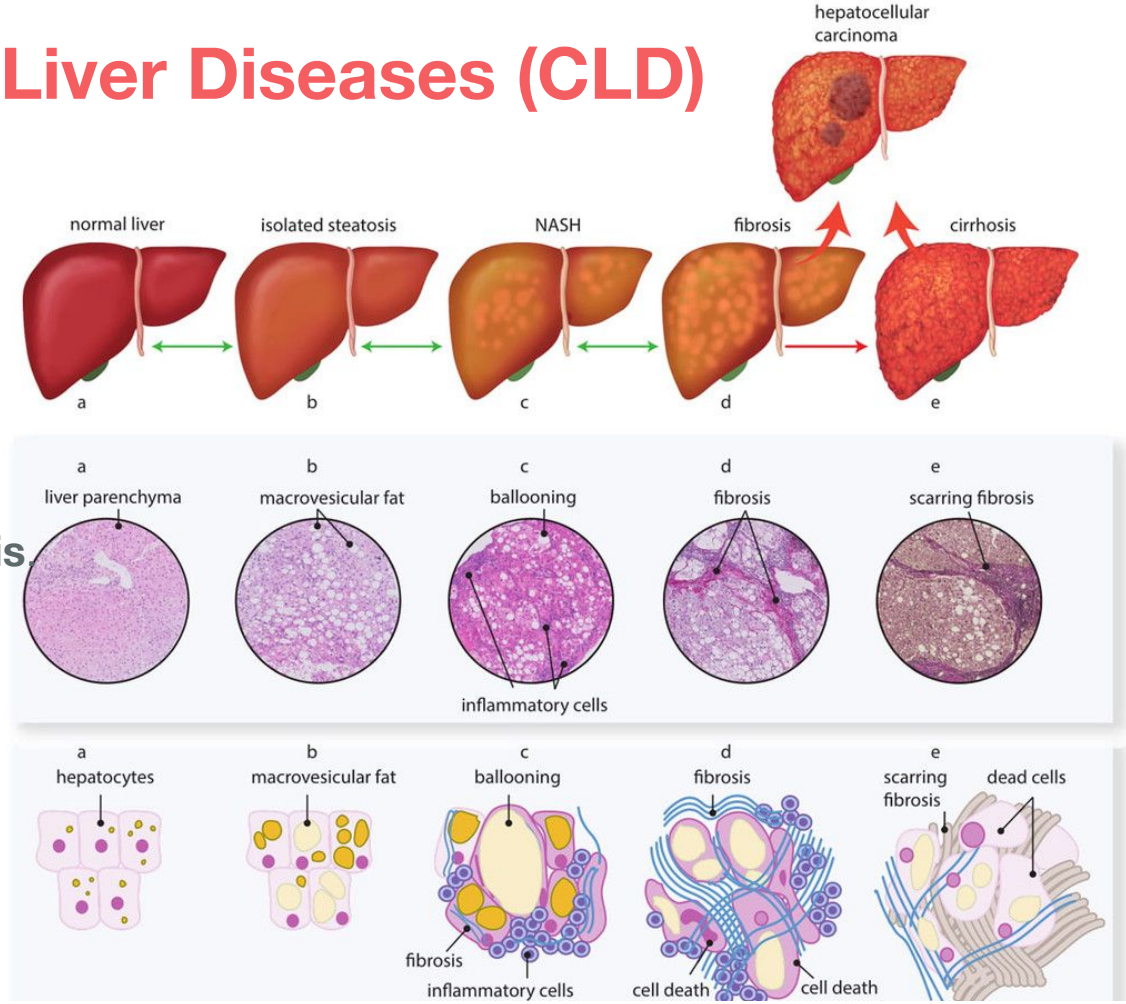
Predicting the Cause of Chronic Liver Diseases using Clinical Data

Yukti Makhija
2019BB10067

Prof. Ishaan Gupta
IIT Delhi

Overview of Chronic Liver Diseases (CLD)

- Continuous **worsening of liver functions** which lasts more than **6-7 months**.
- There is a generation of harmful proteins and **clotting factors**, **inflammation** of liver parenchyma that results in **cirrhosis** and **fibrosis**.
- Around **50%** of patients with Chronic Liver Failure develop **life-threatening** conditions.



Existing Scoring Systems

- Due to the high mortality rate observed worldwide, there are many scoring systems which use clinical parameters at the time of diagnosis:
  a. **MELD (Model of End stage Liver Disease)**

     Internal Normalized Score, Creatinine, Bilirubin, Serum Sodium

  b. **European Association's CLIF-C ACLF**
- In most cases, it has been observed that we get a confirmed prognosis between the third and seventh day after hospitalisation.

Changes in the prevalence of the causes of CLD from 1988 to 2008

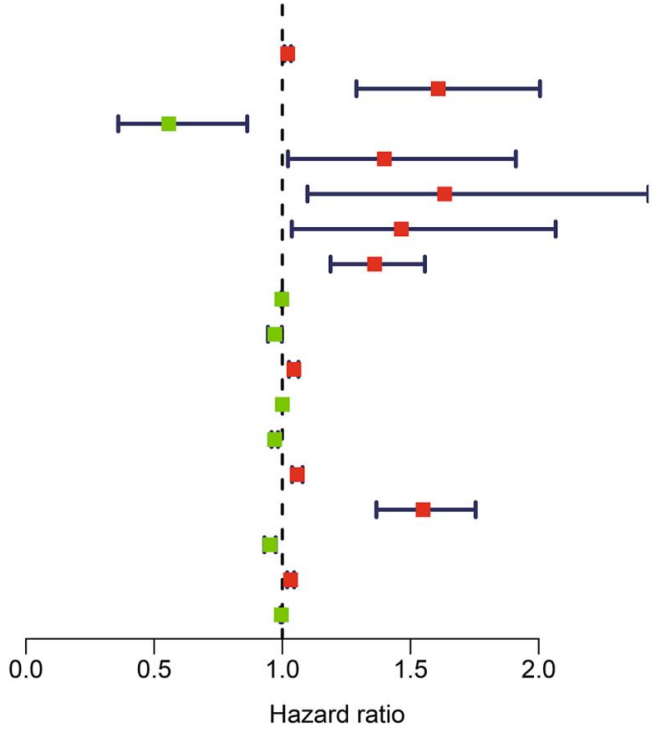
- Increase in the number of CLD cases in these twenty years.
- CLD cases due to **Non-alcoholic Fatty Liver Disease (NAFLD)** were drastically increasing with the increase in **obesity**.
- **Decrease** in the percentage of cases caused by **HCV** infections.
- **NASH (Non-Alcoholic SteatoHepatitis)** is the most severe type of NAFLD and often evolves into cirrhosis. NASH is very common in **diabetic** and **obese** patients and is the reason behind the **high mortality rate**.

Existing Models

Dynamic prediction model for the prognosis of ACLF given by Yu et al. (2021)

Univariate COX regression analysis

	pvalue	Hazard ratio
Age	0.001	1.021(1.008−1.034)
WGO type (A/B/C)	<0.001	1.608(1.289−2.005)
Etiology (Alcoholic)	0.009	0.557(0.360−0.863)
Infections	0.036	1.398(1.022−1.911)
Ascites	0.015	1.633(1.098−2.428)
Gastrointestinal Bleeding	0.030	1.464(1.037−2.066)
Hepatic Encephalopathy	<0.001	1.360(1.188−1.556)
Glutamate Transpeptidase	0.026	0.998(0.997−1.000)
Albumin	0.037	0.971(0.944−0.998)
Total Bilirubin	<0.001	1.045(1.026−1.063)
Cholinesterase	0.045	0.9999(0.9997−1.000)
Prothrombin Activity	<0.001	0.970(0.958−0.983)
Blood Urea Nitrogen	<0.001	1.058(1.038−1.079)
Creatinine	<0.001	1.549(1.367−1.754)
Serum Sodium	<0.001	0.952(0.930−0.974)
Neutrophil percentage	<0.001	1.032(1.019−1.046)
Platelet	0.006	0.996(0.993−0.999)



Objectives

- Using longitudinal clinical data on liver disease patients to predict Chronic Liver Diseases and Liver Failure.
- Feature engineering and enhance model interpretability towards clinical deployment.
- Making a parser that extracts useful information from EMRs.

Data Description

- Data was collected for **3747 liver failure patients** at the Institute of Liver and Biliary Sciences (ILBS).
- The initial dataset contained **41 features** for each patients.
- Multiple entries for some patients.
- We categorized our data in two broad groups
  - ➔ **Clinical Parameters**
  - ➔ **Physical Parameters:** Age, Gender

Clinical Parameters

Parameters	Description
Clotting time (CT)	Time after which clot formation starts after adding the start reagent to the blood.
Clot Formation time (CFT)	Time taken for the clot firmness to reach 20mm after CT.
A5, A10, A15, A20, A25, A30 values	Clot amplitude (firmness) after 5,10,15,20,25,30 mins.
Maximum Clot Firmness (MCF)	Largest observed amplitude. (absolute strength of the platelet clot)
Alpha-angle	Tangential angle 0 and the curve when clot firmness has reached 20mm.
Maximum Lysis (ML)	Percent of clot stability lost wrt MCF at the end of the test.
Lysis Index after 30 mins (LI 30)	Clot stability wrt MCF (%) is measured thirty minutes after clotting time.
MaxV	Maximum velocity of clot formation

Data Preprocessing

- All columns with more than 15% missing values were deleted.
- We imputed the missing values with mean for continuous variables.
- The data obtained after preprocessing contains **1104 patients and 26 features**. This dataset is in a systematic form and has been used for further analysis and machine learning.

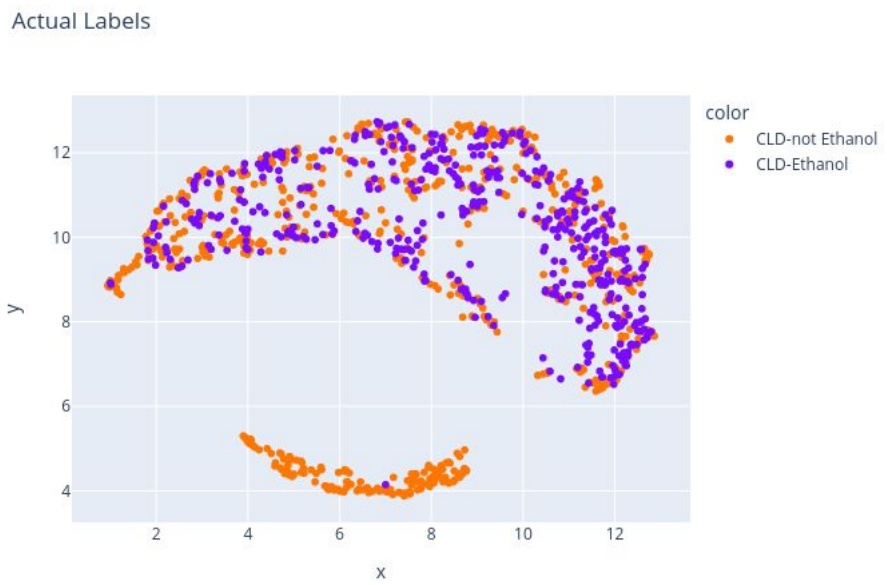
Types of CLD and Distribution of Patients

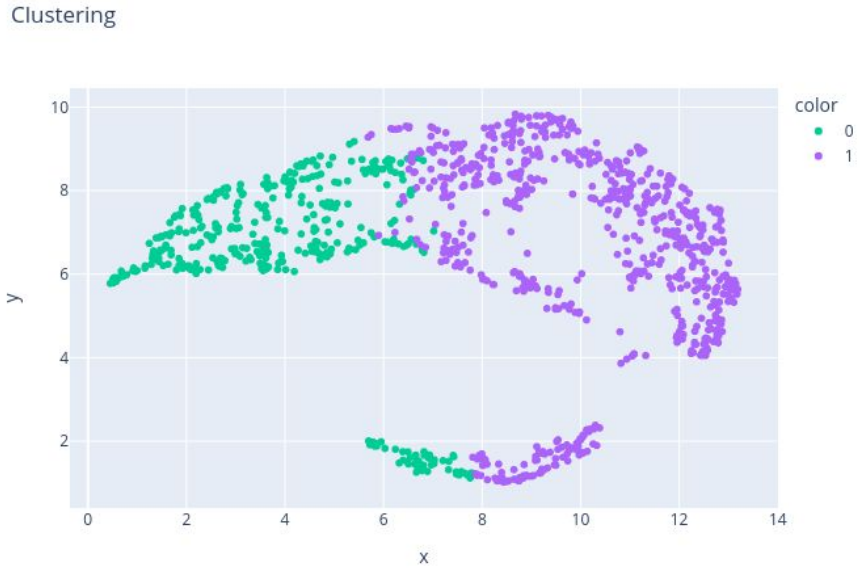
Type of CLD	Cause	Number of Patients
CLD Ethanol	Excessive drinking for several years	459 (62.28%)
CLD NASH	Obesity, patient's lifestyle	159 (21.57%)
CLD HBV	Hepatitis B viral infection	41 (5.56%)
CLD HCV	Hepatitis C viral infection (hepatotropic RNA virus)	37 (5.02%)
CLD Cryptogenic	Cause is unknown	41 (5.56%)

Development of the models

- Types of classification:
  - CLD Ethanol Prediction (Binary)
  - CLD NASH/Ethanol/Infectious (Multiclass)
- Trained different ML prediction models on the curated data
  - Logistic Regression
  - Decision Trees
  - Support Vector Machines
  - XGBoost

Projection and Clustering



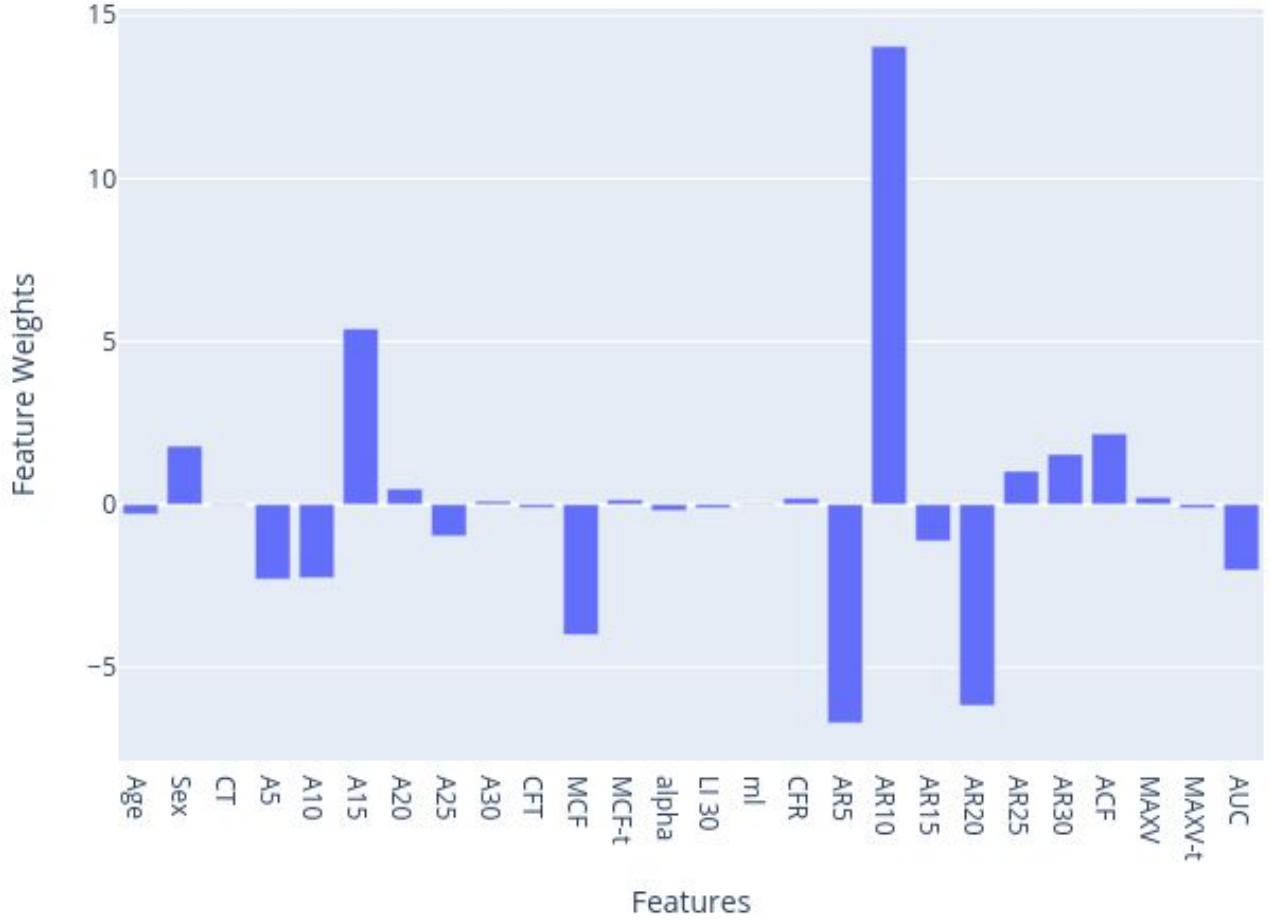


UMAP projection colored with (a): actual labels (b): K-Means clustering labels

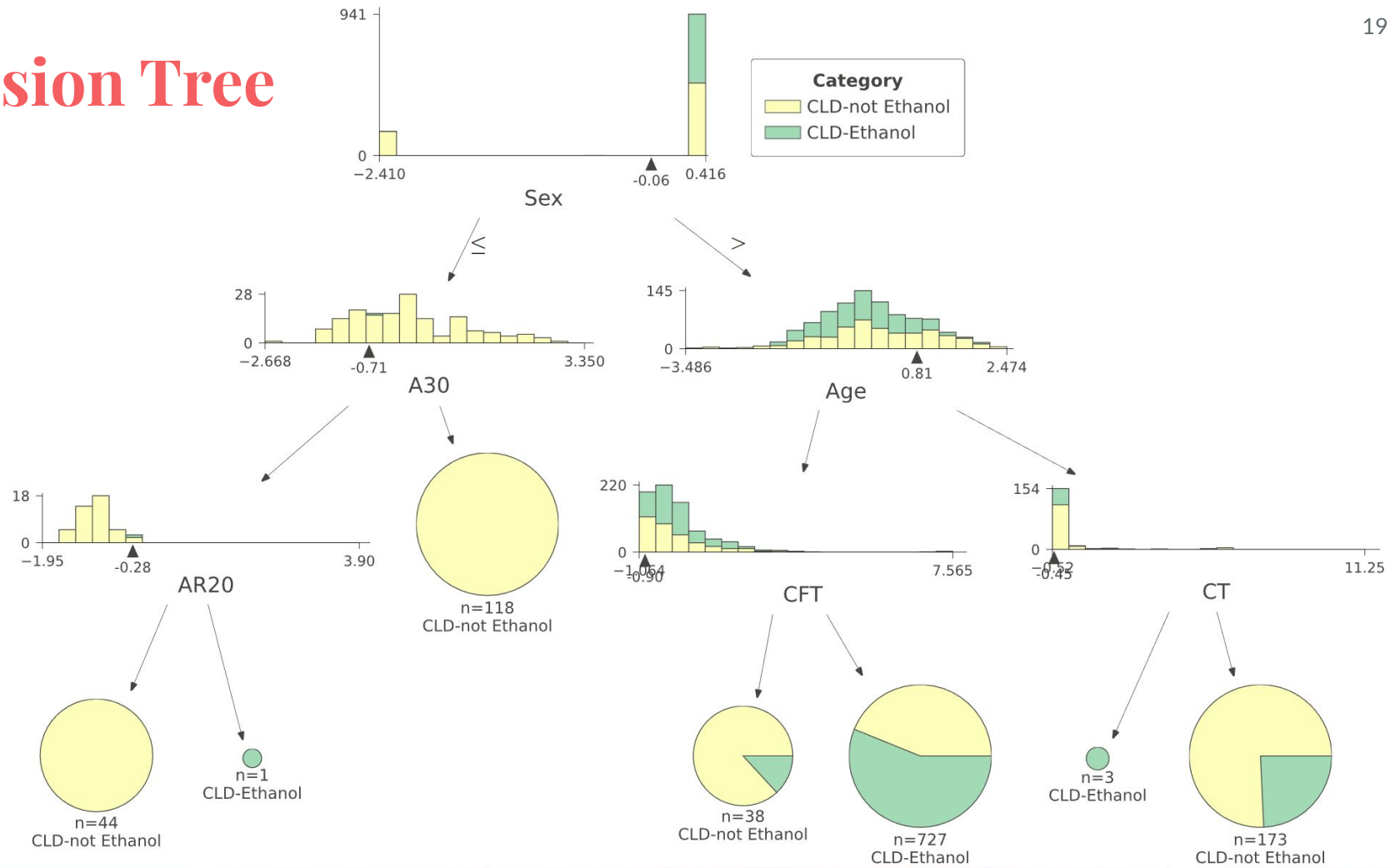
Results: CLD-Ethanol Prediction (Binary)

Metrics	Logistic Regression	Decision Tree	SVM (linear kernel)	SVM (RBF Kernel)	XGBoost
Training Accuracy	67.381	66.576	66.938	70.853	**78.14**
Testing Accuracy	63.578	63.758	65.659	66.567	**66.96**
Precision	0.559	0.540	0.573	0.593	**0.617**
Recall	0.588	0.865	0.682	0.623	**0.543**
F-score	0.573	0.665	0.623	0.608	**0.578**
AUC-ROC	0.629	0.670	0.660	0.660	**0.651**

Logistic Regression: Learnt Feature Weights

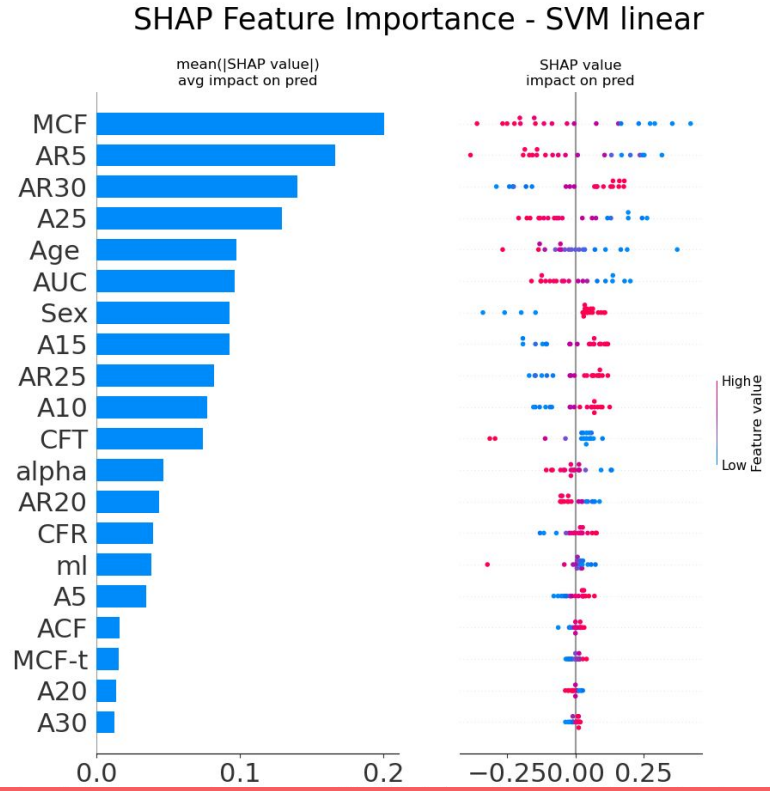


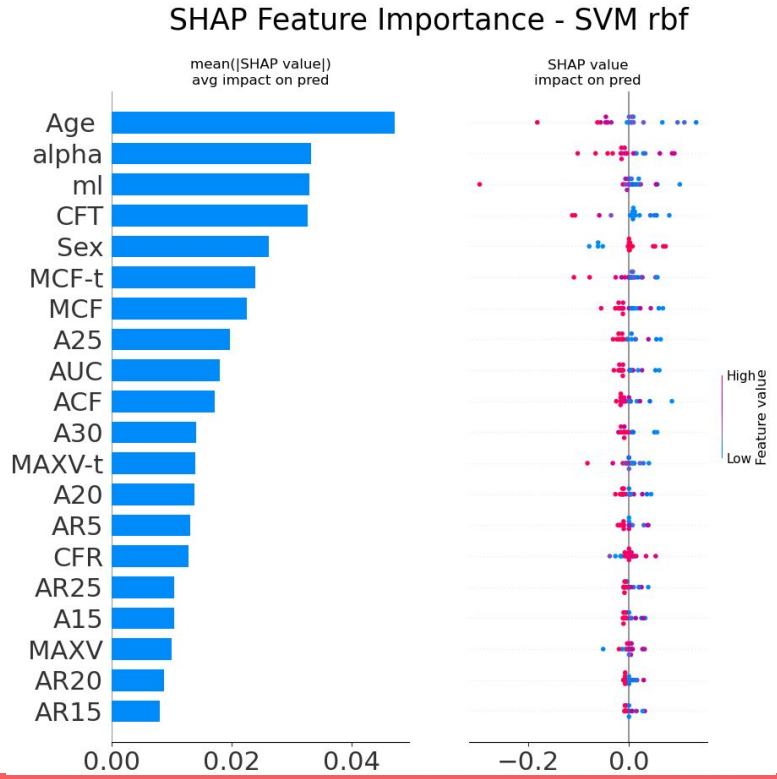
Decision Tree



SHAP Feature Importance

We performed feature engineering using post-hoc techniques like SHAP to explain the working of ML models.





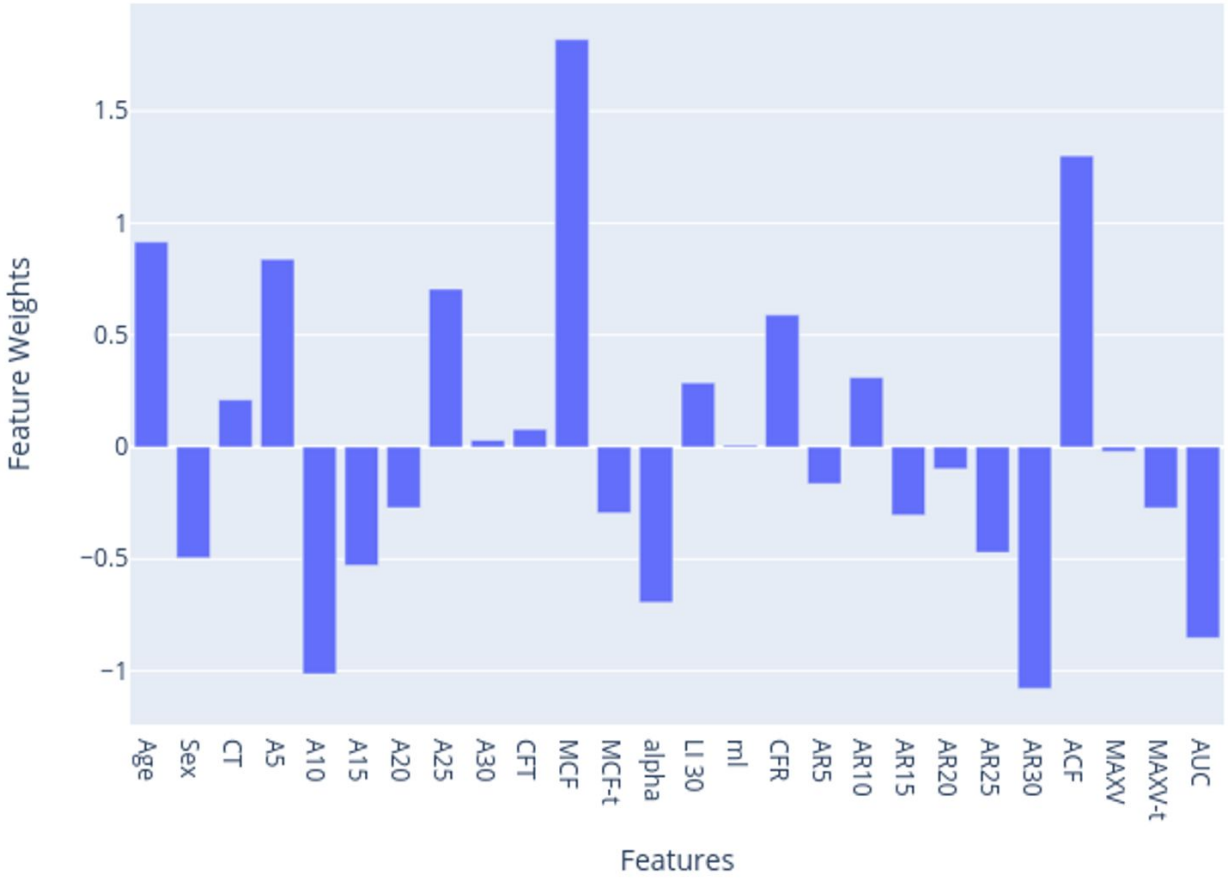
Projection and Clustering

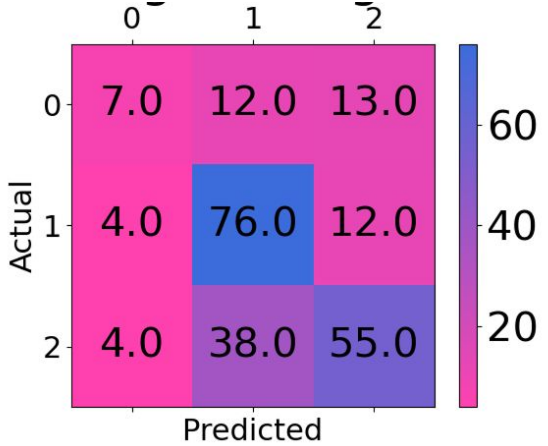
UMAP projection colored with (a): actual labels (b): K-Means clustering labels

Results: CLD NASH/Ethanol/Infectious (Multiclass)

Metrics	Logistic Regression	Decision Tree	SVM (linear kernel)	SVM (RBF Kernel)	XGBoost
Training Accuracy	58.75	57.41	59.10	60.02	65.74
Testing Accuracy	62.44	52.48	61.54	59.28	59.11
Precision	0.585	0..492	0.544	0.404	0.472
Recall	0.537	0.456	0.516	0.466	0.467
F-score	0.539	0.429	0.509	0.418	0.465
AUC-ROC	0.663	0.601	0.651	0.615	0.613

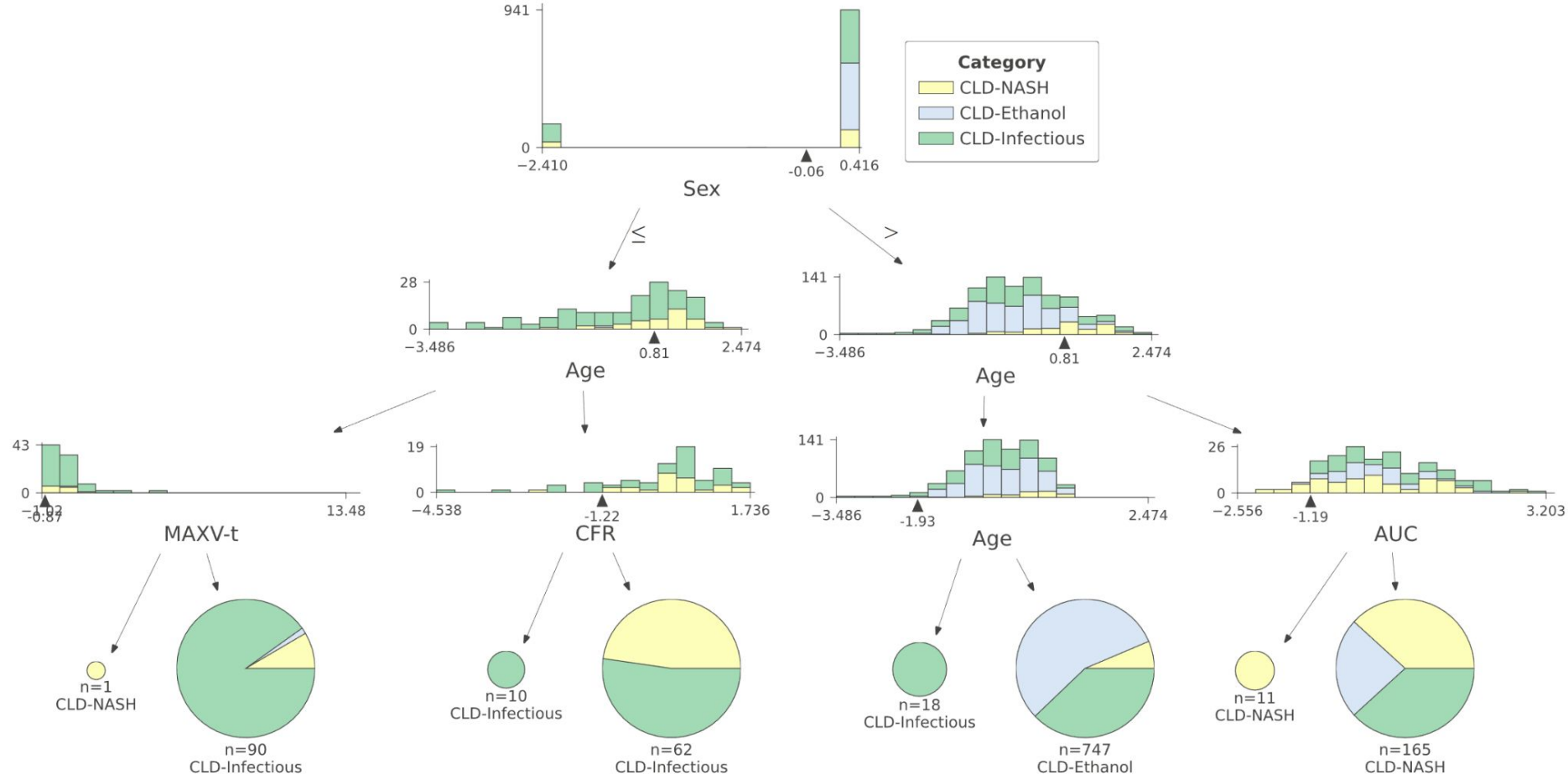
Logistic Regression





0 - CLD-NASH
1 - CLD-Ethanol
2 - CLD-Infectious

Decision Tree: Splits Visualized



Parser to Extract Information from EMRs

SciSpacy's **en_ner_bc5cdr_md** pre-trained model used for entity recognition.

- Symptom Recognition

Chief Complaint

Fever since past since past 15 days

Non productive cough since past 15 days

Altered sensorium since past 2 days

```
{'bc5cdr': {'Fever': 'DISEASE'}, 'craft': {}, 'nlp': {'past 15 days': 'DATE'}}
{'bc5cdr': {'cough': 'DISEASE'}, 'craft': {}, 'nlp': {'past 15 days': 'DATE'}}
{'bc5cdr': {}, 'craft': {}, 'nlp': {'past 2 days': 'DATE'}}
```

- # Extraction of Information from Patient History

History

Mrs XXX XXXX 50 yrs old female who is a known hypertensive , and had hypothyroidism , non diabetic had an index issue in the form of fever since past 15 days which was high grade , intermittent associated with gernalised body aches (Backaches +) . patient also complaint of non productive cough since past 15 days . later patient also developed bleeding PV since past 2-4 days , fever settled . following that patient again developed cough and also he had episode of altered sensorium since past 2 days. patient was evaluated outside and was found to have dengue NS1 postive and typhoid IgM positive . Patient was admitted and was being managed conservatively with RDPC, iv fluids and other supportive measures . patient came to ILBS with above mentioned complaints for further evaluation and management . There is no , vomiting, abdominal pain, altered bowel habits, hematemesis, and malena, burning micturition or decreased urine output. There is no h/o any intoxications, indigenous medications, major surgeries, blood transfusions or IV drug abuse prior to onset of the disease. There is no h/o CAD/TB/COPD.

{'bc5cdr': {'hypertensive': 'DISEASE', 'hypothyroidism': 'DISEASE', 'diabetic': 'DISEASE', 'fever': 'DISEASE', 'aches': 'DISEASE', 'cough': 'DISEASE', 'bleeding': 'DISEASE', 'dengue': 'DISEASE', 'typhoid': 'DISEASE', 'vomiting': 'DISEASE', 'abdominal pain': 'DISEASE', 'hematemesis': 'DISEASE', 'malena': 'DISEASE', 'drug abuse': 'DISEASE'}, 'craft': {'PV': 'GGP', 'drug': 'CHEBI'}, 'nlp': {'50 yrs old': 'DATE', 'past 15 days ': 'DATE', 'past 15 days': 'DATE', 'PV': 'ORG', '2-4 days': 'DATE', 'past 2 days': 'DATE', 'RDPC': 'ORG', 'malena': 'GPE', 'CAD/TB/COPD': 'ORG'}}

Next Steps

- Use NLP models to complete the information of remaining 2643 patients using EMRs.
- Test the performance of the current models on these new samples and retrain if required.
- Perform bleeder vs non-bleeder classification.

References

- Website. Available: Sharma A, Nagalli S. Chronic Liver Disease. [Updated 2022 Jul 4]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554597/
- Kamath PS, Kim WR. The model for end-stage liver disease (MELD). Hepatology. 2007;45. doi:10.1002/hep.21563
- Engelmann C, Thomsen KL, Zakeri N, Sheikh M, Agarwal B, Jalan R, et al. Validation of CLIF-C ACLF score to define a threshold for futility of intensive care support for patients with acute-on-chronic liver failure. Crit Care. 2018;22: 1–8.
- Understanding MELD Score for Liver Transplant. In: UPMC | Life Changing Medicine [Internet]. [cited 23 Sep 2022]. Available: https://www.upmc.com/services/transplant/liver/process/waiting-list/meld-score
- Yu Z, Zhang Y, Cao Y, Xu M, You S, Chen Y, et al. A dynamic prediction model for prognosis of acute-on-chronic liver failure based on the trend of clinical indicators. Sci Rep. 2021;11: 1–13.
- Younossi ZM, Stepanova M, Afendy M, Fang Y, Younossi Y, Mir H, et al. Changes in the prevalence of the most common causes of chronic liver diseases in the United States from 1988 to 2008. Clin Gastroenterol Hepatol. 2011;9. doi:10.1016/j.cgh.2011.03.020