

LiverMatch

Modeling Acceptance Behavior to Optimize Liver Allocation

Proposal Midterm Final

Midterm Checkpoint

Background

Literature Review

Center variability in liver offer acceptance for the highest priority patients varies from 15.7% to 58.1% [2]. This reluctance significantly impacts patient survival, as a 5% absolute decrease in a center's adjusted acceptance rate correlates with a 27% increased odds of waitlist mortality [1]. Institutional rejection patterns are so varied that candidates are up to three times as likely to receive frequently declined grafts from older donors by listing elsewhere [3][4]. Acceptance practices and the donor-to-candidate ratio independently cause geographic disparities in the median allocated MELD score. To address utilization issues, the OPTN implemented the risk-adjusted offer acceptance rate ratio performance metric, which relies on predictive modeling using donor characteristics.

Dataset

The primary datasets for LiverMatch are from the [Scientific Registry of Transplant Recipients \(SRTR\) Database](#) and the [Organ Procurement Transplantation Network \(OPTN\)](#).

Our first dataset is the DECEASED_DONOR_DATA dataset, which details demographic and organ information from deceased organ donors. This dataset contains 308,308 datapoints with 539 total features and covers individuals that have at least 1 organ recovered for transplant since October 1, 1987.

The data was initially broken into two separate files: the dataset as a .DAT and the respective labels as a .htm. This presented the first challenge with the data, where each dataset needed to be merged with their labels to create human readable files.

Obs	LABEL	FORMAT	LENGTH	TYPE	START	END
1	HIV_DON	\$	2	Character	1	2
2	DA1		6	Numeric	3	8
3	DA2		6	Numeric	9	14
4	ABO_DON	\$	3	Character	15	17
5	DB1		6	Numeric	18	23
6	DB2		6	Numeric	24	29
7	DDR1		6	Numeric	30	35
8	DDR2		6	Numeric	36	41
9	GENDER_DON	\$	1	Character	42	42
10	HOME_STATE_DON	\$	2	Character	43	44
11	RECOVERY_DATE_DON	MMDDYY	10	Numeric	45	54
12	WARM_ISCH_TM_DON		6	Numeric	55	60
13	RECOV_COUNTRY_DON	\$	3	Character	61	63
14	RECOV_OUT_US_DON	\$	1	Character	64	64
15	ETHNICITY_DON		6	Numeric	65	70
16	HBSAB_DON	\$	2	Character	71	72
17	EBV_IGG_CAD_DON	\$	2	Character	73	74
18	EBV_IGM_CAD_DON	\$	2	Character	75	76
19	LT_KI_GLOMERUL_COUNT		11	Numeric	77	87
20	RT_KI_GLOMERUL_COUNT		11	Numeric	88	98

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29						
1	N	A	2	3	B	44	35	4	97	M	OH	10/06/1987	.	N	A	N	0	N	A	A	.	.	N	A	N	A	A	.	.	N	A					
2	N	24	.	O	7	21	4	7	M	NM	10/04/1987	.	N	A	N	1	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	N	A		
3	N	26	1	A	8	27	2	3	M	NC	10/11/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	0	N	A	
4	N	3	28	O	53	58	2	5	M	NC	10/05/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	0	N	A	
5	N	32	11	O	44	53	5	6	F	AL	09/30/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	0	0	N	A
6	N	1	.	O	8	57	3	6	F	AL	09/30/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	0	0	N	A
7	N	2	24	B	27	44	3	4	F	AL	10/03/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	0	0	N	A
8	N	11	28	A	7	53	2	9	M	AL	10/01/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	999	0	0	0	0	0	0	N	A
9	N	1	11	O	8	35	3	.	F	NY	10/03/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	1	0	0	0	0	0	0	N	A
10	N	2	.	A	8	27	3	6	F	MD	10/09/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	0	0	N	A
11	N	2	3	O	7	97	2	4	M	NM	10/07/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	0	0	N	A
12	U	.	.	A	.	.	.	.	M	NV	10/10/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	2	0	0	0	0	0	0	N	A
13	N	1	.	O	7	8	4	5	M	VA	10/03/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	3	0	0	0	0	0	0	N	A
14	N	2	11	A	7	41	2	.	M	NC	10/02/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	2	0	0	0	0	0	0	N	A
15	N	2	24	O	8	51	4	5	M	LA	10/13/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	2	0	0	0	0	0	0	N	A
16	N	28	33	A	60	.	2	4	M	N	10/13/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	2	0	0	0	0	0	0	N	A
17	N	2	3	O	7	37	2	.	M	MO	10/06/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	2	0	0	0	0	0	0	N	A
18	N	24	11	A	62	35	4	.	F	KS	10/05/1987	.	N	A	N	1	N	A	A	A	.	.	N	A	N	N	1	999	0	0	0	0	0	0	N	A
19	N	1	28	A	8	57	2	3	M	MS	10/11/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	2	0	0	0	0	0	0	N	A
20	N	1	31	B	8	62	3	4	M	MS	10/08/1987	.	N	A	N	0	N	A	A	A	.	.	N	A	N	N	1	2	0	0	0	0	0	0	N	A

The readable file is as shown below:

	HIV_DON	DA1	DA2	ABO_DON	DB1	DB2	DDR1	DDR2	GENDER_DON	HOME_STATE_DON	RECOVERY_DATE_DON	WARM_ISCH_TM_DON	RECOV_COUNTRY_DON	RECOV
1	NA	2	3	B	44	35	4	97	M	OH	10/06/1987	.	NA	N
2	N	24	.	O	7	21	4	7	M	NM	10/04/1987	.	NA	N
3	N	26	1	A	8	27	2	3	M	NC	10/11/1987	.	NA	N
4	N	3	28	O	53	58	2	5	M	NC	10/05/1987	.	NA	N
5	N	32	11	O	44	53	5	6	F	AL	09/30/1987	.	NA	N
6	N	1	.	O	8	57	3	6	F	AL	09/30/1987	.	NA	N
7	N	2	24	B	27	44	3	4	F	AL	10/03/1987	.	NA	N
8	N	11	28	A	7	53	2	9	M	AL	10/01/1987	.	NA	N
9	N	1	11	O	8	35	3	.	F	NY	10/03/1987	.	NA	N
10	N	2	.	A	8	27	3	6	F	MD	10/09/1987	.	NA	N
11	N	2	3	O	7	97	2	4	M	NM	10/07/1987	.	NA	N
12	U	.	.	A	.	.	.	.	M	NV	10/10/1987	.	NA	N
13	N	1	.	O	7	8	4	5	M	VA	10/03/1987	.	NA	N
14	N	2	11	A	7	41	2	.	M	NC	10/02/1987	.	NA	N
15	N	2	24	O	8	51	4	5	M	LA	10/13/1987	.	NA	N
16	N	28	33	A	60	.	2	4	M	NA	10/13/1987	.	NA	N
17	N	2	3	O	7	37	2	.	M	MO	10/06/1987	.	NA	N
18	N	24	11	A	62	35	4	.	F	KS	10/05/1987	.	NA	N
19	N	1	28	A	8	57	2	3	M	MS	10/11/1987	.	NA	N
20	N	1	31	B	8	62	3	4	M	MS	10/08/1987	.	NA	N

Another important dataset is the LIVER_DATA dataset, which includes records for livers that have been registered for transplant or for the waiting list. The dataset includes demographic information of the donor of the liver, and may not include waitlist information if the liver was listed immediately for transplant without being placed on the waiting list.

	TXPAN	TXVCA	CLSTR_OLD	CLSTRTP_OLD	HBEAB_OLD	HEPD_OLD	MRCREATG_OLD	PREV_PI_TX_TCR_ARCHIVE	STATUS_TCR	TOT_SERUM_ALBUM	PREV_MALIG_TY	AMIS	B
1	NA	NA	N	997	U	U	NA	.	V	.	.	1	2
2	NA	NA	N	997	U	U	NA	0	V	.	.	2	2
3	NA	NA	NA	NA	NA	NA	NA	.	NA	.	.	.	.
4	NA	NA	N	NA	ND	ND	NA	.	V	.	.	.	.
5	NA	NA	NA	NA	NA	NA	NA	.	NA	.	.	.	.
6	NA	NA	N	997	C	C	N	.	V	.	.	1	2
7	NA	NA	N	NA	C	C	NA	.	V	.	.	0	1
8	NA	NA	NA	NA	NA	NA	NA	.	NA	.	.	.	.
9	NA	NA	N	997	N	ND	NA	.	V	.	.	1	1
10	NA	NA	N	NA	N	ND	NA	.	V	.	.	1	1
11	NA	NA	N	997	ND	ND	NA	.	V	.	.	1	2
12	NA	NA	N	997	ND	ND	NA	.	V	.	.	2	2
13	NA	NA	N	997	N	N	NA	.	V	.	.	1	1
14	NA	NA	N	997	N	N	NA	.	V	.	.	2	2
15	NA	NA	NA	NA	NA	NA	NA	.	NA	.	.	.	.
16	NA	NA	N	997	C	C	NA	.	V	.	.	2	2
17	NA	NA	NA	NA	NA	NA	NA	.	NA	.	.	.	.
18	NA	NA	NA	NA	NA	NA	NA	.	NA	.	.	.	.
19	NA	NA	NA	NA	NA	NA	NA	.	NA	.	.	.	.
20	NA	NA	NA	NA	NA	NA	NA	.	NA	.	.	.	.

Once each dataset was processed and converted into a more readable state, the next challenge involved cleaning up messiness in the data. Numerous values were missing and were denoted with '.', 'Unknown', or 'U'. These values needed to be changed into NaN values for easier data processing and analysis later.

Problem Definition

Problem

The process of liver transplantation involves a sequence of offers for each donor organ, but transplant centers vary in their acceptance thresholds [2]. Some organs are accepted at the top of the match run, while others cascade through multiple centers before being selected. This creates variability in allocation efficiency, and any delay in the process can result in viable organs expiring or reaching patients later than optimal, increasing patient mortality.

Motivation

Eliminating this inefficiency in the liver allocation process reduces waitlist times, maximizes utility of donor organs, and increases allograft survival rates. By introducing a standardized prediction model, the project shortens waiting times for scarce organs and can reduce the disparity in waiting times and geographical variability between different transplant centers [2][3]. Streamlining the selection process also reduces operational costs and logistical burden on hospitals and organ procurement organizations.

Methods

Given our datasets and large problem scope we broke it down to three integral subproblems - Donor Characterization, Medical Center Profiling, and Recipient Population Analysis.

Donor Characterization:

This step essentially allows us to create a taxonomy of donor liver profiles given the features in our dataset. This step is important because moving forward creating profiles for donors will be simpler than rows of numbers that summarize to a meaningful profile.

The procedure of creating these profiles consists of:

- 1. Exploratory Data Analysis
- 2. Data Preprocessing
 - Replacing unknown values ('Unknown', 'U', '.') with NaN values.

- Removing rows with NaN values once analysis begins.
 - Converting string type data into a binary representation using one-hot encoding.
3. Clustering
- We began by using Kmeans to create our clusterings.
 - Using the elbow method over various values of K (1-10), we then measured it against each Kmeans distortion score. From the elbow method rule, we identified K = 3 as the "elbow".
 - We then ran a Kmeans and GMM with K = 3 and added the cluster assignment back to the original dataset.
4. Model
- Given the labels provided from the two clustering algorithms, we will use supervised learning and evaluate if our data can classify these donor profiles.

```
Summary for the NeuralNetwork model:
-----
Layer (type)                Output Shape      Param #
-----
Linear-1                    [-1, 128]         2,432
ReLU-2                      [-1, 128]         0
Dropout-3                   [-1, 128]         0
Linear-4                    [-1, 128]         16,512
ReLU-5                      [-1, 128]         0
Dropout-6                   [-1, 128]         0
Linear-7                    [-1, 5]           645
-----
Total params: 19,589
Trainable params: 19,589
Non-trainable params: 0
-----
Input size (MB): 0.00
Forward/backward pass size (MB): 0.01
Params size (MB): 0.07
Estimated Total Size (MB): 0.08
-----
```

- We will implement a neural network, specifically a ResNet18, with the architecture shown below.

Medical Center:

Here, we want to analyze trends in transplant centers and look at the features of organs that they accept for transplants. Our ultimate goal here is to find two things:

- We want to create a MELD score, which is a value that represents a patient's urgency for a liver transplant, with a higher MELD score being more urgent
- We also want to find a way to encode the likelihood or tendency of a transplant center to take a liver based on its features

The progress for this step is broken into:

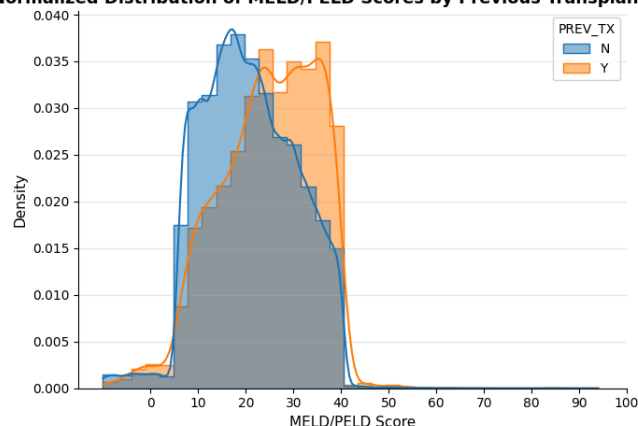
1. Data preprocessing
 - Replacing unknown values ('Unknown', 'U', '.') with NaN values, similar to the donor characterization step, then removing those NaN values after.
 - No one-hot encoding was necessary for this step, as the features selected were all numeric values
2. Aggregation
 - Grouping by each center (represented by center code OPO_CTR_CODE), we are able to obtain mean and median values for several important variables like donor age and SGOT test results (to determine if a liver has any health complications).

Recipient Population Analysis:

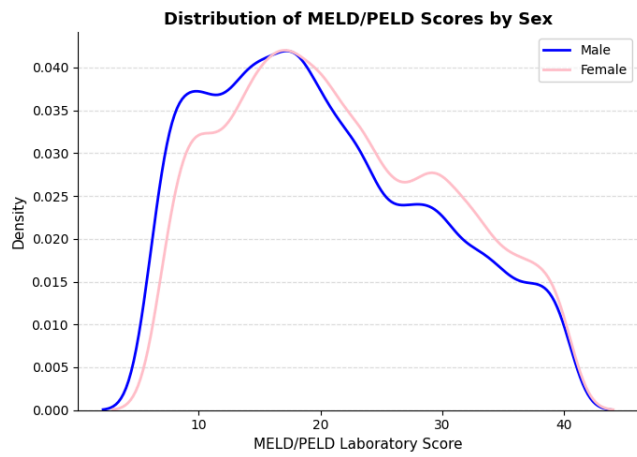
Here, we seek to understand what features of our recipients correlate to a higher MELD/PELD score, and how close that correlation is. We followed this process to find correlations:

1. Perform data preprocessing on LIVER_DATA by:
 - Removing rows with missing values in the MELD/PELD columns.
 - Dropping administrative or non-clinical features unrelated to MELD/PELD calculation.
 - Validating data ranges for numeric values such as MELD/PELD.
 - Replacing numeric diagnosis codes with their corresponding descriptive diagnosis labels.
2. Plot the relationship between relevant features and determine any correlation.
 - We primarily used probability density plots where we mapped the probability of a feature being a certain value given the final MELD/PELD score. The example below shows the probability that a patient on the transplant waitlist has had a prior transplant given they were assigned a particular MELD/PELD score.

Normalized Distribution of MELD/PELD Scores by Previous Transplant Status



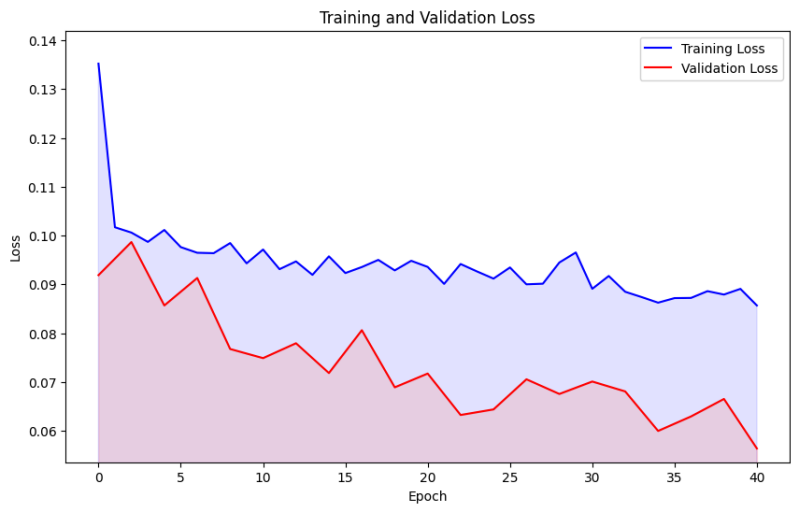
3. Reflect upon any observations to eliminate any external influence on results such as improper data sampling and presence of missing values.
 - One of our plots showed a slight correlation between being a woman and having a higher MELD/PELD score, but the lack of sufficient sex data for the majority of data points led us to believe this correlation was very weak.



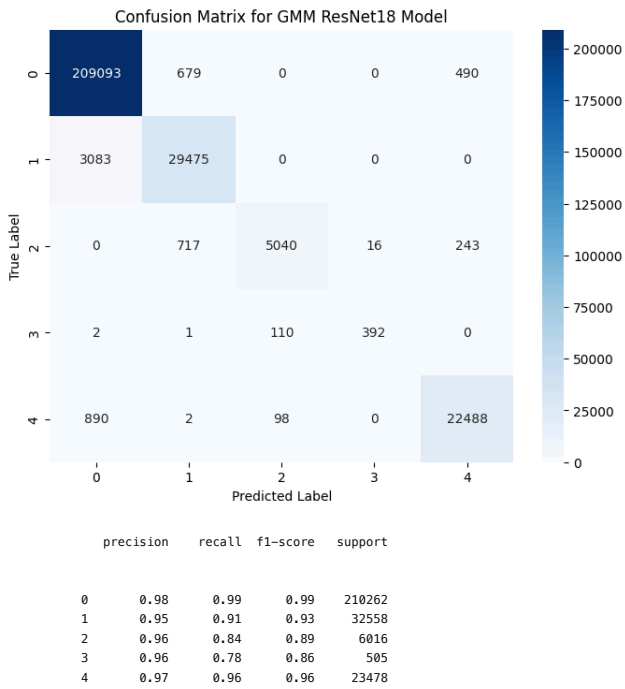
Results and Discussion

Donor Characterization:

Classifications with model assigned by GMM:



- Epoch 40/40, Train Loss: 0.0857, Validation Loss: 0.0564, Accuracy: 97.68%
- Reports:

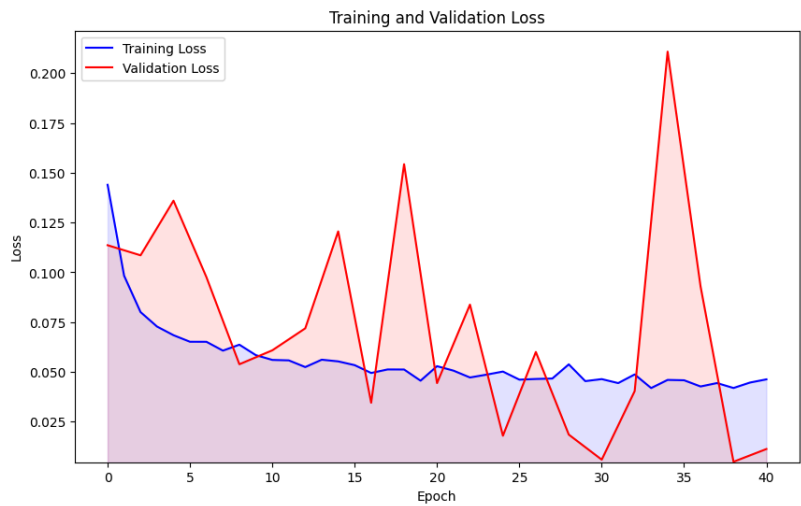


accuracy			0.98	272819
macro avg	0.97	0.89	0.93	272819
weighted avg	0.98	0.98	0.98	272819

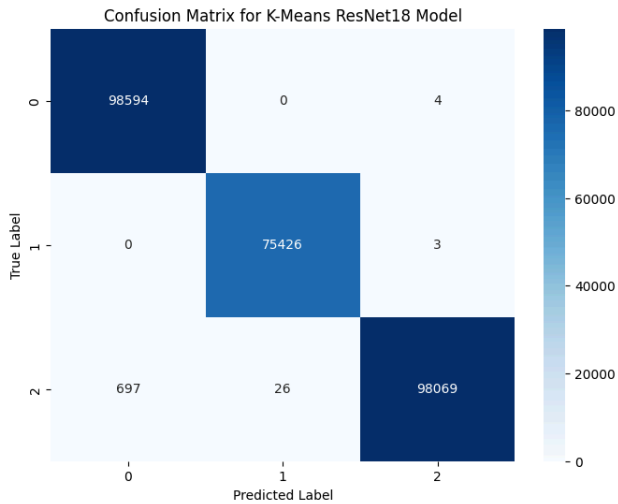
* Based on report, precision, recall, and f1-score, the model performed best on cluster 0 and worst on cluster 3.

* Overall, the model performed well since for the cluster with the worst performance, precision = 0.96, recall = 0.78, and f1-score = 0.86 which is all above 0.5

Classification with model assigned by K-Means:



- Epoch 40/40, Train Loss: 0.0462, Validation Loss: 0.0112, Accuracy: 99.71%
- Reports:



	precision	recall	f1-score	support
0	0.99	1.00	1.00	98598
1	1.00	1.00	1.00	75429
2	1.00	0.99	1.00	98792

accuracy			1.00	272819
macro avg	1.00	1.00	1.00	272819
weighted avg	1.00	1.00	1.00	272819

* Based on the report, precision, recall, and f1-score, the model performed equally well on all 3 clusters Overall, the model performed well since all clusters have precision, recall, and f1-score cl

* From this we came to a decision that K-means labeling caused overfitting because of the high validation loss and insanely high accuracy. This was not the case for GMM labeling, so moving forward we

Medical Center:

Selecting relevant features:

	DONOR_ID	OPO_CTR_CODE	AGE_DON	ABO_DON	GENDER_DON	COD_CAD_DON	TBILL_DON	SGOT_DON	SGPT_DON	HEP_C_ANTI_DON
0	100223.0	3441.0	0.0	B	M	NaN	NaN	NaN	NaN	None
1	12687.0	8866.0	16.0	O	M	3	NaN	NaN	NaN	None
2	187082.0	14911.0	2.0	A	M	3	NaN	NaN	NaN	None
3	218478.0	14911.0	24.0	O	M	3	NaN	NaN	NaN	None
4	76722.0	11346.0	16.0	O	F	3	NaN	NaN	NaN	None
...	...	...	...	...	...	...	...	...	...	...
308303	422202.0	3348.0	8.0	B	F	1	1.60	35.00	41.00	N
308304	422350.0	25172.0	55.0	A	F	1	0.40	65.00	26.00	N
308305	422211.0	527.0	20.0	O	F	999	0.30	25.00	10.00	N
308306	422461.0	12617.0	57.0	A2	M	3	0.20	17.00	13.00	P
308307	422384.0	5239.0	20.0	A	F	3	0.20	419.00	201.00	N

The table of aggregated values is shown below:

	avg_age_don	avg_tbill	avg_sgot	median_age_don	median_tbill	median_sgot	n_donors	hep_c_percent_pos
OPO_CTR_CODE								
93.0	38.765377	1.001713	101.635328	39.0	0.6	43.0	2016	11.016949
217.0	35.261625	1.156853	123.908122	36.0	0.7	56.0	1957	5.895564
527.0	37.764620	0.906117	135.727656	38.0	0.6	49.0	1368	4.442582
1054.0	42.461488	1.184096	234.615465	44.0	0.7	54.0	15294	10.098207
1116.0	44.724138	1.083714	79.272807	49.0	0.7	42.0	2552	2.938840
...	...	...	...	...	...	...	...	...
24614.0	38.858479	0.981018	124.281036	40.0	0.6	45.0	5879	6.404360
24986.0	39.928215	1.202203	125.171700	42.0	0.7	44.0	2605	5.664395
25172.0	41.118068	0.969091	118.363474	43.0	0.6	46.0	4472	4.135338
25389.0	39.325137	0.990191	120.129835	40.0	0.7	45.0	2562	9.211654
25581.0	39.940374	1.056168	140.746954	42.0	0.7	53.0	2784	4.889780

Recipient Population Analysis:

We found the following features were associated with higher MELD/PELD values:

- 1. Age - Surprisingly, we found that younger and middle aged patients (20-50) were most likely to have the highest MELD/PELD values.
- 2. Previous Transplant Status - Patients with prior transplant history were more likely to have higher MELD/PELD scores.
- 3. Amount of Bilirubin and Serum Creatinine were correlated with higher MELD/PELD scores.
- 4. Severity of Ascetis & Encephalopathy
- 5. Presence of Cirrhosis.

Next Steps

Donor Characterization:

We plan to explore the centroid-based clustering methods that improve on current deep neural network-based K-means approaches [5], aiming for more meaningful patient groupings. Inaddition, We will also consider using longitudinal features to observe the patient conditions over time. We hope that by doing this, we can gain more insights into the dataset.[6]

Medical Center:

After aggregation of the data has been completed, we will want to consult with Dr. West and Emory to determine if any other features may be worthwhile to analyze and aggregate into the data, as well as create features to represent the distribution of donor profiles across each recipient center. We will connect the centers data with clusters found in the recipient data and determine which centers are more likely to accept certain donor/liver profiles.

Recipient Population Analysis:

Our next steps will be to implement **clustering** to identify groups of patients with similar clinical and demographic profiles. We can use key variable such as lab values (BILIRUBIN, ALBUMIN), clinical records (PREV_TX), and patient demographics (AGE, SEX) to identify natural subpopulations within the waiting list. After standardizing numeric features and encoding categorical ones, we can use methods like K-Means to group patients based on similar lab/clinical history and disease severity. Comparing clusters can reveal deeper patterns beyond just correlation such as which combinations of characteristics occur frequently together and how these relate to sickness level.

Gantt Chart

See our Gantt Chart [here](#).

Contributions

Name	Contribution
Thinh	Donor characterization model, Model Training, Analysis, Results
Arjun	Recipient Profiling, Analysis, Results
Alex	Center Profiling, Data cleaning, Analysis, Results
Jordan	Dataset preparation for subproblem, task creation, website
Kim Phuoc	Donor characterization clustering, Model Training, Analysis, Results

References

[1] C. E. Haugen et al., "Offer Acceptance Patterns for Liver Donors Aged 70 and Older," Liver Transplantation, vol. 28, no. 4, pp. 571–580, Apr. 2022, doi: <https://doi.org/10.1002/lt.26309>.

[2] D. S. Goldberg et al., "Liver transplant center variability in accepting organ offers and its impact on patient survival," Journal of Hepatology, vol. 64, no. 4, pp. 843–851, Apr. 2016, doi: <https://doi.org/10.1016/j.jhep.2015.11.015>.

[3] A. Wey et al., "Offer acceptance practices and geographic variability in allocation model for end-stage liver disease at transplant," Liver Transplantation, vol. 24, no. 4, pp. 478–487, Mar. 2018, doi: <https://doi.org/10.1002/lt.25010>.

[4] "New pre-transplant performance metric now in effect, offer acceptance rate ratio - OPTN," Hrsa.gov, 2025. <https://optn.transplant.hrsa.gov/news/new-pre-transplant-performance-metric-now-in-effect-offer-acceptance-rate-ratio/>.

[5] D. Roy, "An Approach Towards Learning K-means-friendly Deep Latent Representation," arXiv, 2024. <https://doi.org/10.48550/arXiv.2411.19496>.

[6] X. Gao et al., "Predicting Long-Term Allograft Survival in Liver Transplant Recipients," arXiv, 2024. <https://doi.org/10.48550/arXiv.2408.05437>.

Special Reward Consideration:

Yes, we opt-in for Outstanding Project.

GitHub Pages Link

The link to our project page can be found [here](#).