Predicting Survival Status of Patients with Liver Cirrhosis

Matt Wang, Lena Wang, Yiyu Yao, Chuan Lin

Outline

1. Cirrhosis Overview
2. Dataset and Challenges
3. Data Preprocessing
4. Feature Engineering
5. Model Selection and Results
6. Feature Importance
7. Conclusion and Future Work

Cirrhosis Overview

- Cirrhosis: A chronic liver disease with high mortality.
- Traditional diagnostics are invasive and costly.
- Aim: Use machine learning to determine the effectiveness of D-penicillamine and to predict the survival status of patients.

Dataset Description

Cirrhosis Dataset Overview

- 424 patient records from Mayo Clinic (1974–1984).
  - 312 patients in clinical trial (112 did not join but agreed to record basic metrics)
  - 6 patients dropped out
- Features: Demographics, clinical metrics, and survival data.

Dataset Challenges

Limitations in Data

- Small dataset: 418 usable records (424 - 6 that dropped out).
- Missing data: $\sim 33\%$ of values missing (mainly due to 106 patients not participating in the trial).

Missing Data Handling

Analysis and Solution

Method Evolution

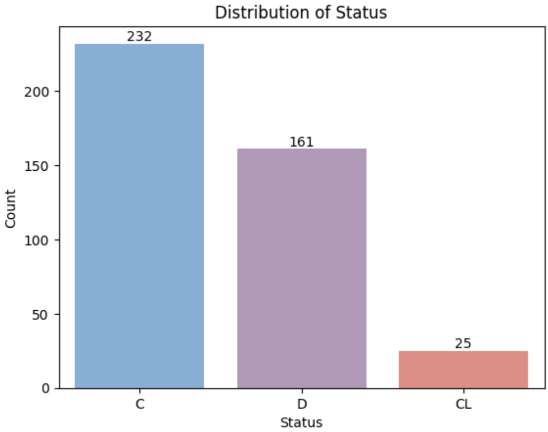
- Simple deletion → Data loss
- Median/Mode → Bias
- EM only → Limited
- Combined approach ✓
  - EM
  - Proportional & Statistical Imputation

```
ID                0
N_Days            0
Status            0
Drug            106
Age               0
Sex               0
Ascites         106
Hepatomegaly    106
Spiders         106
Edema             0
Bilirubin         0
Cholesterol     134
Albumin           0
Copper          108
Alk_Phos        106
SGOT            106
Tryglicerides   136
Platelets        11
Prothrombin       2
Stage             6
```

Categorical Encoding

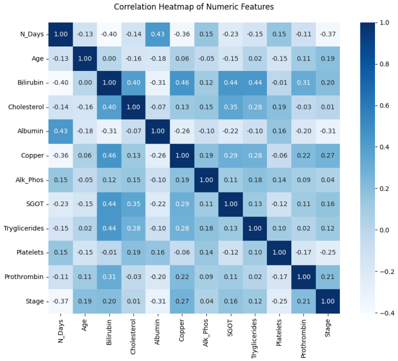
Transforming Features

- One-hot encoding applied to categorical variables.
- Survival status simplified to binary: 0 (Alive), 1 (Deceased).



Feature Engineering

New Features Introduced



- **DiagnosedDay**
- **Age Group**
- **BA Ratio**
- **CA Ratio**
- **RiskScore**
- **Liver Complication Index**

Models Evaluated

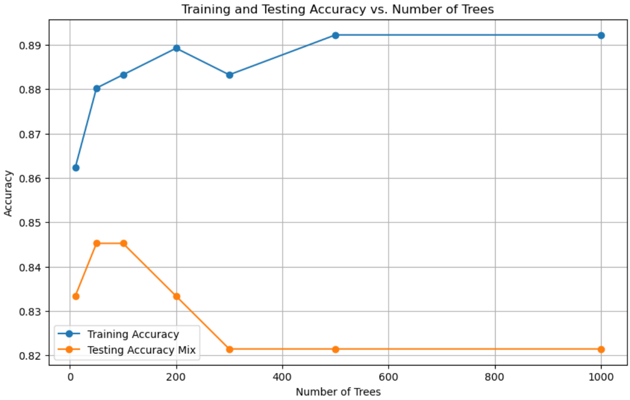
Machine Learning Techniques

- Random Forest (RF)
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Multilayer Perceptron (MLP)

Random Forest Results

Optimized Tree Numbers

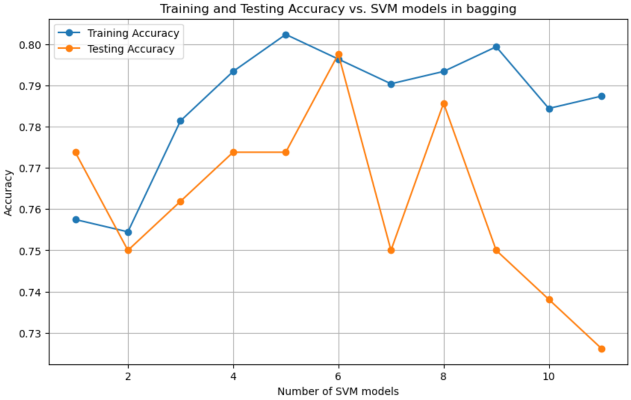
- Best performance: 21 trees (Accuracy: 87%).
- Higher trees led to overfitting.



SVM Results

with Bagging and Kernel

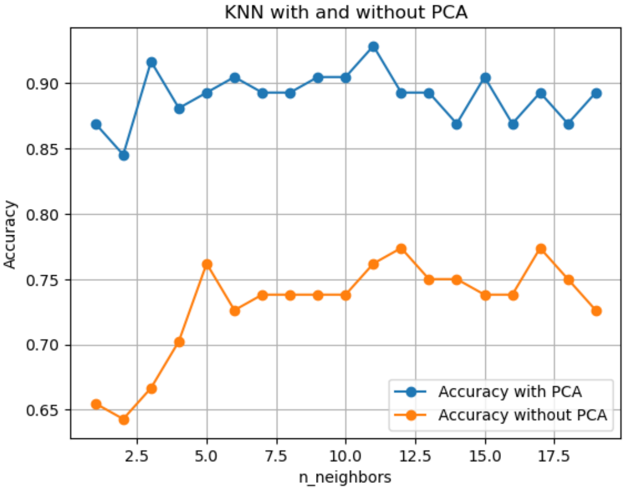
- Best accuracy: 79.8% with linear kernel.
- Non-linear kernels reached an even lower accuracy (rbf = 61.9%).



KNN Results

with/withoPCA

- PCA improved accuracy to 92.8% with k=11.



MLP Results

Neural Network Findings

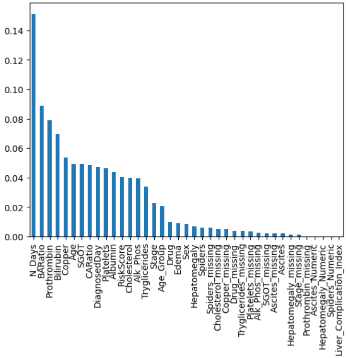
- Optimal learning rate: 0.001.
- Simple architectures performed better: Single layer (128 units).

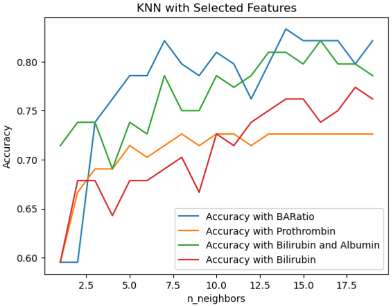
hidden_units	learning_rate	test_accuracy
[64]	0.001	0.761905
[64]	0.010	0.730159
[64]	0.100	0.746032
[128]	0.001	0.809524
[128]	0.010	0.777778
[128]	0.100	0.746032
[64, 64]	0.001	0.809524
[64, 64]	0.010	0.761905
[64, 64]	0.100	0.682540
[128, 64]	0.001	0.761905
[128, 64]	0.010	0.730159
[128, 64]	0.100	0.761905

Feature Importance Analysis

Key Insights

- Drug feature (D-penicillamine) has a minimal correlation with the survival status.
- Top three informative features: BA Ratio (Bilirubin/Albumin), Prothrombin, Bilirubin.





Conclusion

Key Takeaways

- KNN achieved highest accuracy (92.8%).
- D-penicillamine had limited effect on cirrhosis treatment.
- BA Ratio as a cost-effective and non-invasive predictor.

Limitations and Future Work

Next Steps

- Expand dataset to improve robustness.
- Validate BA Ratio in real-world clinical settings.
- Using prediction results to generate the probability of death