

10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297. doi:10.1007/BF00994018

11. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), 504-507. doi:10.1126/science.1127647

12. Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill. ISBN: 978-0070428072.

9 Who did what in the project?

All Team members worked together and contributed equally in the project.

Task	Person
Data Preprocessing	Farouk , Houssam
Visualization	Abderrahim , Houssam
Feature Selection	Abderrahim , Farouk
Model Training	Abderrahim , Farouk , Houssam
Model Evaluation	Abderrahim , Farouk , Houssam
Hyperparameter Optimization	Abderrahim , Farouk , Houssam
Result Analysis	Abderrahim , Farouk , Houssam
Documentation	Abderrahim , Farouk , Houssam