Decision Tree Methodology for Electronic Health Record (EHR) Clinical Data Endpoints

Kale Kasdorf BSc, Dillon Chrimes PhD
Health Information Science, University of Victoria, Victoria, BC, Canada

Introduction

EHR data plays an important role in clinical decision-making. In health informatics, the combination of human expertise and data-driven support tools should improve decision-making beyond what either can achieve alone. Artificial intelligence (AI) tools have the potential to increase productivity and support decision-making in clinical settings. One approach is through decision trees, which provide interpretable models for predictions.

To test the use of decision trees in health informatics, EHR datasets are required. Massachusetts Institute of Technology (MIT) MIMIC-IV eICU collaborative database is an openly available dataset that spans across 20 hospitals between 2014-2015. The dataset contains over 35,000 deidentified patient encounters, with charting information from electronic health record systems.

Recent studies have used machine learning to predict ICU outcomes, but there is limited exploration of the use of decision trees for this purpose. Therefore, this study investigates the application of decision trees on the eICU dataset to address the following research questions:

- Can decision trees be created from the eICU dataset?
- What methodology is needed to produce usable decision tree models?
- Can the eICU dataset be used to predict ICU mortality?

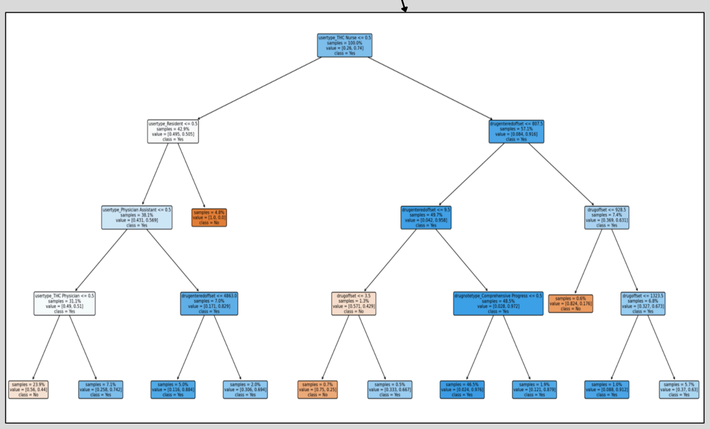
Methods

To address the research questions, we established a series of methodological steps over a five-month period, starting in October 2025. These methods determined the feasibility of using AI decision trees. We identified and explored which tables and variables in the eICU data are useful for clinical or operational decisions.

The MIT eICU dataset contains:

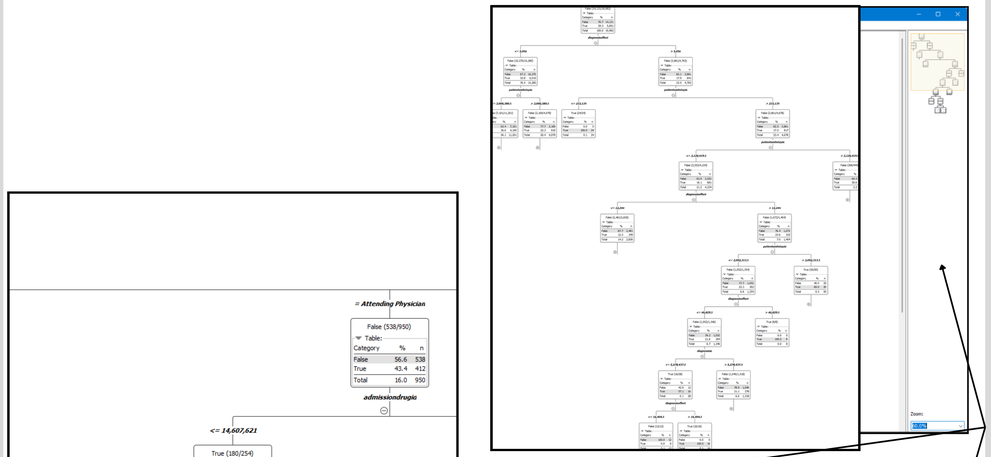
- 30 tables within a relational database
- A database schema describing all tables and variables (https://eicu-crd.mit.edu/)
- Variables derived from EHR modules cleaned and curated by the MIT research group

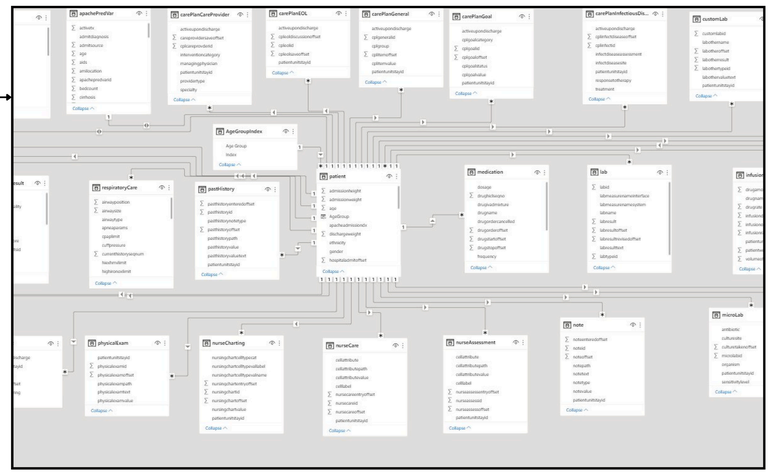
We reviewed the past four years of dashboard visualizations of this eICU dataset in our undergraduate course HINF 310 to verify which tables and variables showed trends. Then we used these identified trends to test the creation of decision trees in ChatGPT.

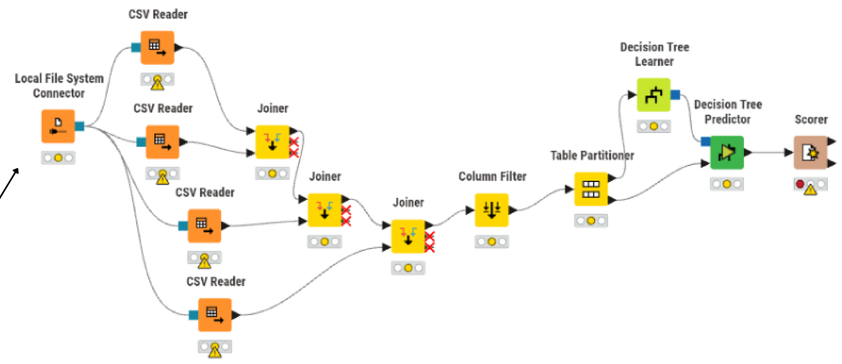


Next, decision tree models were developed in the KNIME analytics platform. Tables were joined to combine data ad the decision tree learner node was used to generate decision tree models.

EHR Data-Driven Decision Trees

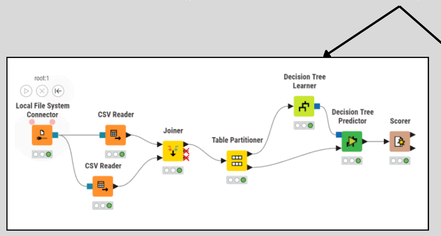


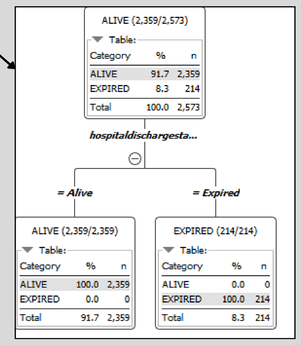




Results

From 30 tables, appropriate decision trees were derived from 10 tables (e.g. admissionDrug, apachePatientResult, patient, diagnosis, treatment, etc.). Early prediction results were strong, models produced accuracy of 96%+ and error up to 3.13%.





Despite these results, there was evidence of overfitting, skewed decision trees, and a lack of clinical or operational significance. Decision trees ranged from 2-10 nodes in both depth and width.

Inspecting the decision tree showed splitting based on factors that were related to health outcomes and many were not. For example, the unit type or user type, diagnosis offset, Rxincluded, ProgressNotes included were all significant threshold in the trees. However, offset times did repeat often in the trees and sometimes even the patient unit stay ID, which should not be a prediction factor. Furthermore, there was a noticeable quality difference between the trees produced by predicting binary values vs continuous/multi-class values.

Finally, implementing correct table joins quickly scaled the data and improved results. A combination of four tables combine to produce 7.8 million rows with 55 columns. Further filtering of the data significantly improved results and prediction accuracy. One model was able to predict ICU mortality with over 95% accuracy and around 6% error. Although these results are worse than initially, these models are no longer overfitting and are much more realistic prediction models.

Discussion

Establishing practicality and methodology was easier and quicker than initially thought. Decision trees could be created within seconds. However, many of the trees produced in ChatGPT and KNIME were either useful for decision-making or not useable at all. Therefore, we establish an initial pipeline for using ICU data from EHRs, and incrementally adjusted and improved models throughout the process. This represents the first step in assessing the feasibility of decision trees in this context.

Further data wrangling and validation of trends and generated decision trees are required to ensure reliable predictions, particularly in clinical settings. While health outcomes and organizational insights can be extracted from the dataset, the current decision trees alone are limited. Further organization and manipulation of the data and features is necessary to binary decision trees capable of reliably predicting ICU mortality from EHR data.

References

This research was supported by the Jamie Cassels Undergraduate Research Awards, University of Victoria.
Supervised by Dr. Dillon Chrimes, Health Information Science.

Questions? Contact kalekasdorf@uvic.com

Produced: March 6, 2026