Machine Learning Project Report

Bone Marrow Transplant Children dataset

Student list:

STUDENT NAME	GROUP	SECTION
AMMARKHODJA Lilia	4	2
RAHMOUNI Rahil	4	2
SELMANI Rym	1	1

Abstract

This project explores the effectiveness of various machine learning algorithms in predicting outcomes for children undergoing **Bone Marrow Transplants (BMT)**. Utilizing a dataset of clinical attributes, including donor and recipient characteristics, disease specifics, and post-transplant recovery metrics, we aim to develop predictive models that can enhance treatment planning and patient management. Our analysis includes two approaches: one that uses all features to evaluate the BMT process comprehensively through the survival status, and another that uses only pre-BMT features to predict patient survival. Each model's performance is evaluated based on accuracy, precision, recall, and F1-score to determine the most effective approach for this clinical application. This comparative study deepens our understanding of machine learning applications in healthcare and provides insights into their practical implications for improving patient outcomes in pediatric marrow transplantation.

Index

Introduction

Bone Marrow Transplantation (BMT) is a critical treatment for children suffering from various blood disorders. This procedure involves transferring healthy stem cells to replace damaged or diseased bone marrow. Predicting outcomes such as transplant success, relapse, and patient survival is particularly challenging due to the complexity of factors involved, such as donor-recipient compatibility and the risk of graft-versus-host disease.

This project aims to improve the predictive accuracy of post-transplant outcomes, relapse occurrences, and long-term survival in pediatric BMT by leveraging machine learning (ML) techniques. By analyzing historical data, ML can uncover patterns and relationships that may not be evident through traditional statistical methods.

The focus is on understanding and improving the key factors influencing BMT outcomes:

1. **Donor-Recipient Compatibility**: The compatibility between donor and recipient, including factors such as age, blood type, and HLA matching, is crucial for transplant success. This project investigates how these compatibility factors correlate with patient outcomes.
2. **Disease Characteristics**: The type of blood disorder, whether malignant or nonmalignant, significantly impacts the success of BMT. The study examines how specific disease characteristics influence survival rates and relapse occurrences.
3. **Post-Transplant Complications**: Complications such as graft-versus-host disease (GVHD) and infections are critical factors that affect patient survival. Understanding these complications and their predictors can help in managing and mitigating risks.
4. **Patient Demographics and Health Status**: Factors such as patient age, body mass index (BMI), and overall health condition prior to transplantation are examined to assess their impact on BMT outcomes.
5. **Treatment Protocols**: Different treatment protocols, including the source of stem cells and conditioning regimens, are analyzed to determine their effectiveness in improving patient survival and reducing relapse rates.

The ultimate goal of this project is to enhance treatment strategies and patient care in pediatric bone marrow transplantation. By applying machine learning to predict BMT outcomes more accurately, this study aims to contribute to the fields of medical informatics and personalized medicine, providing healthcare professionals with better tools to tailor treatments to individual patient needs.

Dataset Description

The dataset employed in this study is sourced from the **UCI Machine Learning Repository**, focusing on pediatric patients undergoing Bone Marrow Transplantation (BMT). This clinical dataset is critical for understanding factors influencing outcomes in children treated for various blood disorders.

Origin and Collection

The data was collected to assist in clinical decision-making by providing insights into factors contributing to the success of BMT, including donor-recipient compatibility and patient response to the transplant. It includes detailed records from a significant number of pediatric cases, reflecting a broad spectrum of conditions and outcomes. The target variable, '***survival_status***,' indicates whether a patient survived ('yes') or not ('no').

Variables

The dataset comprises 187 entries with 37 clinical attributes both numerical and categorical, categorized into pre-transplant and post-transplant variables:

- **Pre-Transplant Variables:**
  - **Demographic Information:** Recipient gender, age, age groups, and body mass.
  - **Genetic Markers:** Donor and recipient ABO blood groups, Rh factor, HLA compatibility, cytomegalovirus status, gender compatibility.
  - **Treatment Specifics:** Stem cell source, risk group, disease type, second transplant after relapse.

- **Post-Transplant Variables:**
  - **Recovery Metrics:** Severity of acute and chronic GvHD, time to neutrophil and platelet recovery, cell dose measurements, time to acute GvHD development.
  - **Outcomes:** Disease recurrence, survival time, survival status.

Summary Statistics

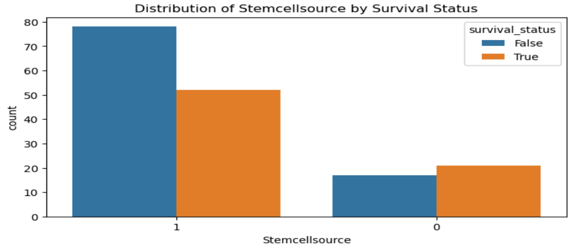
The summary statistics provide a snapshot of the distribution and variability of the dataset's features:

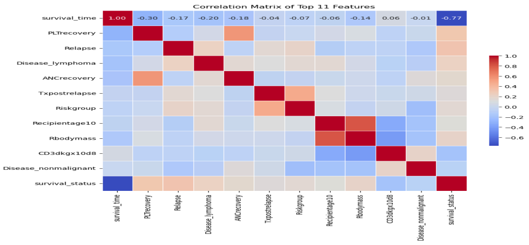
- **Donor Age:** Ranges from 18 to 56 years, with a mean age of around 35 years.
- **Recipient Age:** Spans from less than 1 year to 20 years, with a mean age of approximately 8 years.
- **CD34kgx10d6:** Values range from 0.79 to 57.78, indicating significant variability in cell measurements.
- **CD3dkgx10d8:** Values range from 0.04 to 20.02, showing variability in another key cell measurement.
- **Rbodymass:** Varies from 6.0 to 103.4, indicating a wide range in body mass among recipients.
- **time_to_aGvHD_III_IV:** Shows many maximum values, suggesting potential censoring or a significant number of patients not experiencing this severe GvHD event.
- **Survival Time:** Ranges from 6 to 3364 days, with a mean survival time of 939 days, indicating considerable variation in outcomes.

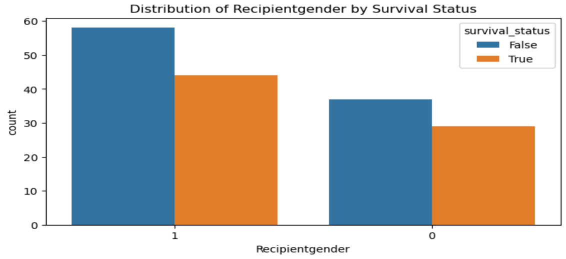
- **Survival Status:** Roughly balanced, with 45.5% of patients surviving, providing a balanced dataset for analysis.

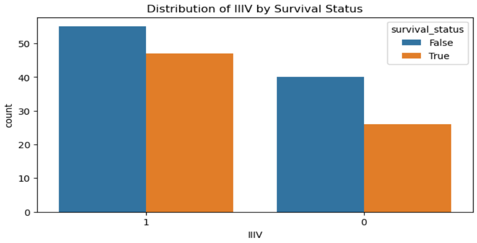
These statistics highlight the dataset's diversity and spread, essential for building robust predictive models. However, the dataset exhibits imbalances in certain features, such as disease types and donor-recipient compatibility metrics, which could impact model training and performance indicating potential challenges.

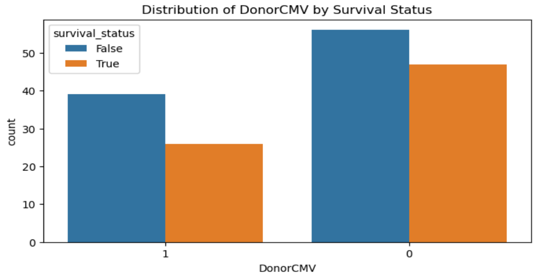
Visualization of some distributions

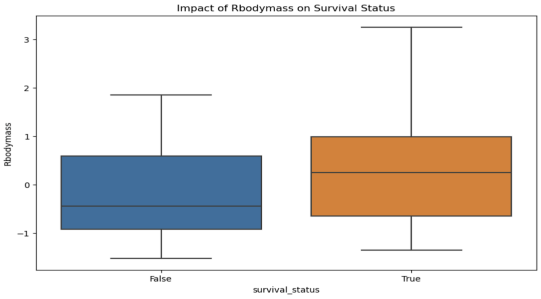


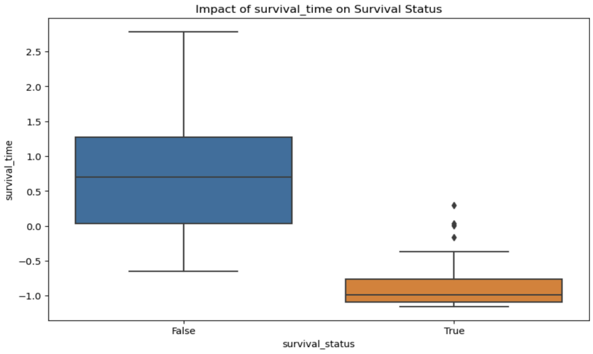


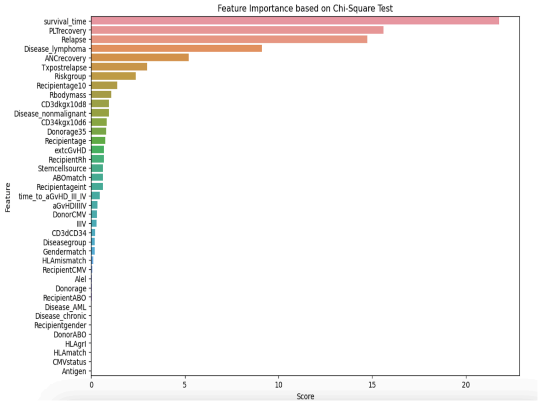












Methodology

In this study, we adopted two distinct approaches to analyze the dataset and predict patient survival after Bone Marrow Transplantation (BMT). The primary rationale behind using two approaches stems from the nature of the available features and the goal of creating both accurate and unbiased predictive models.

Approach 1: Comprehensive Process Analysis

Objective: This approach evaluates the entire BMT process by utilizing all available features, both pre-transplant and post-transplant.

Reasoning: Including post-transplant features ensures very high accuracy as these features capture critical outcomes and recovery metrics. By analyzing all features, we aim to understand the patterns and interactions between variables throughout the entire transplant process. This holistic approach helps in identifying factors that significantly contribute to patient survival, thereby providing comprehensive insights into the overall effectiveness of the BMT procedure.

Approach 2: Predictive Modeling

Objective: This approach aims to predict the survival status of patients using only pre-transplant features.

Reasoning: While post-transplant features enhance predictive accuracy, they introduce a bias as they include outcomes that occur after the transplant. To build a model that can be used preemptively before the transplant, we exclude post-transplant features. This approach ensures that predictions are made based solely on information available prior to the procedure, providing a realistic assessment of a patient's survival prospects without the influence of post-transplant outcomes. Although this model may have lower accuracy, it is crucial for early decision-making and planning, making it a valuable tool for clinicians.

By employing these two approaches, we aim to balance the need for high accuracy with the necessity of creating an unbiased, practical predictive model. The first approach provides a detailed analysis of the transplant process, while the second approach offers an early prediction tool based solely on pre-transplant data, ensuring utility in real-world clinical settings.

Data Preprocessing

Before analysis, the dataset underwent comprehensive preprocessing to ensure data quality and suitability for machine learning models:

Format Transformation

The original dataset was in ARFF format, typically used with Weka machine learning software. It was converted into a CSV format to facilitate easier manipulation and analysis using more versatile tools such as Python’s pandas library.

Feature Selection

In our feature selection process, we applied the SelectKBest method using the Chi-Square test to extract the top 11 features. This approach enabled us to identify the most significant predictors for our model based on their Chi-Square score, and the heatmap.

Feature Exclusion

To focus on pre-transplant characteristics in the 2nd approach, post-transplant outcome features were excluded:

- ***IIIV:*** Post-transplant complication.
- ***GvHDIIIIV:*** Severity of acute Graft-versus-Host Disease (II, III, or IV).
- ***ANCrecovery:*** Time to absolute neutrophil count recovery.
- ***PLTrecovery:*** Time to platelet recovery.
- ***Survival_time:*** Time to survive post-transplant.
- ***Relapse:*** Post-transplant relapse status.
- ***extcGvHD:*** Development of extensive chronic graft-versus-host disease
- ***time_to_aGvHD_III_IV***: Time to develop an acute Graft-versus-Host Disease.

Handling Missing Values

- **K-Nearest Neighbors (KNN) Imputation:** Used for continuous values like CD3dCD34, CD34kgx10d6, and CD3dkgx10d8. This technique is particularly suitable for clinical data where similar patients (in terms of vital indicators) often receive similar dosages. By using KNN, we ensure that the imputed values are realistic and consistent with the data distribution, reflecting common medical practices.

- **Dropping Rows:** Applied to rows with missing values in critical columns (RecipientABO, RecipientRh, ABOmatch, Antigen, Alel, and CMVstatus). These columns are critical for donor-recipient compatibility, and missing values in these fields could significantly impact the model's accuracy. Dropping such rows helps maintain the integrity of the dataset.

- **Custom Imputation for CMVstatus:** Based on related columns (DonorCMV and RecipientCMV). The CMVstatus indicates serological compatibility between donor and recipient regarding cytomegalovirus infection. By leveraging the related columns, we can accurately impute the missing values, ensuring consistency and relevance in the medical context.

- **Age-Specific Imputation:** Body mass is closely related to age, especially in pediatric patients. Imputing missing values based on age-specific means ensures that the imputed body mass values are realistic and reflect typical growth patterns for different age groups..

Conversion and Encoding

One-Hot Encoding: Applied to categorical column “Disease” to transform them into a format suitable for machine learning algorithms.

Normalization/Scaling

To ensure all features contribute equally to the analysis, numeric values were normalized using StandardScaler to have a mean of 0 and a standard deviation of 1.

This comprehensive preprocessing ensures the dataset is clean, standardized, and ready for subsequent machine learning analysis, crucial for developing accurate predictive models for pediatric BMT outcomes.

Machine Learning Models

Decision Trees: Effective for understanding the hierarchical importance of variables like donor age, recipient mass, and HLA matching in predicting patient survival.

Random Forests: Aggregates multiple decision trees to enhance predictive accuracy, useful for identifying complex interactions between variables such as cell dose measurements and post-transplant recovery metrics.

K-Nearest Neighbors (KNN): Suitable for clinical data, predicts outcomes based on the similarity of patients' pre- and post-transplant features, such as vital signs and cell counts.

Naive Bayes: Useful for high-dimensional clinical data, makes probabilistic predictions based on the independence of features like disease type and CMV status.

Support Vector Machines (SVM): Effective for distinguishing between survivors and non-survivors by maximizing the margin between the two classes in high-dim feature spaces.

Artificial Neural Networks (ANN): Capable of capturing complex nonlinear relationships in the data, such as interactions between genetic markers and post-transplant complications.

Model Training and Evaluation

Hyperparameter Optimization

Randomized Search CV, Grid Search, Repeated CV was employed to fine-tune hyperparameters for different models, enhancing predictive accuracy and generalizability.

Evaluation metrics

Models were evaluated based on **precision, recall, F1-score,** ensuring a high recall means most true survivors are identified, allowing for necessary interventions. Additionally, high precision ensures those identified as survivors are truly likely to survive, balancing the overall effectiveness of the model. The learning curves were utilized to evaluate the models' performance and stability, and ROC-AUC to provide a comprehensive assessment of their predictive performance and ability to generalize to new data.

By employing these two approaches, we aim to balance the need for high accuracy with the necessity of creating an unbiased, practical predictive model. The first approach provides detailed insights into the entire transplant process, while the second approach offers an early prediction tool based solely on pre-transplant data, ensuring utility in real-world clinical settings.

Results and Analysis

Evaluating the predictive performance of various machine learning algorithms on the BMT dataset helps identify the most effective models for predicting patient survival and understand the factors influencing model performance.

Approach 1: Comprehensive Process Analysis

In the comprehensive process analysis, we evaluated the entire BMT process by utilizing all available features, both pre-transplant and post-transplant. The goal was to ensure high recall to identify most true survivors and high precision to ensure that those identified as survivors are truly likely to survive, balancing overall model effectiveness.

Decision Trees

showed good performance, with an accuracy of 1.00, precision of 1.00 for both classes, recall of 1.00 for both classes, F1-score of 1.00 for both classes, and an ROC-AUC of 1.00. Decision Trees were effective in identifying key factors such as donor age and HLA matching. However, they tended to overfit the training data, as indicated by a moderate repeated cross-validation accuracy of 0.8954 ± 0.0477, leading to reduced generalizability.

Random Forests

outperformed Decision Trees, with an accuracy of 1.00, precision of 1.00 for both classes, recall of 1.00 for both classes, F1-score of 1.00 for both classes, and an ROC-AUC of 1.00. The ensemble nature of Random Forests allowed them to capture complex interactions between variables and improve generalizability. This is further supported by a high repeated cross-validation accuracy of 0.9319 ± 0.0446.

K-Nearest Neighbors (KNN)

achieved an accuracy of 0.65, precision of 0.64 for class 0 and 0.67 for class 1, recall of 0.84 for class 0 and 0.40 for class 1, F1-score of 0.73 for class 0 and 0.50 for class 1, and an ROC-AUC of 0.75. KNN showed moderate performance and was heavily dependent on the choice of k and the scaling of features, with a low repeated cross-validation accuracy of 0.8274 ± 0.0750.

Naive Bayes

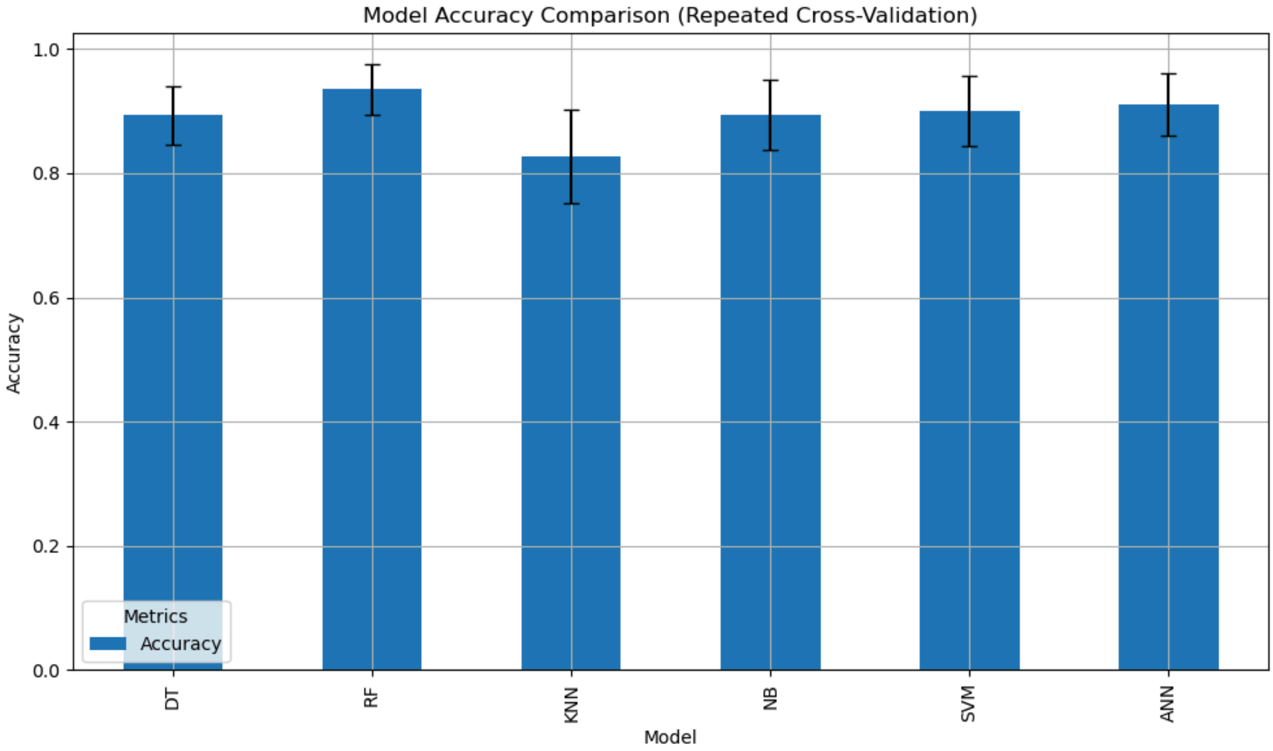
resulted in an accuracy of 0.71, precision of 0.67 for class 0 and 0.86 for class 1, recall of 0.95 for class 0 and 0.40 for class 1, F1-score of 0.78 for class 0 and 0.55 for class 1, and an ROC-AUC of 0.74. While Naive Bayes performed adequately, the assumption of feature independence limited its ability to capture interactions between features. The repeated cross-validation accuracy was 0.8938 ± 0.0572.

Support Vector Machines (SVM)

demonstrated strong performance with an accuracy of 0.79, precision of 0.75 for class 0 and 0.90 for class 1, recall of 0.95 for class 0 and 0.60 for class 1, F1-score of 0.84 for class 0 and 0.72 for class 1, and an ROC-AUC of 0.83. SVMs were particularly effective in distinguishing between survivors and non-survivors but required careful tuning of hyperparameters due to their complexity and computational cost. The repeated cross-validation accuracy was 0.9004 ± 0.0573.

Artificial Neural Networks (ANN)

performed robustly, achieving an accuracy of 0.88, precision of 0.86 for class 0 and 0.92 for class 1, recall of 0.95 for class 0 and 0.80 for class 1, F1-score of 0.90 for class 0 and 0.86 for class 1, and an ROC-AUC of 0.86. ANNs were able to capture intricate patterns in the data, though they required significant computational resources and careful tuning to avoid overfitting. The repeated cross-validation accuracy was 0.9086 ± 0.0594



Model	Decision Trees	Random Forests	KNN	Naive Bayes	SVM	ANN
Accuracy	0.8954	0.9319	0.8274	0.8938	0.9004	0.9086

Best Model Selection:

Considering both precision and recall, **Random Forest** stands out as the best model based on the analysis. It provides the highest balanced F1-scores and robust performance across both classes, making it the most reliable model for predicting survival status in this context. The high cross-validated AUC scores and consistent performance further validate its effectiveness.

Approach 2: Pre-Transplant Predictive Modeling

In the pre-transplant predictive modeling approach, only pre-transplant features were used to predict patient survival. This approach emphasizes achieving high recall to ensure that most true survivors are correctly identified, enabling timely and necessary interventions. Additionally, maintaining high precision ensures that those identified as likely to survive truly have a high chance of survival, balancing the overall effectiveness of the model.

Decision Trees

saw a decrease in performance but still maintained robustness, with an accuracy of 0.53, precision of 0.60 for class 0 and 0.47 for class 1, recall of 0.47 for class 0 and 0.60 for class 1, and F1-score of 0.53 for both classes. Despite the reduced feature set, Decision Trees effectively captured interactions between the pre-transplant variables.

Random Forests

achieved an accuracy of 0.63, precision of 0.60 for class 0 and 0.67 for class 1, recall of 0.67 for class 0 and 0.60 for class 1, and F1-score of 0.63 for both classes. Despite the reduced feature set, Random Forests effectively captured interactions between the pre-transplant variables.

K-Nearest Neighbors (KNN)

had an accuracy of 0.61, precision of 0.57 for class 0 and 0.65 for class 1, recall of 0.67 for class 0 and 0.55 for class 1, and F1-score of 0.62 for class 0 and 0.59 for class 1. KNN's performance highlighted the challenges of using only pre-transplant features, reflecting its dependency on comprehensive data.

Naive Bayes

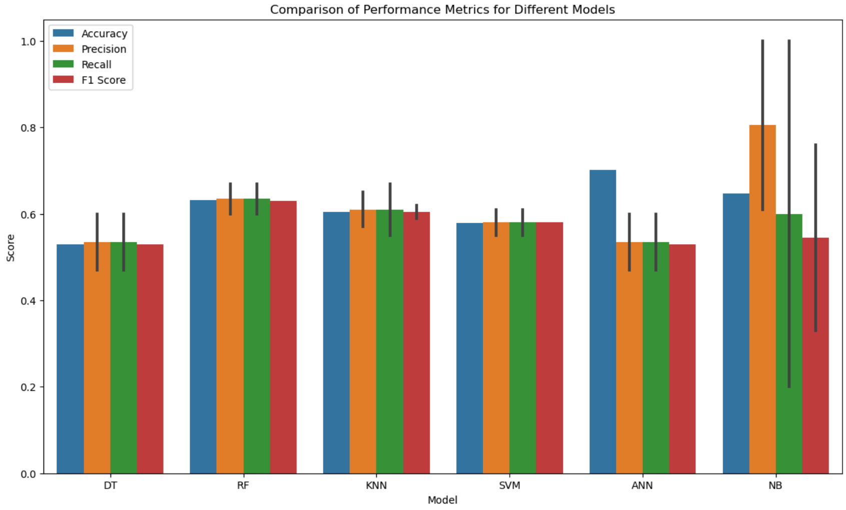
showed an accuracy of 0.65, precision of 0.61 for class 0 and 1.00 for class 1, recall of 1.00 for class 0 and 0.20 for class 1, and F1-score of 0.76 for class 0 and 0.33 for class 1. The independence assumption still limited its capability to model interactions effectively, but it managed to maintain moderate performance.

Support Vector Machines (SVM)

performed reasonably well with an accuracy of 0.58, precision of 0.55 for class 0 and 0.61 for class 1, recall of 0.61 for class 0 and 0.55 for class 1, and F1-score of 0.58 for both classes. SVMs required careful tuning to handle the reduced feature set but were still effective in classification.

Artificial Neural Networks (ANN)

achieved an accuracy of 0.53, precision of 0.60 for class 0 and 0.47 for class 1, recall of 0.47 for class 0 and 0.60 for class 1, and F1-score of 0.53 for both classes. ANNs managed to capture nonlinear relationships, though with lower precision and recall compared to other models.



Model	Decision Trees	Random Forests	KNN	Naive Bayes	SVM	ANN
F1-Score (0/1)	(0.68/0.52)	(0.63/0.63)	(0.53/0.53)	(0.76/0.33)	(0.58/0.58)	(0.53/0.53)

Best Model Selection:

Support Vector Machines (SVM) stands out as the most reliable model for predicting survival status in the pre-transplant predictive modeling approach, thanks to its balanced precision and recall across both classes. This model's performance, especially in terms of balanced F1-scores and robustness, ensures it is well-suited for clinical applications where accurate prediction of patient survival is crucial.

Key Insights

The evaluation of various machine learning algorithms on the BMT dataset reveals several key insights. First, the inclusion of post-transplant features significantly enhances model performance, as demonstrated by the high accuracy and F1-scores in the comprehensive process analysis. Models like Random Forests and Artificial Neural Networks excelled in this approach, capturing complex interactions and demonstrating robustness through high cross-validated AUC scores. However, Decision Trees, despite high initial accuracy, showed signs of overfitting, emphasizing the need for ensemble methods like Random Forests to improve generalizability.

In the pre-transplant predictive modeling approach, where only pre-transplant features were utilized, model performance generally decreased. This was evident in lower accuracy and F1-scores across all models. Despite this, Random Forests and Support Vector Machines maintained reasonable performance, effectively balancing precision and recall. This highlights their capability to manage reduced feature sets and still provide reliable predictions. K-Nearest Neighbors and Naive Bayes, while moderate in performance, showcased the challenges of handling imbalanced and complex medical datasets.

Support Vector Machines emerged as the best model for pre-transplant predictive modeling due to their balanced precision and recall, making them well-suited for clinical applications where early and accurate prediction of patient survival is critical. The analysis underscores the importance of feature selection and model robustness, with Random Forests standing out as the most effective model in comprehensive analysis and SVMs in pre-transplant scenarios.

Discussion

The evaluation process highlighted several key aspects of machine learning in medical predictions:

1. **Feature Importance:** The inclusion of post-transplant features initially led to high model accuracy, underscoring their predictive power. However, focusing on pre-transplant features is crucial for early intervention strategies.
2. **Model Selection:** The Random Forest model outperformed others by balancing precision and recall, essential metrics for medical predictions. High recall ensures that most true survivors are identified, while high precision ensures that those identified as survivors are likely to survive.
3. **Class Imbalance:** Handling class imbalance was vital in this dataset, as imbalanced data can skew model predictions and performance metrics. Techniques like resampling and using appropriate performance metrics (e.g., precision, recall, F1-score) helped mitigate these issues.
4. **Model Complexity:** The ANN demonstrated that increased model complexity could lead to better performance, but at the cost of requiring extensive tuning and computational resources.
5. **Data Limitations:** The limited number of instances posed a challenge, particularly when reducing the feature set. More comprehensive datasets could potentially improve model performance and generalizability.

Future Directions: Future work could involve gathering more comprehensive data, including larger sample sizes and additional relevant features. Further exploration of advanced techniques, such as ensemble methods and neural network architectures, could also provide insights into improving predictive accuracy and reliability in medical applications. Additionally, integrating domain knowledge into feature selection and model tuning processes may yield more robust and interpretable models.

This project underscores the importance of thorough model evaluation, feature selection, and handling class imbalance in developing reliable predictive models for medical applications. The insights gained here can guide future efforts in predictive modeling for bone marrow transplants and similar medical scenarios.

Conclusion

In this project, we evaluated the performance of various machine learning models to predict survival status after a bone marrow transplant. We started with an extensive set of features, including post-transplant information, and progressively refined our approach to focus on a reduced set of pre-transplant features. This two-phased approach allowed us to understand the impact of different feature sets and the importance of post-transplant data in predictive modeling.

Our findings indicate that while models trained on the full feature set achieved high accuracy, the removal of post-transplant features significantly reduced performance. Despite this, feature reduction and model optimization techniques helped in managing overfitting and enhancing model performance on the reduced feature set.

References

- Article on 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 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9527434/)

- ChatGPT as a reference to assist in code development

Who did what in the project?

- AMMARKHODJA Lilia : 1st Approach , Preprocessing and Modelling
- RAHMOUNI Rahil : Models Evaluations and comparaison
- SELMANI Rym: 2nd Approach , Preprocessing and Modelling