# TEAM INSIGHT - FINAL PROJECT REPORT

## Team Members: Personal Information

Name: Aibo Sun
Address: 2595 Hoover street  Apt #324,Los Angeles,C.A 90007
Email Id: aibosun@usc.edu

Name: Rasika Guru
Address: 1204 W.Adams Blvd, Apt #4, Los Angeles, CA 90007
Email Id: rguru@usc.edu

Name: Roshani Mangalore
Address: 720 W 27th street, Apt #228, Los Angeles, CA, 90007
Email Id: rmangalo@usc.edu

## Project Information

**Project Title:** Prediction Model for Liver Transplantation Surgery
**Date Started:** January 25, 2018
**Date Completed:** April 26,2018
**Project Sponsor/Champion:** University of Southern California - Keck School of Medicine

## Executive Summary: Six Sigma Project

Approximately 14,000 patients are listed and waiting for orthotopic liver transplantation (OLT) but only 7,000 OLTs are performed annually. The number of patients who require liver transplantation surgery are almost twice the number of donors available. There is an obvious need to identify the potential recipient for every donor. Our predictive model not only determines the influential clinical features but also matches every donor with a potential recipient justified by graft survival probability. The end product/deliverable is a user friendly web application which can be used by medical professionals.

**Problem Statement:**

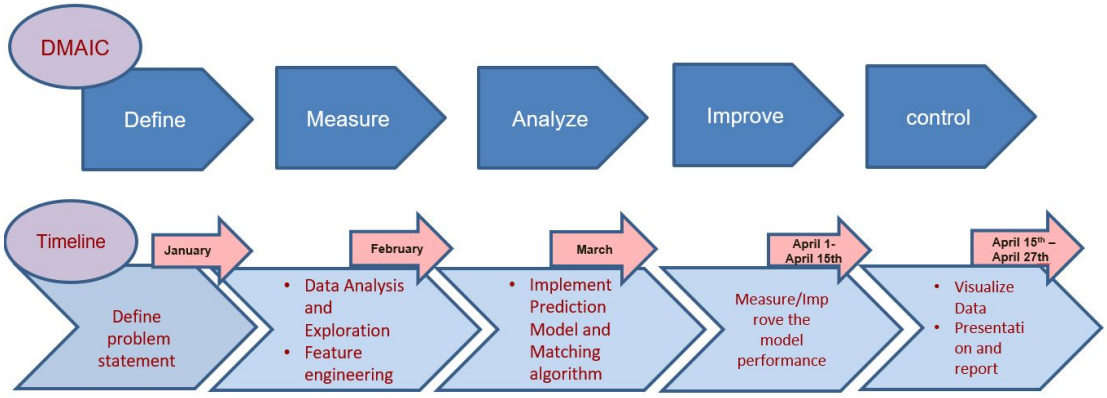
Our focus is to develop a predictive model for liver transplantation which:

- Identifies the key clinical attributes that influences the identification of potential recipient
- Identifies 5-best recipients for every donor based on the matching score between the donor and recipient.
- If two recipients have the same matching score, graft survival probability is used to break the ties.

**Project Scope:**

- Perform preliminary data analysis and data cleaning to eliminate the attributes which do not contribute to the prediction
- Perform feature engineering which includes categorical feature encoding, handling missing values and normalization of Data
- Identify key clinical attributes which influences the Identification of potential liver recipient
- Implement matching algorithm to match every donor with the potential recipient
- Develop a highly accurate model that predicts the graft survival probability of the potential recipient for the Liver Transplantation Surgery.
- Measure the Performance of the Model using different Evaluation Matrices
- Improve the performance of the Model by incorporating Cross Validation and Over sampling techniques
- Develop a user-friendly web application to be used by Medical professionals

**Major project phase (DMAIC):**




**Key Learnings:**

1. Delivering a product to the customers by adopting six sigma principles and techniques.
2. Collaborating with team members and keeping up with timelines.
3. Working with private patient information by incorporating the learnings from HIPPA training

**Test Scenario:**

A user friendly web application with the below features:

**Input**: CSV files for donor and recipient with their clinical information.

**Output**:

1. Top 20 clinical features that influence the identification of potential recipient.

2.The top 5 recipients for every donor can be viewed along with the matching and graft prediction scores of each recipient.

**Conclusions and Project Recommendations:**

A predictive model that matches the donor with recipient justified by the graft survival probability. The top 5 recipients for every donor is available in the preferred MELD range of the recipients. We recommend medical experts to use this application because it can efficiently

match the liver donor with the potential recipient who has higher survival rate. This assists the utilization of scarcely available liver donors and also ensures that the liver is donated to the right recipient.

**Actions Taken:**
The challenge of scarce donors is addressed by our project that identifies the potential recipient for every donor**.** The identified potential recipient has the highest graft survival probability. A web application is also built to help the medical professionals have easy access to the prediction tool.

**Benefits Realized:**
The medical experts can use the web application that is easy to use and user friendly. The top clinical features influencing the predictive analysis and the top 5 recipients for every donor is displayed along with their matching scores and graft survival prediction probability based on which the decision about the recipient can be made.

## Lean Six Sigma Project


## DEFINE PHASE

### Cost of Poor Quality Statement

The Define phase is the first phase of the DMAIC roadmap. It is the base on which DMAIC rests. The first step is to identify the problem statement.
Outcome of liver transplantation surgery depends upon a complex interaction between donor, recipient and process factors. Driven by the disparity between the increasing number of potential transplant recipients and the limited number of suitable organ donors, there is an increasing use of organs of marginal quality. This shift brings into focus the obvious need to identify the potential recipient for every donor. Add to this the significant financial costs and regulatory pressures with each transplant, a quantitative tool which can help the transplant surgeon optimize the decision-making process in identifying the potential recipient for every donor is urgently required.

### Customer Satisfaction (Voice of the Customer)

The ability to match the liver donor with the potential recipient assists the utilization of scarce resource of donor livers, while ensuring that patients who have the higher survival rate are prioritized.

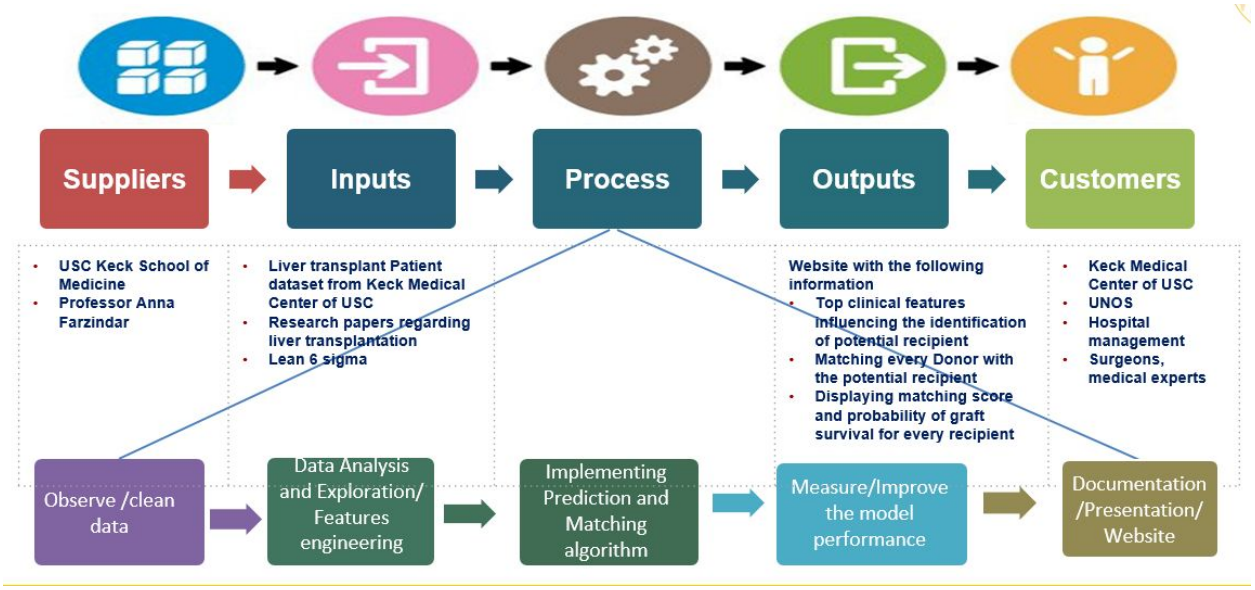
**Tools Application:** Project Charter and SIPOC are the six sigma tools employed during Define phase

## Project Charter:

| USC- INF 560- Data Informatics Professional Practicum | | | | |
|---|---|---|---|---|
| Project Charter | | | | |
| Analysis: | Develop a highly accurate prediction model for liver transplantation surgery | | | |
| Conclusion: | | | | |
| Project Title: | Prediction model for liver transplantation surgery | | | |
| Team Name: | Team Insight | | | |
| Business Case | | Problem/Opportunity Statement | | |
| This project was initiated to address the problem of matching the liver available for a transplantation surgery, (that is scarce in number) to a recepient who has a better survival rate post the surgery. | | To develop a Model that better predicts the potential recipient for a liver transplantation surgery than the existing model. | | |
| Goal Statement | | Scope | | |
| The goal of this project is to identify the key components that help us to identify potential recipients and also develop a model that better predicts the success rate of a transplantation surgery. | | In-scope : Determining the clinical features which influences the accuracy of prediction.<br>Out of scope : Integrating with data systems at multiple institutions | | |
| Team Members | | Project Timeline | | |
| Key Stakeholders | Member Names | Key Milestone | Target Date | Revised Date |
| Team Leader: | Aibo Sun | Start Date: | 30-Jan-2018 | |
| Team Members: | Rasika Guru | Phase B: | 20-Feb-2018 | |
| | Roshani Mangalore | Phase C: | 15-Mar-2018 | |
| | | Phase D: | 30-Mar-2018 | |
| | | Phase E: | 20-April-2018 | |

| Project Budget: | N/A | Project Resources: | Liver transplant Patient dataset from Keck Medical Center of USC, Computers, Data Analysis tools,Human Resources | |
|---|---|---|---|---|
| Constraints, Assumptions, Risks and Dependencies | | | | |
| Constraints | 1. Confidentiality of Patients Data | | | |
| | 2. Time Constraint (To build a better performing Model, It takes time to experiment with the data and come up with an Ideal Model) | | | |
| Assumptions | 1. It is feasible to follow up with patients post surgery to track the survival rate. | | | |
| Risks & Dependencies | Risk : Patient who can better respond to the surgery might lose out a chance on being a potential receipient of a liver transplantation surgery as a result of the poor performance of our Prediction Model | | | |
| | Dependency : We are dependent on the Medical Expert to help us understand the Medical terms | | | |
| Approval | | | | |
| Project Sponsor: | | | Date: | |

**SIPOC:**




**Key Learnings:**

- There is a need to identify all relevant elements for the project
- Map the process in five high level steps.
- Identify the outputs of this process.
- Identify the customers that will receive the outputs of this process.
- Identify the inputs required for the process to function properly.
- Identify the suppliers of the inputs that are required by the process.
- Discuss with project sponsor and other involved stakeholders for verification.

## MEASURE PHASE

**Process Mapping/Process Visualization**

Process maps consist of a sequence of activity steps and also the interactions between individuals or groups. Each participant in the process is displayed on the map - tasks/activity are then articulated in sequence under the column corresponding to that stakeholder. We have defined Common , Detailed and Functional Process maps for this project *[Appendix A]*

**Key learnings from Functional Process Mapping:**

- Displays activity steps in a current process
- Allows you to view department relationships and handoffs
- Helps clarify roles in relation to the flow of events
- Indicates potential areas of delay or rework
- Can be used for improvement or training efforts

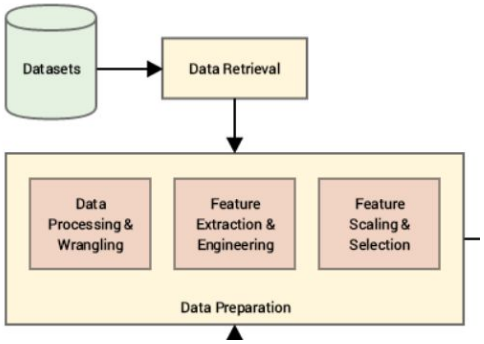
**The Vital few, Data Collection Planning and Execution**

- Dataset used is from United Network for Organ Sharing (UNOS), a tax-exempt, medical, scientific and educational organization which controls the national Organ Procurement

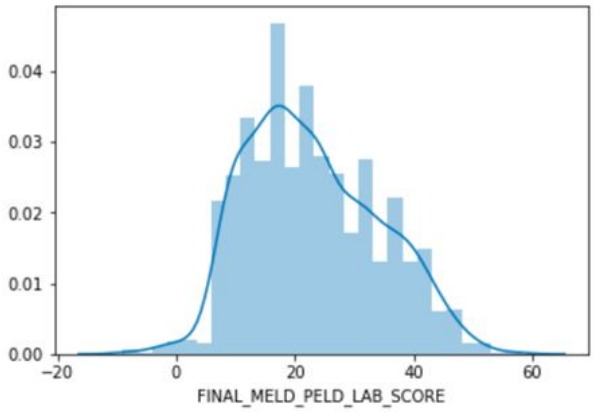
and Transplantation Network (OPTN) under agreement to the Division of Organ Transplantation (DOT) of Department of Health and Human Services (DHHS)

- USC Keck School of Medicine provided us with the multi-organ simulated dataset containing liver patient records since 1st October 1987

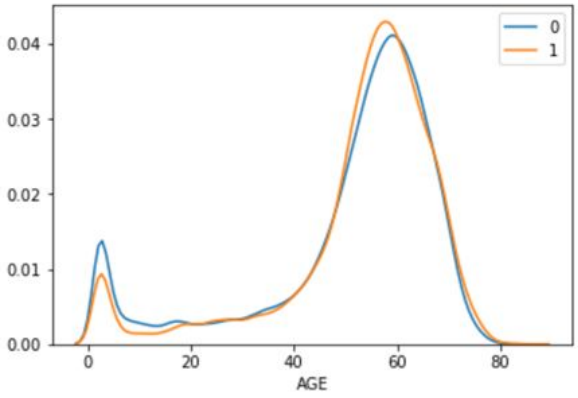
**Data Preprocessing:**



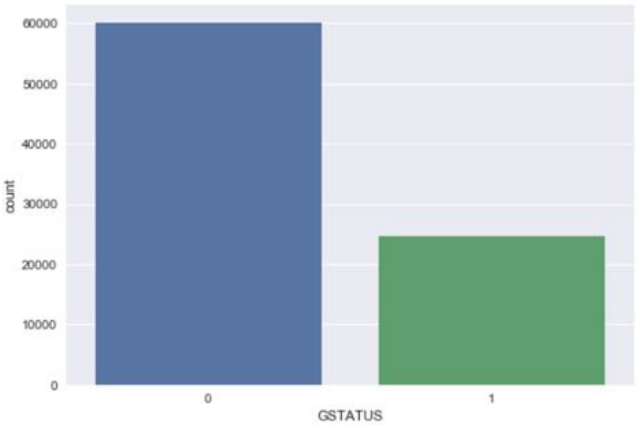

- Perform preliminary data analysis on the available dataset



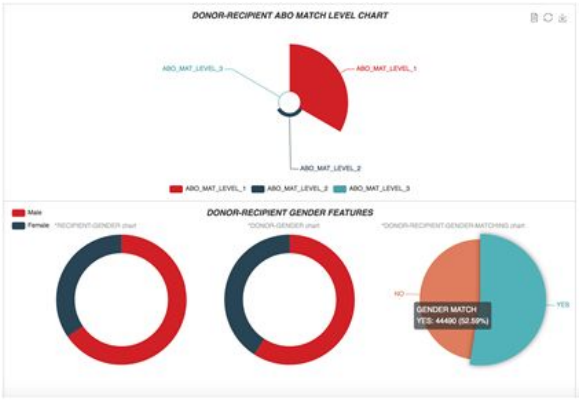

MELD Score distribution



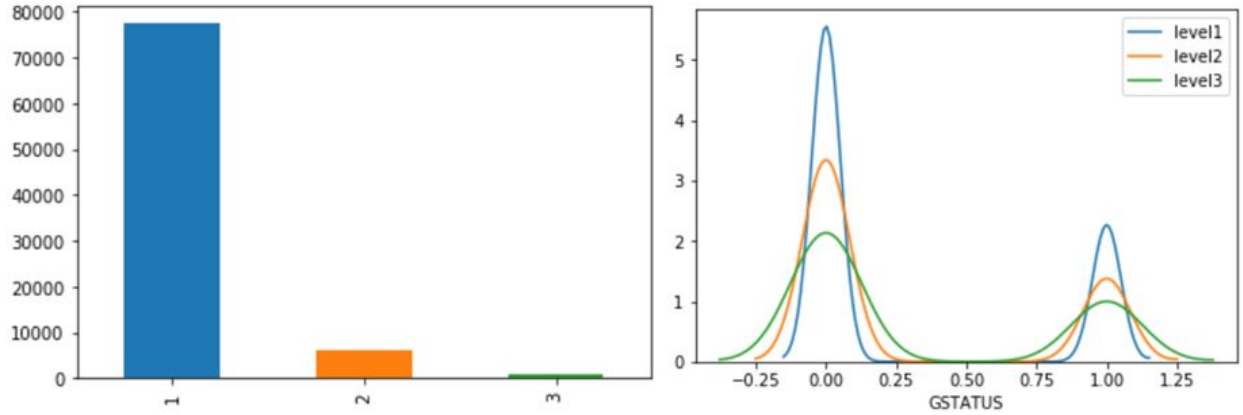

GSTATUS vs AGE




GSTATUS distribution( 0: Success, 1:Failure)




Donor-Recipient Age and gender info




*DONOR-RECIPIENT ABO MATCH LEVEL is directly proportional to GSTATUS*

- Perform data cleaning to eliminate the attributes which do not contribute to the prediction analysis (Ex: Post Transplantation features, State of Residency, Education level of recipients)
- Elimination of records where the age of recipient is below 18 years since this projects aims at dealing only with MELD score
- Elimination of records where PTIME > GTIME marking the outliers or the noisy data
- Handling missing values:
  - ❖ Replace NULL values in categorical columns with the mode

```
imp_cat = CategoricalImputer()
X_cf = pd.DataFrame(imp_cat.fit_transform(np.array(X_cf)),columns = X_cf.columns)
```

  - ❖ Replace NULL values with the mean for non- categorical features

```
imp = Imputer(missing_values='NaN', strategy='mean', axis=0)
imp.fit(X_ncf)
X_ncf = pd.DataFrame(imp.transform(X_ncf),columns = X_ncf.columns)
```

- Perform one-hot encoding for Categorical features as Machine learning models accepts only numerical data
- Perform feature scaling to standardize the range of independent variables or features of data. We have used Min-max scalar from scikit library,

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

```
min_max_scaler = preprocessing.MinMaxScaler()
header=X_ncf.columns
X_train_minmax = min_max_scaler.fit_transform(X_ncf)
X_ncf=pd.DataFrame(X_train_minmax,columns=header)
```

## Measurement System Analysis:

Existing system uses MELD score alone to identify the potential recipient (higher the MELD score, severe is the condition and so the patient with higher MELD score is chosen as a potential recipient for Liver Transplantation surgery).

The utility of the MELD score for predicting three-month mortality among patients awaiting liver transplantation was demonstrated in a study that included 3437 adult liver transplantation

candidates who were listed between 1999 and 2001. Of these, 412 died during the three-month follow-up period. Waiting list mortality was directly proportional to the MELD score at the time of listing, with mortality being 1.9 percent for patients with MELD scores less than 9, and 71 percent for patients with MELD scores ≥40 *[Reference:1]*

Proposed system develops a matching score based on donor-recipient features that augments the existing MELD score in identifying the potential recipient. The 5-best recipients for every donor is listed from which the potential recipient is identified by our survival analysis prediction. The implemented model has 77.2% accuracy with 71.2 auc-roc score.

**Tools Used:** Functional Process Maps for developing process flowcharts and Python Jupyter Notebook for Data Pre-processing

### Evaluation Metrics:

- F1 Score: F1 score (also F-score or F-measure) is a measure of test's accuracy. It considers both the precision and recall to compute the score. The F1 score is the harmonic average of the precision and recall, where an F1 score reaches the best value at 1 and worst at 0.
- AUC ROC Score: The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- Matching Score: The model develops a matching score for a donor and a recipient. This matching score must be highly correlated with the survival rate.

## ANALYSIS PHASE

### Charts used for Analysis:

Cause and Effect Fishbone Diagram - The root causes are listed. ‘Why’ questions are asked to each of the cause till a solution is achieved.

### Formula:

$$Y = f(X)$$

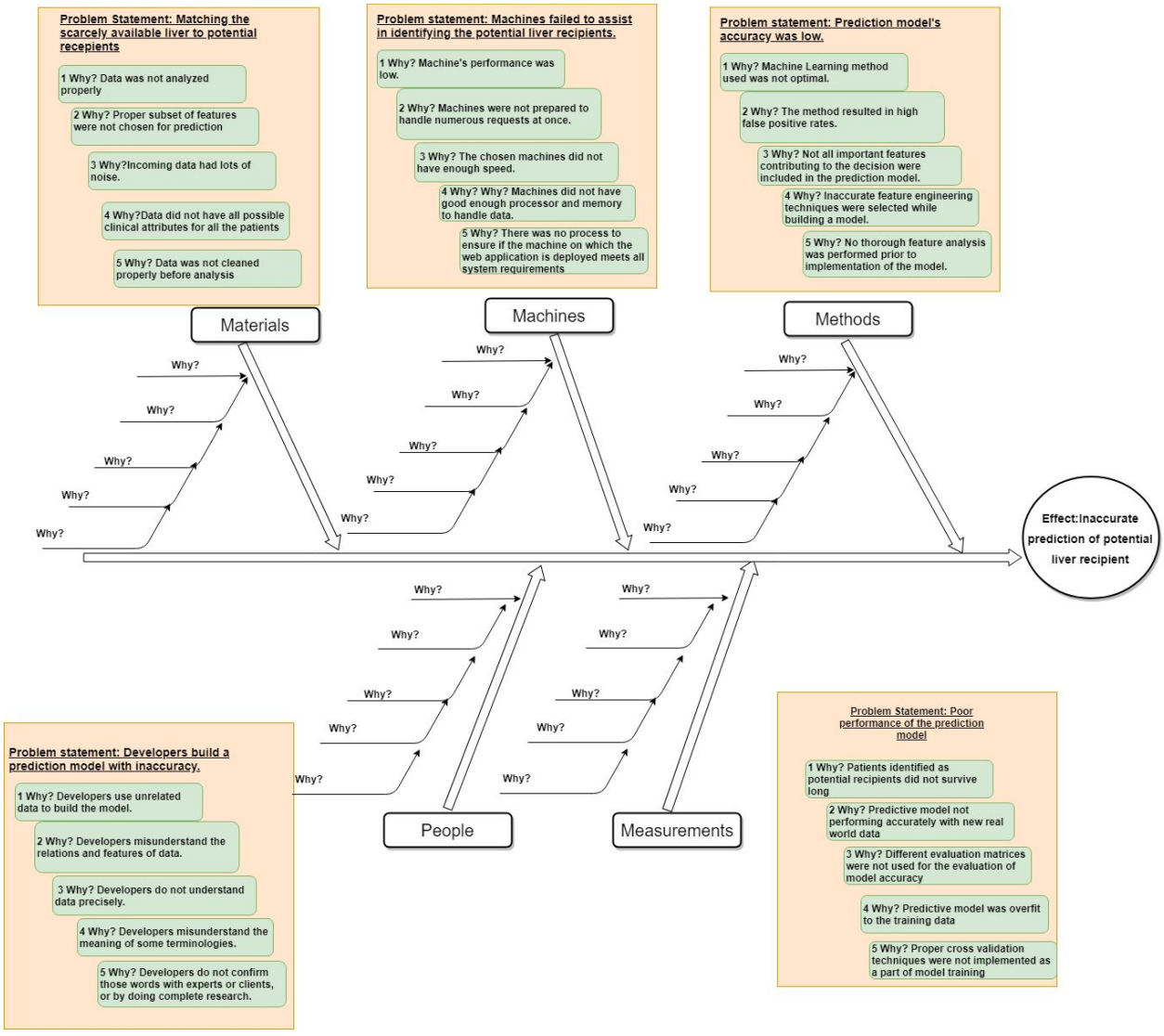
**Output,** Y: Predictive model for liver transplantation surgery that matches every donor with potential recipient.
**Input,** X: Influential clinical features of donors and recipients.
**Function,** f:
1. A machine learning model that is a combination of three algorithms - Logistic Regression, Random forest and Deep Learning. The model is trained on 39162 donor/recipient information. Upon receiving the input, the model predicts the graft survival probability of the recipient. This model performs with an accuracy of 77.2% and has an AUC-ROC of 0.712
2. A matching algorithm that takes the common features between donor and recipient and calculates a matching score that highly correlates with the graft survival probability.

## SWOT Analysis:

1. **Strengths:** A highly accurate predictive model that has an accuracy of 77.2% and AUC-ROC of 0.712. The matching scores of the two classes of recipients (graft status = 0 and graft status = 1) have difference in mean that is significant (p -val = 5.7e-06) suggesting the two scores belong to two different distributions. This score also highly correlates with the graft survival probability.
2. **Weakness:** The model currently runs on CSV Input files. This architecture is not scalable when the input data is voluminous.
3. **Opportunities:** Future scope involves having a database in the backend which can efficiently manage large volume of information. The login system in the UI application can also be integrated with the hospital database to ensure only authorized medical professionals have access.
4. **Threats:** If the user enters wrong data, then our predictive model assumes its right and gives results based on the entered data. Currently there is no inbuilt intelligence that can validate the correctness of the input.

## Root Causes:

Fishbone diagram and 5 Whys



### Potential Solutions:

1.Develop a matching algorithm that identifies the best recipient for every donor.
2.Build a predictive model that predicts the graft survival probability of the recipients.
3.Identify key clinical features/attributes that helps in achieving the above two solutions.

The root causes are correlated with the output, which is an accurate predictive model for the transplantation surgery. Analysing the root causes and fixing the issues had significant effect on the performance of the model.

### Sources of Variations:

Dataset provided doesn't have all the necessary information. There are missing values at random, that are handled by data preprocessing techniques. The imputed values might only be a representation of the actual value, but might not be exactly equal to it. This causes variations in the prediction model.

## IMPROVE PHASE

Improve the performance of the predictive model by adopting some of the below potential solutions:

### Alternative Solutions:

1. Implement regularization, cross validation strategies to avoid overfitting to the given test data.
2. Perform undersampling, oversampling or random sampling to handle a imbalanced dataset. (The input dataset has almost twice the number of patients with functioning graft compared to the ones with a graft failure)
3. Perform feature scaling, standardization or normalization techniques on the input data to avoid any dominance of features with values in a different scale/range.
4. Implement Principal Component Analysis (PCA) algorithm to reduce the dimensions of the dataset. This can reduce the computation time.

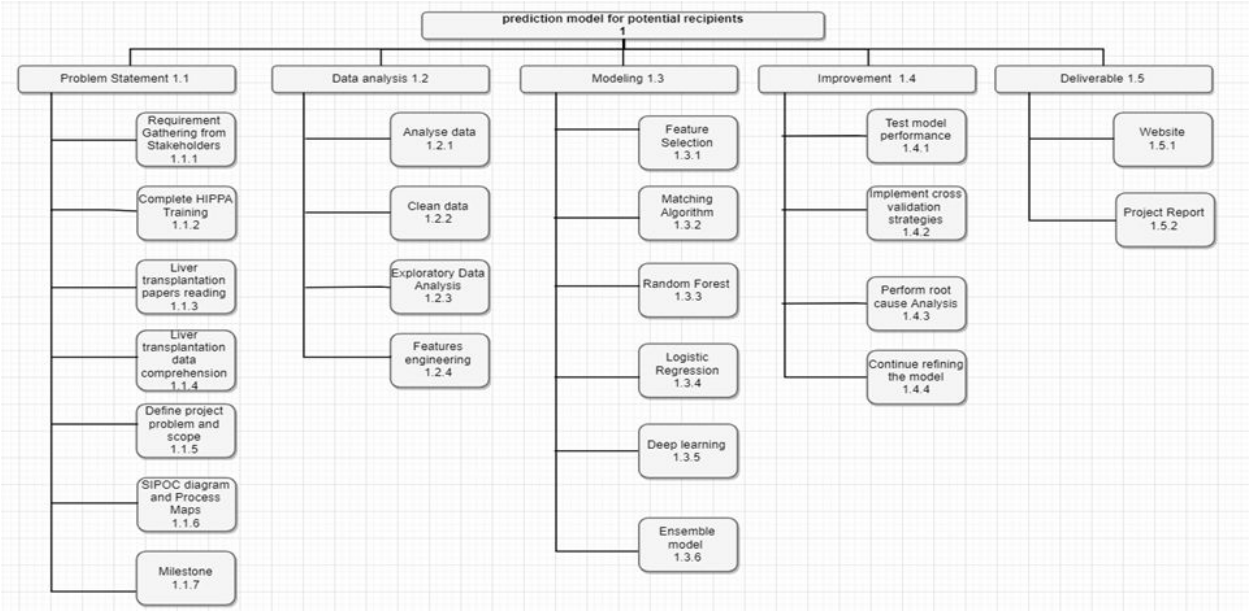
### Recommended Solutions:

1. 10-fold cross validation strategy to avoid overfitting of the model to the input data. Risk: This also increased the computation time required to train the model.
2. Oversampling using Synthetic Minority Over Sampling Technique(SMOTE) to address the imbalance in the dataset. Challenge: It was a challenge to perform SMOTE with cross validation and to ensure no information about test data was leaked to the model during the training phase.
3. Normalization of the input data to avoid any possible bias before training the model. Risk: This process required additional memory since we had to store the actual values of the features as well to cater to some UI requirements.

## Pilot Solution:

1. 10 fold cross validation strategy was implemented on the input data used for training the model which significantly improved the performance.
2. SMOTE was applied to the input dataset which adds simulated records for the minority class. This oversampled data was used for training the model.
3. The input dataset was normalized to ensure all the features had values between 0 and 1 before they were fed to the machine learning models.

## Work Breakdown Structure:




# CONTROL PHASE

Once the idea is implemented, it is to be built into the system as a requirement. This involves putting a system in place to ensure that the performance is controlled as well as measured. The Control phase is the final phase of the Lean Six Sigma. The team focuses on how to sustain the newly achieved improvements by delivering the user manual to the customers.

**Control-Steps:**

Mistake proof the process:

- The web application developed is to be used only by the medical experts. We are ensuring this by creating a login system.
- The user needs to upload the donor and recipient information file in order to view the results. The system throws a warning in case no file is uploaded or an empty file is uploaded.
- The sample format for donor and recipient file is provided. The machine learning model is trained to handle if there is any addition of previously unseen feature in the input file or if any of the expected donor/recipient features are missing in the file.

*Challenge:* The Machine learning models were trained based on certain input features and the model would fail if the test data had different set of features.
*Solution:* We ensured that the test data had same set of features as train data by incorporating the following code

```
def add_missing_dummy_columns(x_test, x_train_columns):
    missing_cols = set(x_train_columns) - set(x_test.columns)
    for c in missing_cols:
        x_test[c] = 0


def fix_columns(x_test, x_train_columns):
    add_missing_dummy_columns(x_test, x_train_columns)
    # make sure we have all the columns we need
    assert(set(x_train_columns) - set(x_test.columns) == set())
    extra_cols = set(x_test.columns) - set(x_train_columns)
    x_test = x_test[x_train_columns]
    return x_test,extra_cols
```

Follow-up actions

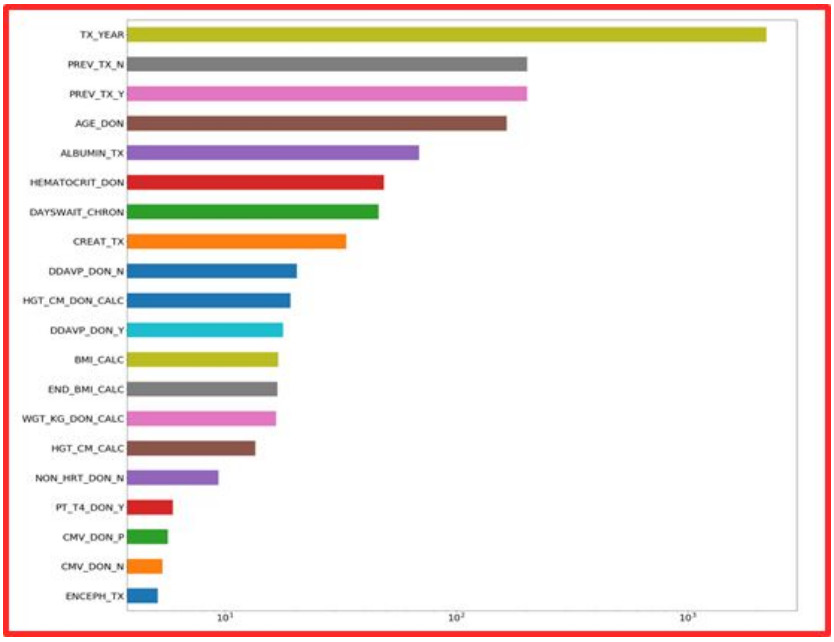
- User manual containing the guidelines on how to use the web application, the input and output expected out of the system is handed over to the end-user
- Project close out includes the delivery of Project Documentation, User Manual and Web Application to the end user

## RESULT AND SYSTEM IMPLEMENTATION

### Part 1: Feature Selection

- Univariate Statistics features selection
    - ○Determines the relationship between each feature and output(target)
    - ○Only the features with highest confidence are selected
    - ○Select Percentile : 75 % of the features
- Random Forest Classifier and Select from Model
    - ○Cross Validation is incorporated to increase the accuracy of the model

*Output : The common top features selected by both the Models*




## Part 2: Matching

| DONOR FEATURES | RECIPIENT FEATURES |
|---|---|
| BMI_DON_CALC | BMI_CALC |
| CREAT_DON | CREAT_TX |
| TBILI_DON | TBILI_TX |
| HGT_CM_DON_CALC | HGT_CM_CALC |
| WGT_KG_DON_CALC | WGT_KG_CALC |
| ABO_DON | ABO |
| GENDER_DON | GENDER |
| HBV_CORE_DON | HBV_CORE |
| HBV_SUR_ANTIGEN_DON | HBV_SUR_ANTIGEN |

- The correlation of the independent common features with the target variable is calculated using Chi square test (for categorical features) and student-t test (for continuous features).
- The metric is then established by incorporating weights to the score corresponding to the correlation.
- The 5- best recipients for every donor is chosen and the potential recipient is then identified by the probability of the graft success rate

$$Score_{r,d} = \sum_{f\ in\ NFs} \sqrt{\{(t_{Rf,gs} \times rf) - (t_{Df,gs} \times df)\}^2} + \sum_{f\ in\ CFs} \left(\chi^2{}_{Rf,gs} + \chi^2{}_{Df,gs}\right)_{norm} \times P(gs = 0 \mid (rf, df))$$

*r – recipient*
*d – donor*
*$Score_{r,d}$ - Matching score for the given donor-recipient pair.*
*gs – Gstatus*
*NFs – Numerical features*
*CFs – Categorical features*
*$rf$ – value of feature f for recipient r*
*$df$ – value of feature f for donor d*
*$t_{Rf,gs}$ - t-statistic between recipient feature Rf and Gstatus from dataset*
*$t_{Df,gs}$ - t-statistic between donor feature Df and Gstatus from dataset*
*$\chi^2{}_{Rf,gs}$ - Chi sq. statistic between recipient feature Rf and Gstatus from data*
*$\chi^2{}_{Df,gs}$ - Chi sq. statistic between donor feature Df and Gstatus from data*
*$P(gs = 0 \mid (rf, df))$ - Probability of Gstatus 0 given feature value pair rf-df, calculated from dataset*

## Part 3: Survival Prediction

### 1. Random Forest:

- Random Forest predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical).
- The data is split into training and test set with a ratio of 70:30
- The target variable is GSTATUS. The input vector to the learning model is of dimensions (39162, 237).
- The output is a binary class - either 0 or 1.
- Incorporated Over Sampling and Cross Validation techniques
- Model Performance:

```
precision score: 0.438
recall score: 0.209
f1 score: 0.283
auc roc score: 0.548
```

### 2. Logistic Regression:

- Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical).
- The data is split into training and test set with a ratio of 70:30
- The target variable is GSTATUS. The input vector to the learning model is of dimensions (39162, 237). The output is a binary class - either 0 or 1.
- Incorporated Over Sampling and Cross Validation techniques
- Model Performance:

```
precision score: 0.442
recall score: 0.634
f1 score: 0.521
auc roc score: 0.648
```

### 3. Deep Learning:

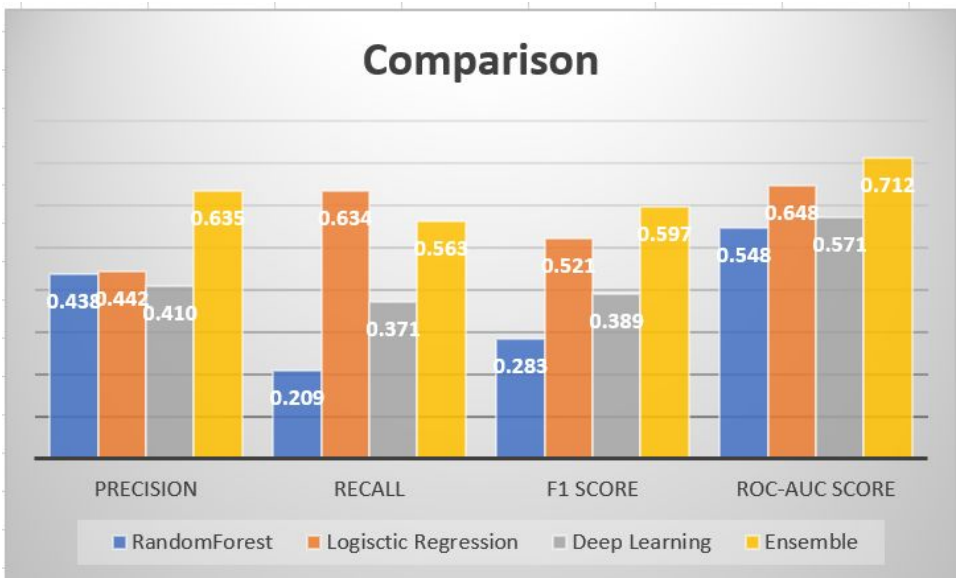
- Deep learning is a specific subfield of machine learning: a new take on learning representations from data that puts an emphasis on learning successive layers of increasingly meaningful representations.
- Model Specifications:
  - ❖ Number of layers:7
  - ❖ Number of hidden units:5
  - ❖ Activation functions for input and hidden layers: relu
  - ❖ Activation functions for output layer: sigmoid
  - ❖ Batch Normalization and drop out layers are added
- Model Performance:

```
precision score: 0.410
recall score: 0.371
f1 score: 0.389
roc auc score: 0.571
```

### 4. Ensemble Model:

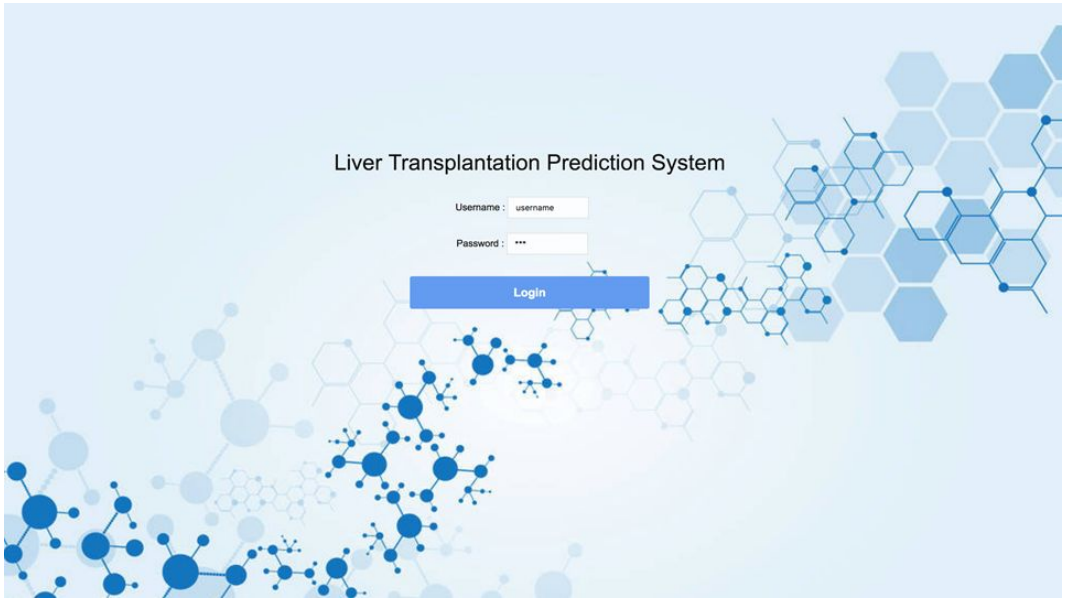
- Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance(bagging), bias (boosting), or improve predictions (stacking).
- We are combining Random Forest, Logistic Regression and Deep Learning in order to build a more efficient prediction Model using predict probability function.
- Model Performance:

```
Accuracy score: 0.772
precision score: 0.635
recall score: 0.563
f1 score: 0.597
roc_auc score: 0.712
```

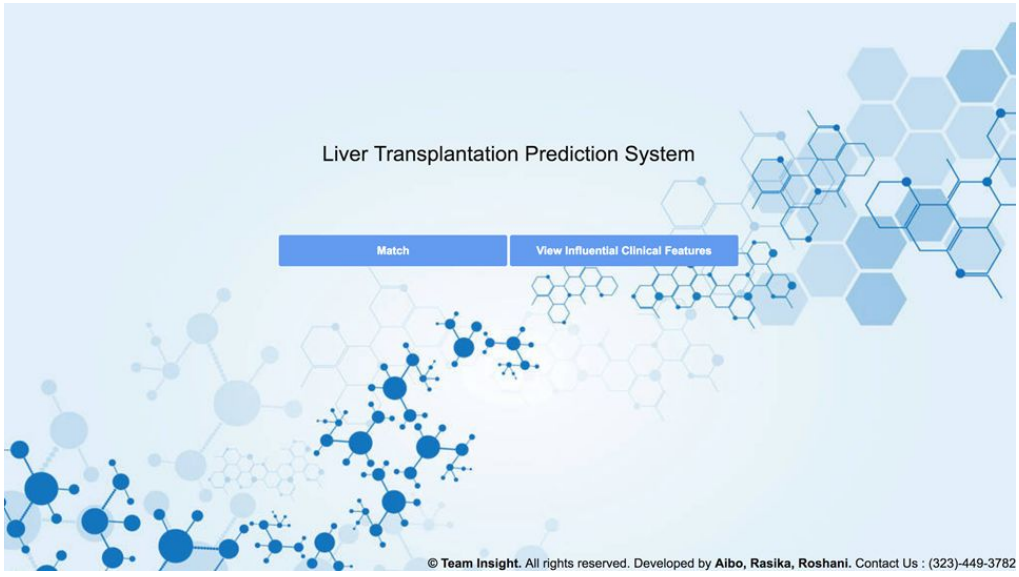


## Part 4: User Interface

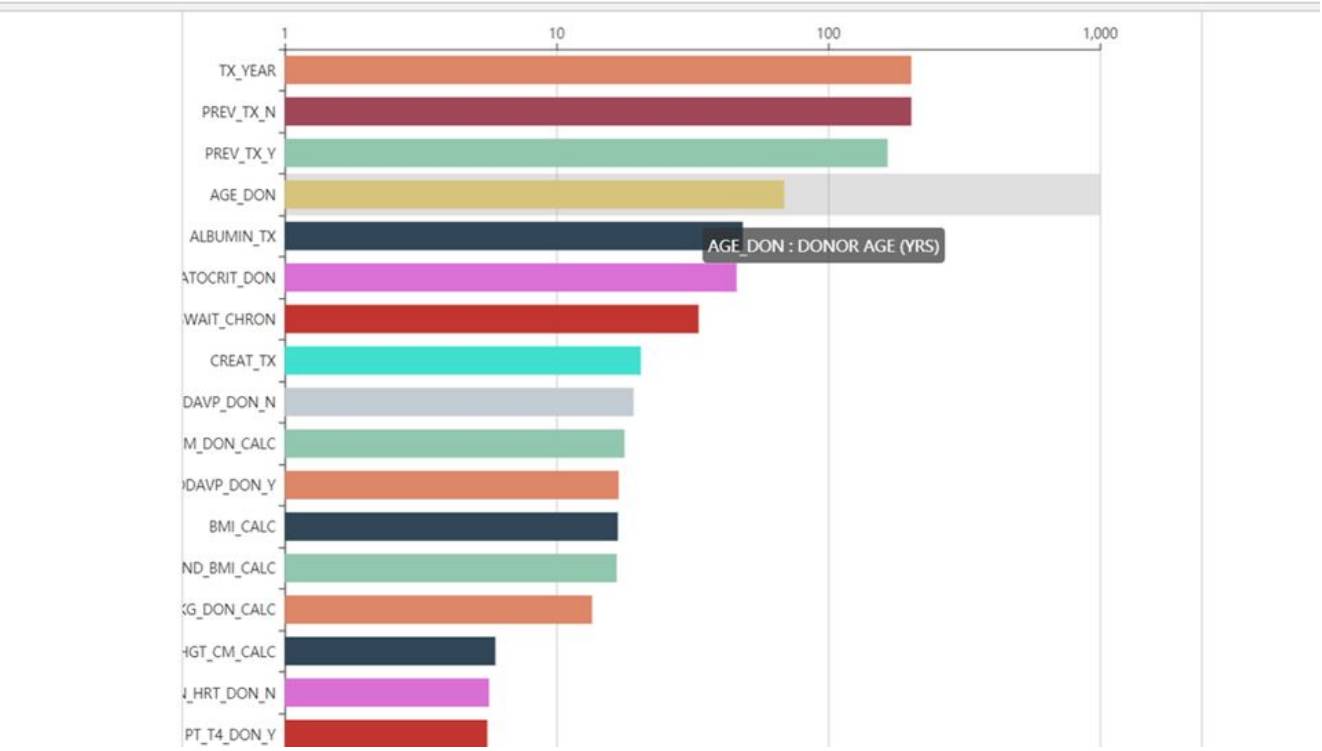
a) Login Page



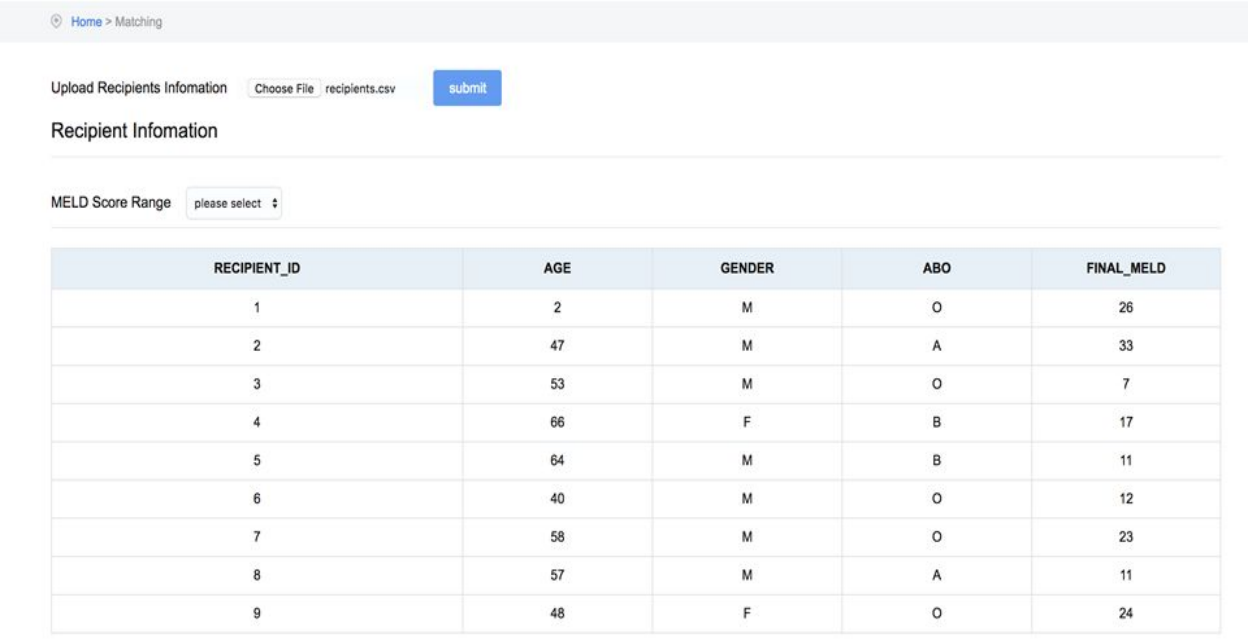

b) Home Page



c) Top Clinical Features influencing the identification of potential recipients




d) Upload Recipient Information




| RECIPIENT_ID | AGE | GENDER | ABO | FINAL_MELD |
|---|---|---|---|---|
| 1 | 2 | M | O | 26 |
| 2 | 47 | M | A | 33 |
| 3 | 53 | M | O | 7 |
| 4 | 66 | F | B | 17 |
| 5 | 64 | M | B | 11 |
| 6 | 40 | M | O | 12 |
| 7 | 58 | M | O | 23 |
| 8 | 57 | M | A | 11 |
| 9 | 48 | F | O | 24 |

## e) Upload Donor Information

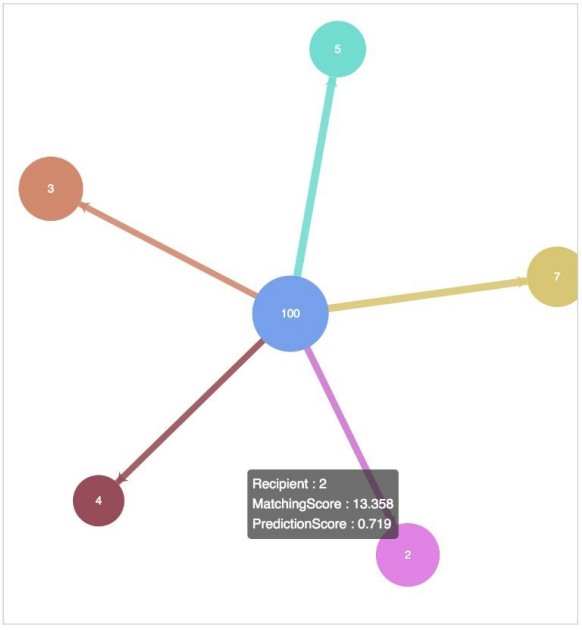
Upload Donor Information Choose File donors.csv submit

Donor Information Match

| DONOR_ID | BMI | ABO_DON | GENDER_DON | RESULT |
|---|---|---|---|---|
| 100 | 15.74124834 | A1 | M | check |
| 101 | 22.00656171 | O | M | check |
| 102 | 26.28780718 | A | F | check |
| 103 | 18.70774927 | A | F | check |
| 104 | 30.13001042 | A | M | check |
| 105 | 19.34851101 | B | F | check |
| 106 | 68.44211285 | O | F | check |
| 107 | 28.34466208 | O | F | check |
| 108 | 25.56610665 | O | M | check |

## f) Results

Final Results



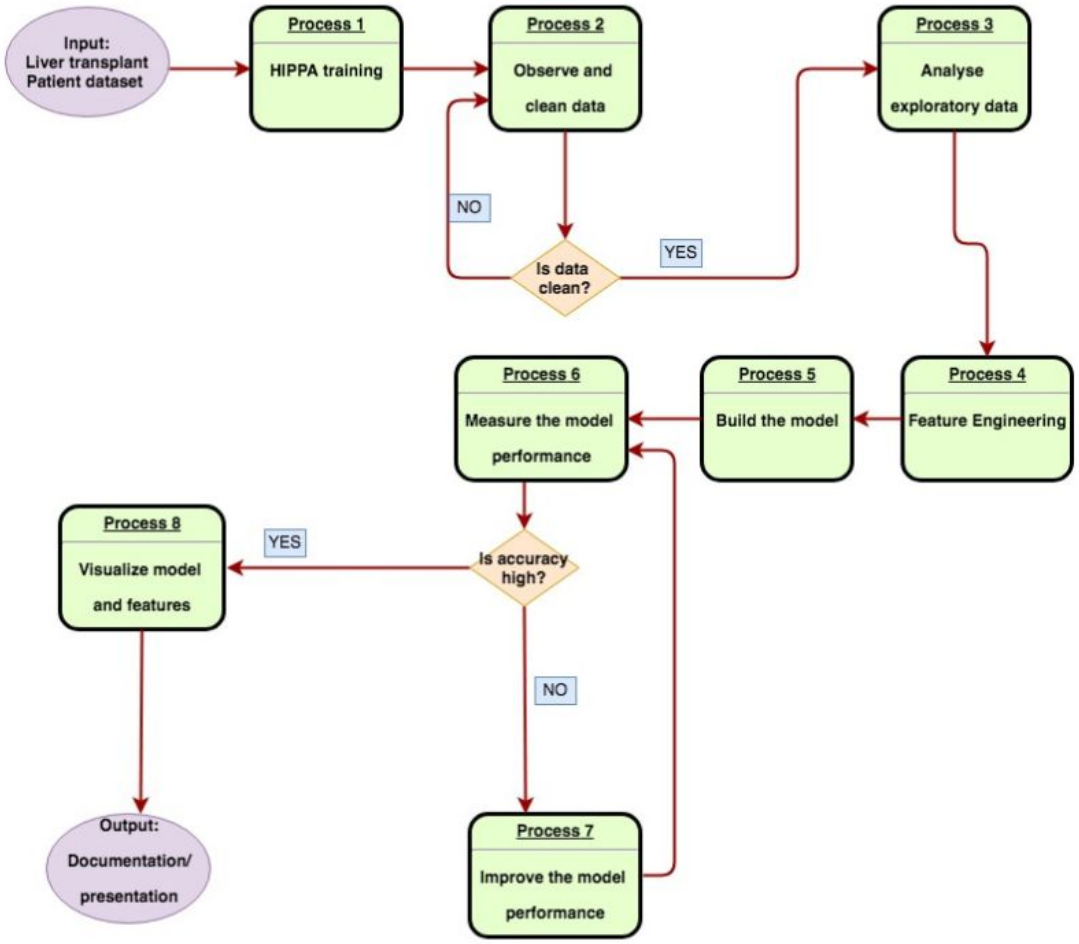

| RECIPIENTID of DONOR: 100 | MATCHINGSCORE | PREDICTIONSCORE | MELDSCORE |
|---|---|---|---|
| 3 | 13.505 | 0.639 | 7 |
| 2 | 13.358 | 0.719 | 33 |
| 7 | 12.69 | 0.788 | 23 |
| 5 | 11.841 | 0.824 | 11 |
| 4 | 10.773 | 0.663 | 17 |

## REFERENCES

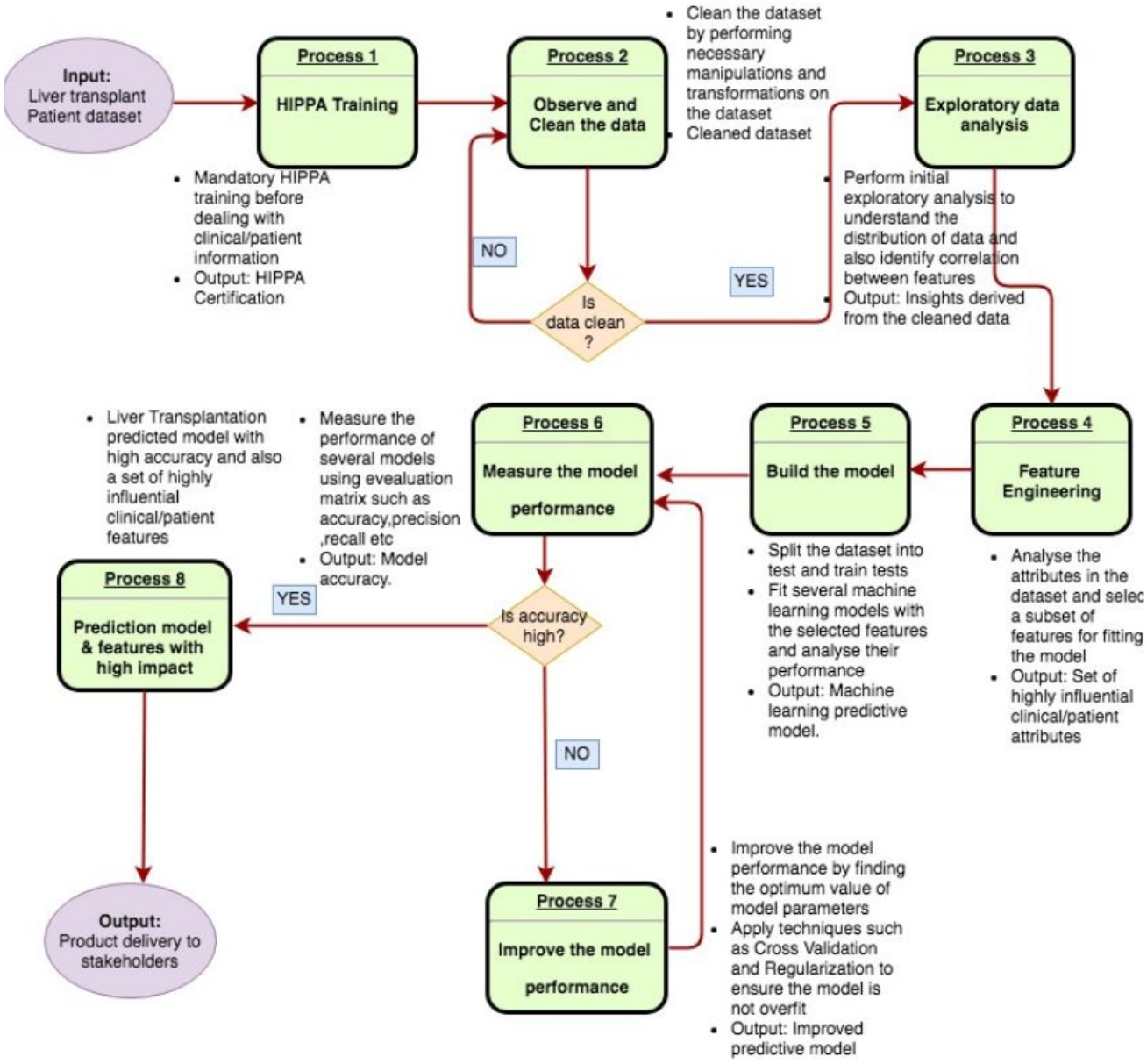

1. https://www.uptodate.com/contents/model-for-end-stage-liver-disease-meld
2. https://www.uptodate.com/contents/liver-transplantation-donor-selection
3. http://primarycare.imedpub.com/machine-learning-in-the-prediction-of-costs-forliver-transplantation.php?aid=20930
4. https://www.uptodate.com/contents/model-for-end-stage-liver-disease-me
5. https://pdfs.semanticscholar.org/c5a9/67eaded74a9fc414de4ad5120b0b66acd2c3.pdf
6. https://www.omicsonline.org/open-access/artificial-neural-networks-in-prediction-of-patient-survival-after-liver-transplantation-2157-7420-1000215.php?aid=67545
7. https://www.uptodate.com/contents/liver-transplantation-donor-selection/
8. https://doi.org/10.1371/journal.pone.0186301


## APPENDICES

A. Common Process Mapping



## B.Detailed Process Map



## C.Functional Process Map