

Supplementary Document

Comparative Analysis of Machine Learning Models for Predicting Hospital- and Community-Associated Urinary Tract Infections Using Demographic, Hospital, and Socioeconomic Predictors

Arash Arjmand¹, Majid Bani-Yaghoub^{1*}, Gary Sutkin², Kiel Corkran¹, Susanna Paschal³

¹Division of Computing, Analytics and Mathematics, School of Science and Engineering, University of Missouri-Kansas City, Kansas City, MO 64110, USA.

²Department of Biomedical and Health Informatics, School of Medicine, University of Missouri-Kansas City, Kansas City, MO 64108, USA

³University Health, Kansas City Hospital 2301 Holmes Street, Kansas City, MO 64108, USA

* Corresponding author baniyaghoubm@umkc.edu

Section S1. Descriptive Statistics

According to Table S1, there are notable differences between (CAI) and (HAI) from 2019 to 2023. HAI patients are consistently older, with a longer length of stay (LOS). HAI cases had a lower percentage of female and married patients compared to CAI. Additionally, a higher proportion of HAI cases occurred at UH Truman Medical Center and were associated with inpatient nursing units.

Table S1. Descriptive statistics of 2019-2023 data organized according to patient demographics, hospital characteristics, and socioeconomic factors. Median values are reported for each group across the years, with minimum and maximum values presented in parentheses.

Model	HA-UTI	CA-UTI	P-value
Average Age	63.5 (58.9, 65.3)	49.7 (48.3, 51.3)	<0.01
Female (%)	65.7 (61.6, 71.0)	84.7 (84.0, 86.4)	Varies
Married (%)	16.4 (12.5, 20.7)	23.6 (22.6, 25.4)	<0.01
Employed (%)	99.7 (92.4, 100)	97.9 (92.1, 98.2)	Varies
Hispanic (%)	8.5 (6.3, 10.1)	9.6 (8.3, 11.0)	Varies
Length of Stay	20.1 (16.2, 27.0)	0.6 (0.5, 0.6)	<0.01
Nurse Unit (%)	58.7 (53.9, 63.7)	4.0 (2.9, 4.4)	<0.01
Location ² (%)	72.9 (62.7, 74.2)	30.1 (29.0, 36.2)	<0.01
Median Income	50,180.1 (49,141.7, 51,620.2)	49,960.0 (49,350.5, 51,988.9)	Varies
Poverty (%)	18.7 (18.1, 19.3)	18.2 (17.6, 18.7)	Varies
Graduate Degree (%)	7.9 (7.5, 8.2)	7.8 (7.7, 8.3)	Varies
Population Density ⁴	2,870 (2,760, 2,936)	2,900 (2,735, 3,075)	Varies

Section S2. Model Performance

This section includes the model performances with respect to positive cases of urinary tract infection in hospitals and community. Table S2 provides a comparison between machine learning models used in this study.

Table S2. Evaluation Results of Decision Tree (DT), Neural Network (NN), Logistic Regression (LR), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) HA-UTI vs. CA-UTI.

Metrics	DT	NN	LR	RF	XGBoost
Overall Accuracy	91.0 (89.8, 94.2)	92.4 (89.8, 93.5)	93.2 (91.3, 94)	92.2 (91.2, 94)	92.6 (90, 94.2)
Sensitivity	95 (92, 96)	88 (84, 92)	93 (89, 100)	94.7 (93.4, 95.6)	94.7 (88.9, 94.7)
Specificity	91.7 (89.4, 94.2)	93.4 (89.8, 94)	92.6 (91.7, 93.3)	91.9 (90.4, 94.1)	92.3 (90.2, 94.2)
Precision	56 (53, 66)	61 (56, 62)	64 (59, 66)	60.2 (55.4, 65.7)	61.5 (57.7, 65.4)
F1 Score	70.7 (68, 77.2)	72.1 (69.6, 72.7)	75.8 (73.7, 78)	73.2 (69.9, 77.2)	72.7 (71.7, 76.5)
Cross Validation	95.2 (92.9, 95.9)	NA	NA	96.2 (95.6, 97.4)	96 (95.5, 97.3)
AUC	94.3 (93.7, 96.9)	89.9 (88.7, 91.9)	96 (95, 97)	96 (95.1, 97.6)	95.7 (95, 97.9)

Note. The median values have been indicated for models applied to the years 2019-2023. The values inside the parenthesis are the upper and lower bounds.

Section S3 Logistic Regression

Table S3 summarizes the stepwise logistic regression estimates for comparison between HA-UTI and CA-UTI across years 2019-2023

Table S3. Odds Ratios (OR) of HA-UTI versus CA-UTI obtained from the stepwise logistic Regression (LR) across various years. The values are associated with different units and facility locations.

Predictors	2019	2020	2021	2022	2023
Inpatient ¹	1.89	—	0.34	1.83	
Other ¹	0.01	0.03	0.00	0.01	0.00
Mental Health ¹	45.32	241.70	—	22.68	—
Rehabilitation ¹	—	0.20	0.18	—	—
Surgery ¹	—	0.20	0.15	—	—
UHTMC ²	2.64	17.63	—	—	—
UHLMC ³	—	31.19	—	—	0.25
Older Age	1.64	1.41	1.42	—	—

¹Nurse Unit, ²University Health Truman Medical Center (UHTMC), ³University Health Lakewood Medical Center (UHLMC)

Section S4. Normalized Importance

Table S4. Normalized Importance Factors for 2019-2023 HA-UTI

2019	2020	2021	2022	2023
Other ¹ (86.6)	Other ¹ (58.2)	Other ¹ (93.6)	Other ¹ (77.4)	Other ¹ (90.7)
Gender-Male (3.3)	Income ² (22.7)	UHTMC ³ (2.1)	Age (6.7)	UHLMC ⁴ (3.0)
Age (2.2)	UHTMC ³ (7.7)	Density ⁵ (1.1)	Gender-Male (3.4)	Gender-Male (1.6)
Ethnicity-NH ⁶ (1.3)	Age (2.9)	MS ⁷ -Single (1.1)	Density ⁵ (3.2)	Poverty (1.5)

¹Nurse Unit, ³University Health Truman Medical Center (UHTMC), ⁴University Health Lakewood Medical Center (UHLMC), ³Percent Graduate Degree, ⁵Population Density, ²Median Household Income, ⁶ Non-Hispanic,

⁷Marital Status

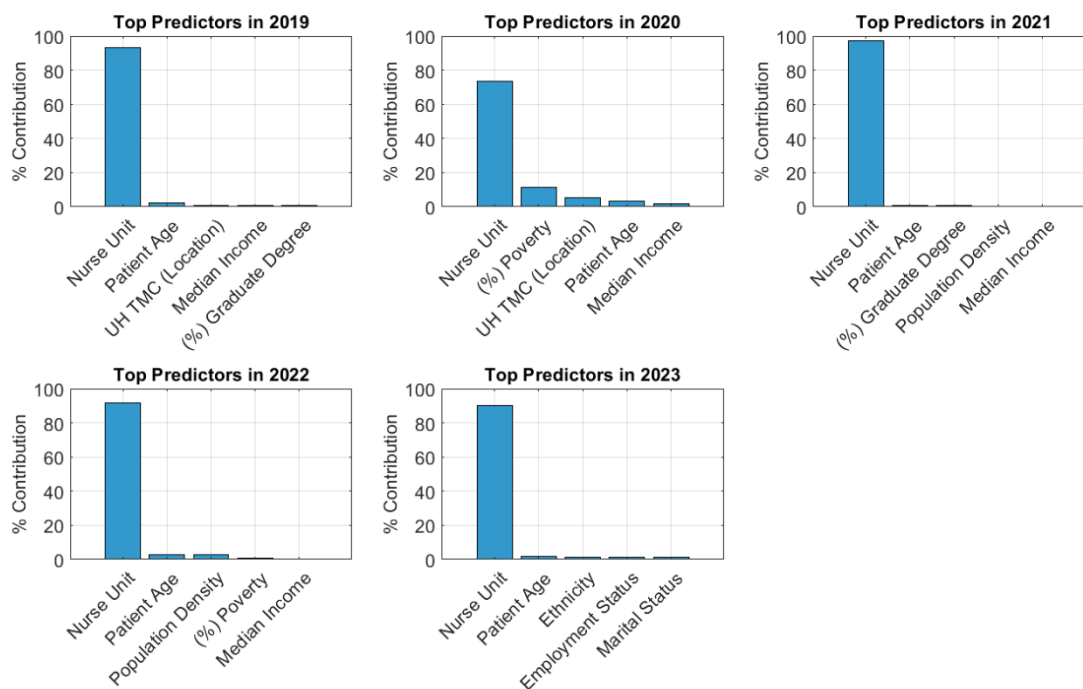


Figure S1. Normalized Importance Factors for HA-UTI vs CA-UTI 2019-2023

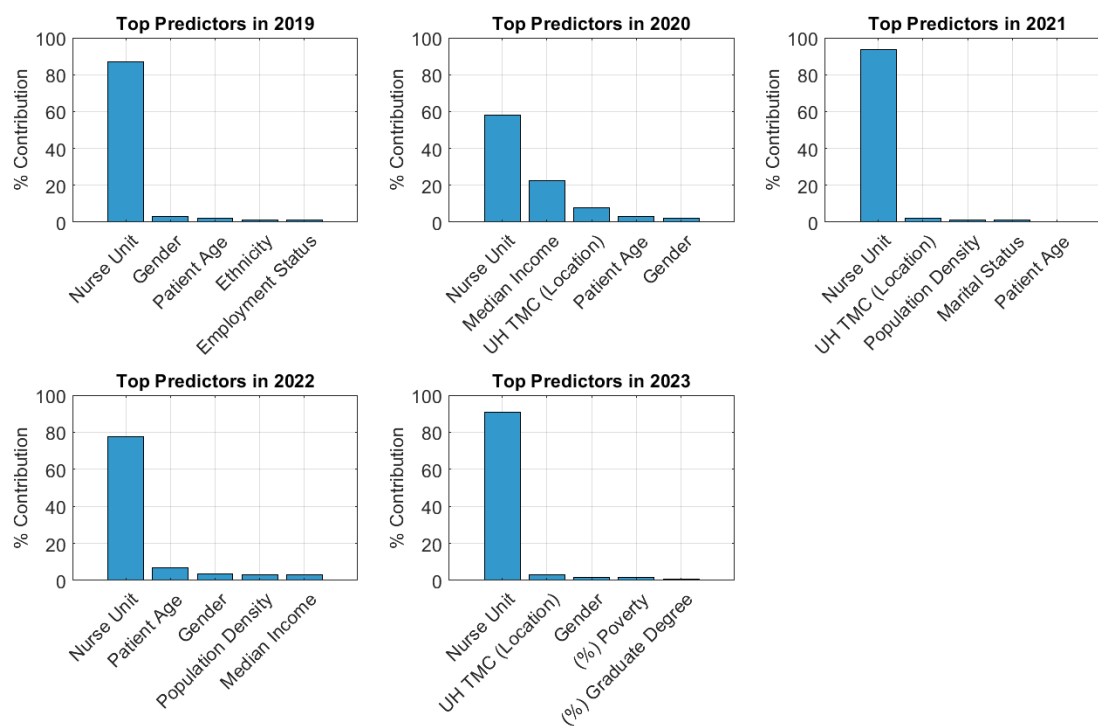


Figure S2. Normalized Importance Factors for HA-UTI versus those without UTI (2019-2023)

Section S5. Decision Trees

This section presents simplified versions of the decision tree diagrams generated from classification and regression tree (CART) modeling. In these trees, each box (node) is proportional in size to the number of samples that reach that point, representing how much of the population belongs to each group. The colors in the tree indicate the dominant class in each node: blue represents healthcare-associated UTI (HAI), and orange/brown represents community-associated UTI (CAI).

Each rectangle (node) in the decision tree contains the following key information:

1. Split Condition (Top): This shows the feature and threshold used to divide the data at that point (e.g., "Nurse Unit_Other $\leq$ 0.99"). If the condition is true, you follow the left branch; if false, you go right.
2. Gini Index: This measures the impurity or diversity of the samples within the node. A Gini index of 0 indicates that all cases belong to a single class (pure), while higher values show a mix of classes. For example, a Gini index of 0.44 indicates a moderate mix of HAI and CAI.
3. Percentage of Total Samples: This shows the percentage of the entire dataset that reaches that node. For example, a node with 51.2% means that over half of the total samples in the dataset reached that point in the tree.
4. Class Distribution (Brackets): This gives the proportions of CAI and HAI cases at the node. For example, [0.03, 0.97] means that 3% of the cases are CAI, and 97% are HAI.

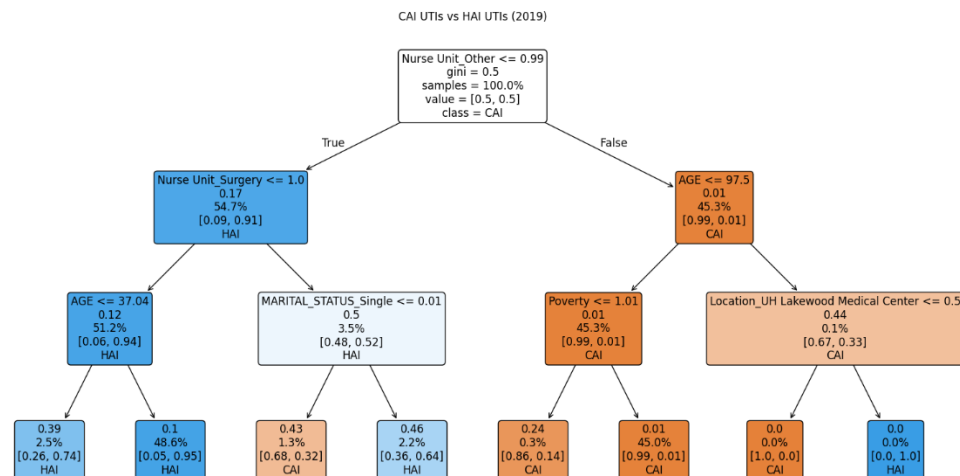


Figure S3: Decision Tree for CAI UTI vs HAI UTI (2019)

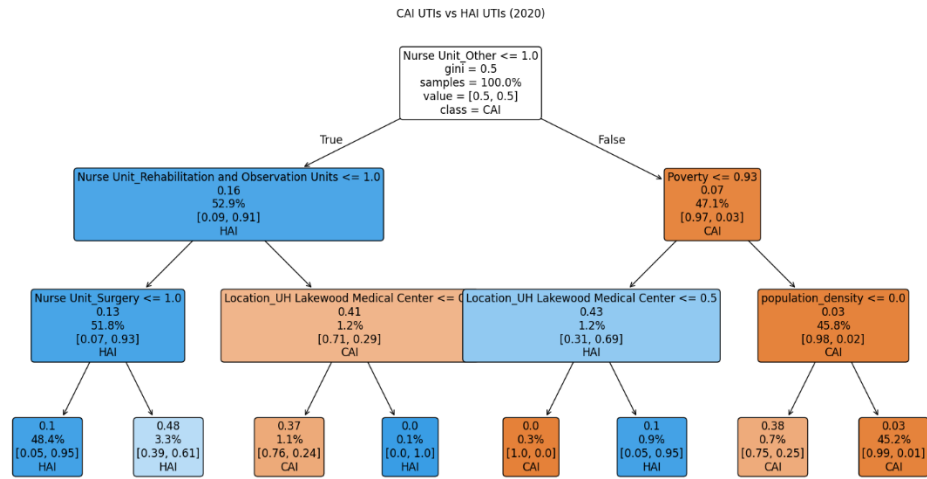


Figure S4: Decision Tree for CAI UTI vs HAI UTI (2020)

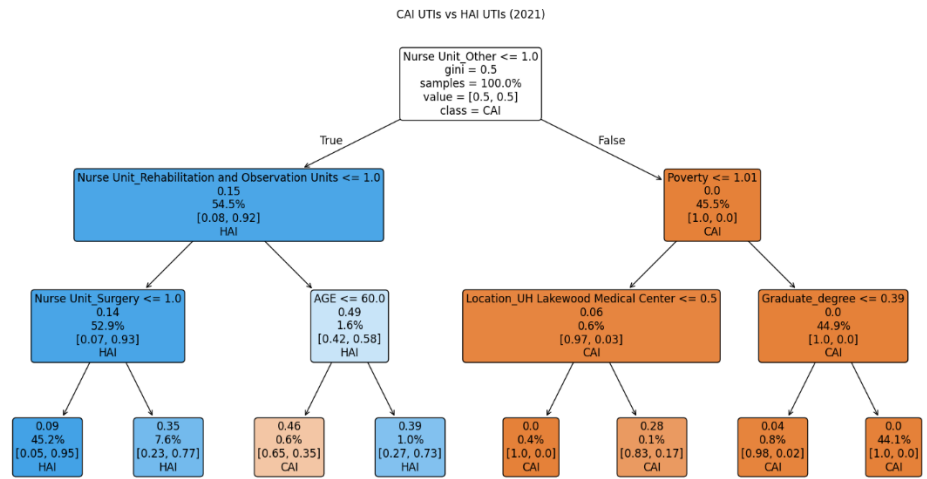


Figure S5: Decision Tree for CAI UTI vs HAI UTI (2021)

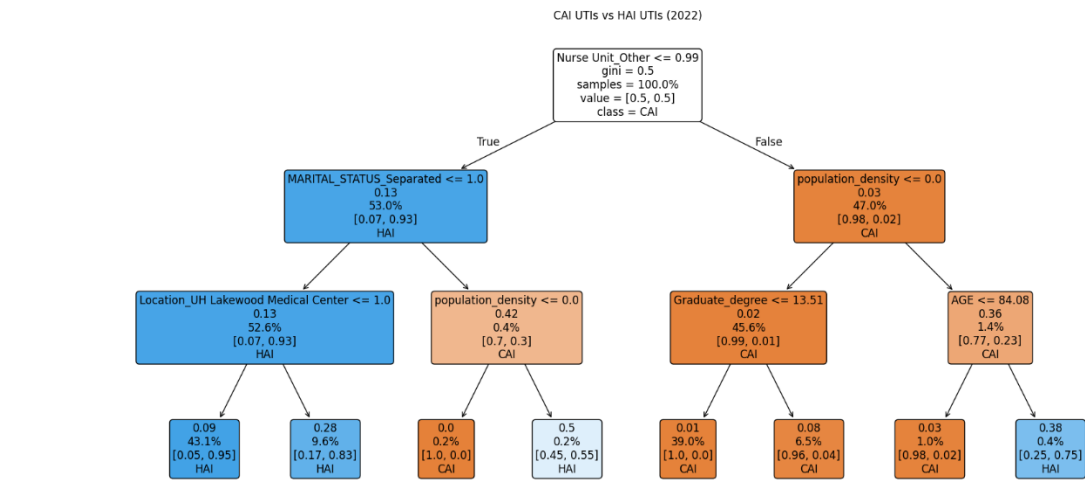


Figure S6: Decision Tree for CAI UTI vs HAI UTI (2022)

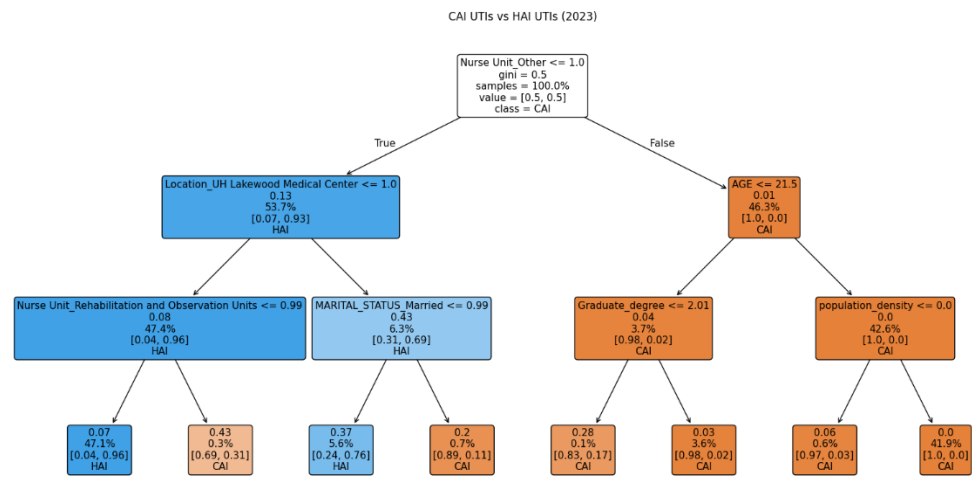


Figure S7: Decision Tree for CAI UTI vs HAI UTI (2023)

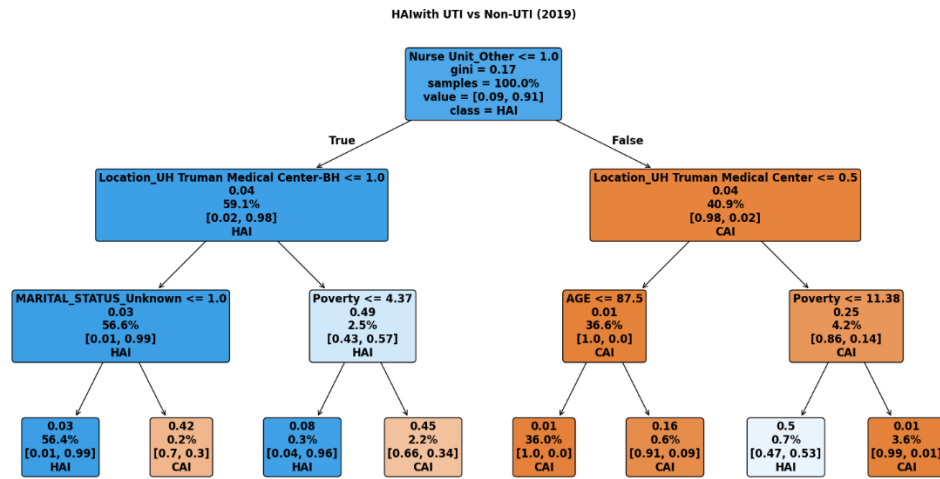


Figure S8: Decision Tree for HA-UTI vs non-UTI (2019)

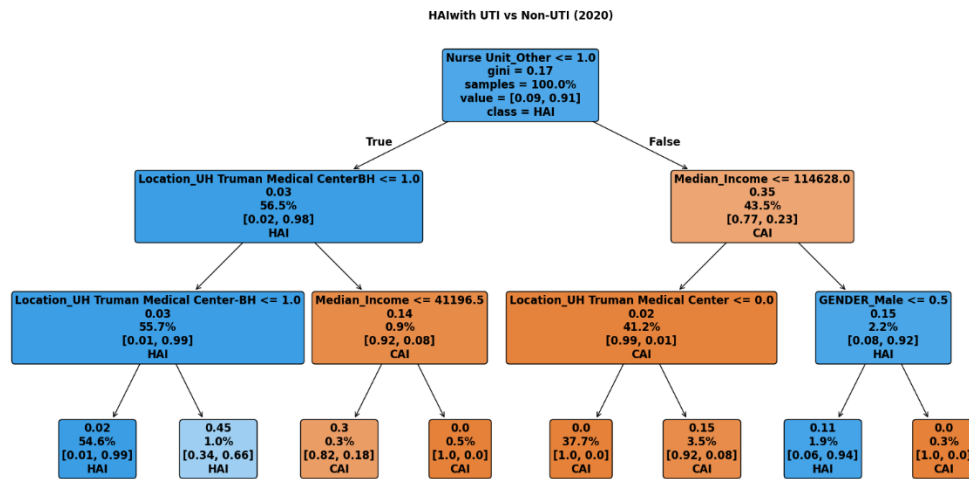


Figure S9: Decision Tree for HA-UTI vs non-UTI (2020)

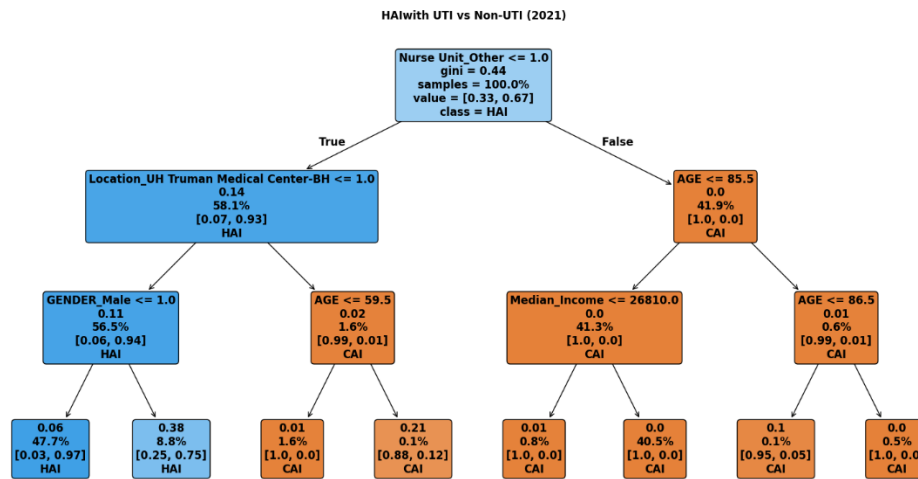


Figure S10: Decision Tree for HA-UTI vs non-UTI (2021)

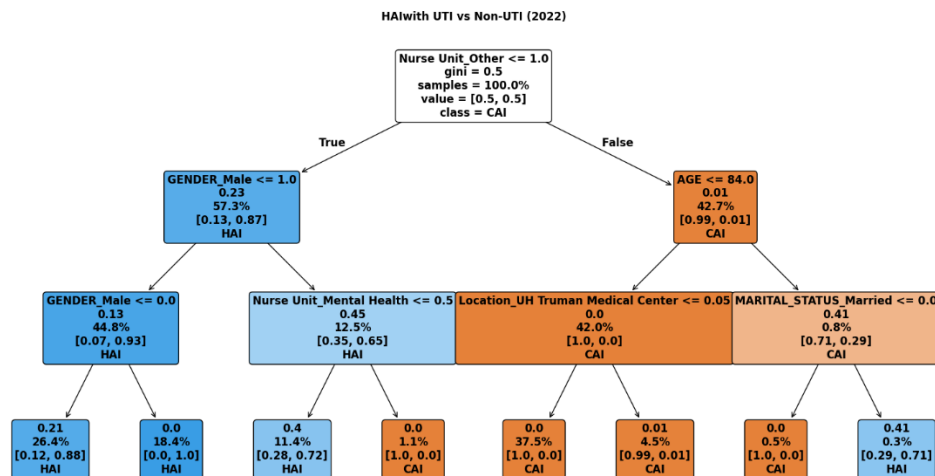


Figure S11: Decision Tree for HA-UTI vs non-UTI (2022)

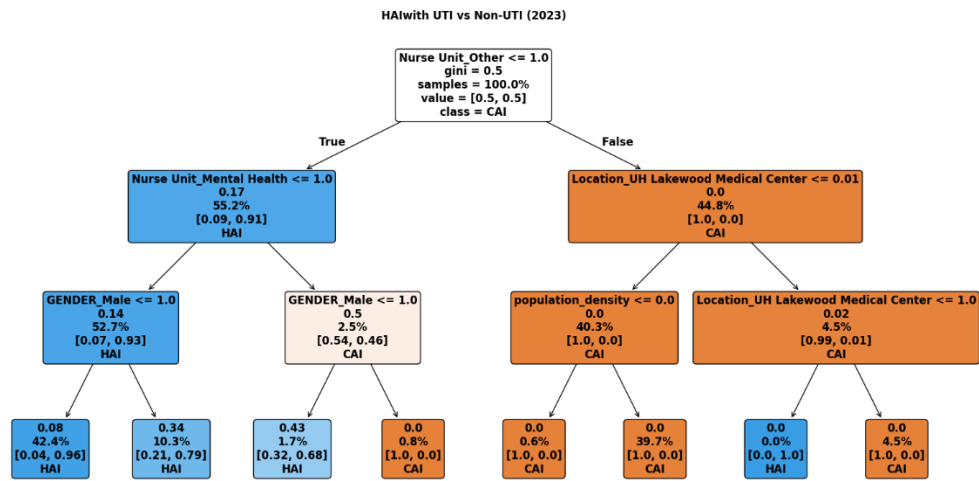


Figure S12: Decision Tree for HA-UTI vs non-UTI (2023)