**PROJECT DESCRIPTION**

**Supervisor Name :** Dr. Robin Ghosh

**Project Overview:**

This project aims to develop a Machine Learning predictive model for hematopoietic stem cell transplantation (HSCT) survival outcomes. The current models fail to address the concerns related to Socioeconomic factors and racial disparities, which leads to bias in the model performance.

**Overview of the Data:**

The dataset consists of 59 variables related to hematopoietic stem cell transplantation (HSCT), encompassing a range of demographic and medical characteristics of both recipients and donors, such as age, sex, ethnicity, disease status, and treatment details.

We are challenging ourselves to develop advanced predictive models for allogeneic HCT that enhance both accuracy and fairness in survival predictions.

|  |  |  |
| --- | --- | --- |
| **Phase** | **Dates** | **Tasks** |
| **Phase -I** | (Jan 27 – Feb 7) | Defining the Problem and Goal, Data Collection and Exploration of the data. |
| **Phase -II** | (Feb 7 – Feb 21) | Data Preprocessing and EDA |
| **Phase -III** | (Feb 21 – Mar 7) | Feature Engineering and Data Visualization |
| **Phase -IV** | (Mar 7 – Mar 21) | Model Selection and Training |
| **Phase -V** | (Mar 21 – April 4) | Model Performance Evaluation and Model Tuning and Refining Parameters |
| **Phase -VI** | (April 4 – April 18) | Model Deployment and Testing. |
| **Phase -VII** | (April 18 – May 1) | Final Presentation |
|  |  |  |

**PHASE -I OVERVIEW:**

At the start of the project, I had a change in my supervisor and selected a new research direction under Dr. Robin Ghosh. I decided to focus on predicting the survival rate of hematopoietic stem cell transplantation (HSCT) with the goal of developing a more accurate and fairer predictive model compared to existing approaches.

To begin, I explored various datasets as part of my internship, ensuring we considered all relevant factors for a comprehensive data analysis process. We considered the ETL process to clean and structure the data before model training. After evaluating multiple datasets, we selected one that had not been extensively modeled before, providing an opportunity to build a novel and effective machine-learning model. We designed a project plan with goals defined for each period to complete the project before the deadline.

Our objective is to enhance survival prediction accuracy while addressing biases related to socioeconomic factors and racial disparities, which are often overlooked in existing models. By integrating advanced feature engineering, data preprocessing, and rigorous model selection, we aim to develop a robust predictive model for allogeneic HSCT outcomes.