Lab 1 - ciTBI Data, STAT 214, Spring 2025

2025-02-21

1. Introduction

Traumatic brain injury (TBI) is a leading cause of death and disability in children worldwide. In the United States, pediatric head trauma results in approximately 7,400 deaths and more than 600,000 emergency department visits each year[1]. Although computed tomography (CT) scans are the gold standard for diagnosing TBI, overuse of CT carries a risk of radiation-induced malignancies[1]. To reduce unnecessary radiation exposure, clinical decision rules (CDRs) have been developed to identify children at very low risk for clinically important TBI (ciTBI) who may not require CT imaging.

The PECARN TBI dataset, derived from Kuppermann et al. (2009), provides a large-scale study of pediatric head trauma that incorporates key clinical indicators for assessing the risk of ciTBI[1]. In this report, we perform an exploratory data analysis (EDA) on the dataset to understand its structure and assess data quality issues. We first describe the dataset and its variables and then identify and address inconsistencies such as missing values and measurement error. Finally, the goal is to refine data preprocessing techniques to improve clinical decision making in pediatric TBI screening.

2. Data

The dataset used in this analysis comes from the Pediatric Emergency Care Applied Research Network (PECARN) Traumatic Brain Injury (TBI) study, which aims to develop clinical decision rules (CDRs) for identifying children at low risk of clinically-important TBI (ciTBI). The dataset includes a variety of clinical and demographic features recorded for pediatric patients presenting with head trauma.

In this report, we focus on the following key variables relevant to TBI risk assessment and clinical decision-making: **GCS Score, Severity of injury mechanism, Presence of loss of consciousness, vomit, headache, Altered mental status (AMS), and Clinically-important traumatic brain injury (ciTBI).**

2.1 Data Collection

The data were collected as part of a multi-center prospective cohort study conducted across 25 emergency departments in North America. Pediatric patients under 18 years old who presented with head trauma were enrolled within 24 hours of injury.

Each patient underwent standardized clinical evaluation, and information was recorded on variables such as **Glasgow Coma Scale (GCS) scores, presence of amnesia, seizures, headache, and mechanism of injury.** CT scans were performed at the discretion of emergency department physicians, and follow-up data were collected to determine whether a patient had **ciTBI (defined as requiring neurosurgery, intubation for** > 24 **hours, hospital admission for** > 2 **nights, or resulting in death).**

2.2 Data Cleaning

The dataset contained several issues, including a high number of missing values, inconsistencies in variable naming, and unexpected discrepancies in certain categorical values. Notably, some variables had 92 values that did not align with their corresponding 0 values. For example, ‘pos_ct’ should have 92 values equal to the sum of 92 and 0 values in ‘edct’, but this was not the case.

1. **Standardizing Variable Names**
    - All variable names were converted to snake_case for consistency.
    - ‘ha_verb’ was renamed to ‘ha’ for easier processing.
2. **Handling Missing Values**
    - Drop all data has ‘gcs_total’< 14
    - ‘amnesia_verb’ missing values were filled with 0 when ‘s_fx_bas’, ‘high_impact_inj_sev’, and ‘gcs_total’ were 0, 0, 15, respectively.
    - ‘loc_separate’ missing values were set to 0 when ‘act_norm’, ‘high_impact_inj_sev’, and ‘gcs_total’ were 0, 0, 15 or 1, 1, 15.
    - ‘seiz’ missing values were set to 0 when ‘s_fx_bas’, ‘high_impact_inj_sev’, ‘gcs_total’, and ‘finding4’ were 0, 1, 15, 0.
3. **Re-mapping Categorical Variables**
    - Based on the dataset documentation, values 0, 91, and 92 were reassigned according to their proper definitions.
4. **Further Missing Value Imputation**
    - ’amnesia_verb’ missing values were filled with 1 when ’s_fx_bas’ == 1.
    - ‘loc_separate‘ missing values were set to 1 when ‘high_impact_inj_sev‘, ’finding4’, and ’s_fx_bas’ were 3, 1, 1.

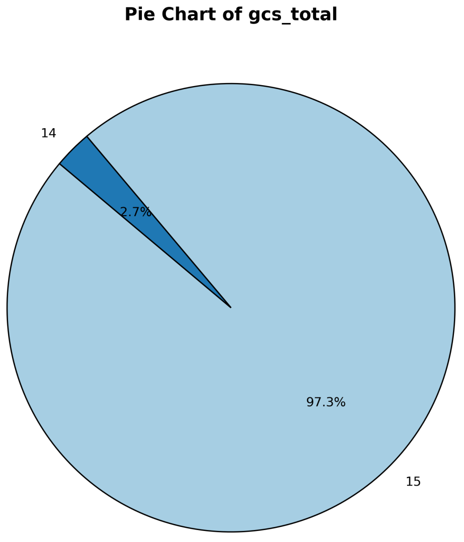
- ’seiz‘ missing values were set to 1 when ’loc_len‘, ’finding20‘, ’s_fx_bas‘, and ’finding4‘ were 4, 1, 1, 1.
- ’ams’ missing values were set to 1 when ’gcs_total’ == 14.
- Variables with a small proportion of missing values were handled by removing rows containing missing data rather than imputing them.

5. **Assigning ‘high_impact_inj_sev’ Based on ‘injury_mech’**
   - Following the dataset definition, ‘high_impact_inj_sev’ values were correctly assigned based on ‘injury_mech’ categories.

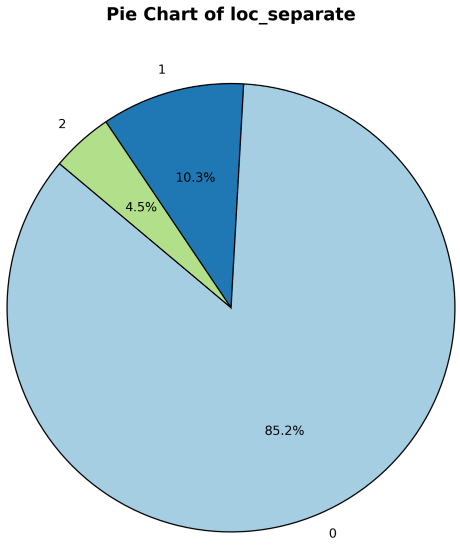
6. **Assigning ‘pos_ct’ Based on findings**
   - ‘pos_ct’ was updated based on whether any findings were present, ensuring consistency with the dataset documentation.

2.3 Data Exploration

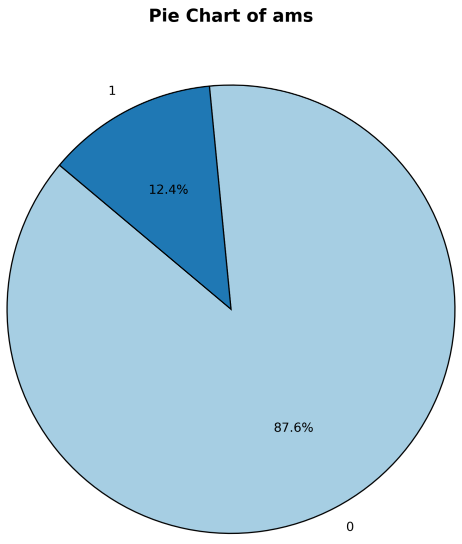
These graphs provide an overview of the six key variables identified as most relevant in the dataset. **In the visualizations, 0 stands for no, 1 stands for yes, and 2 stands for unclear/suspected.**



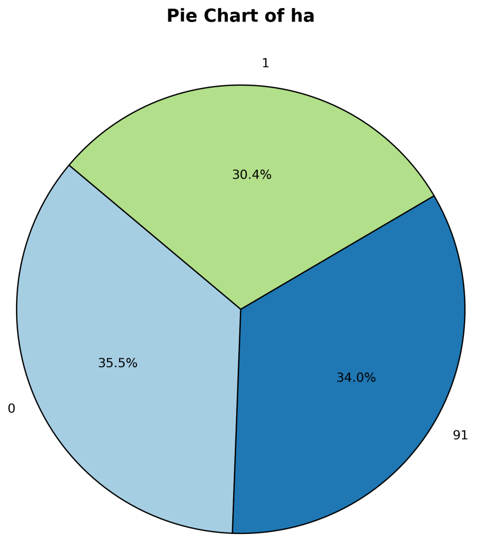
GCS Scores



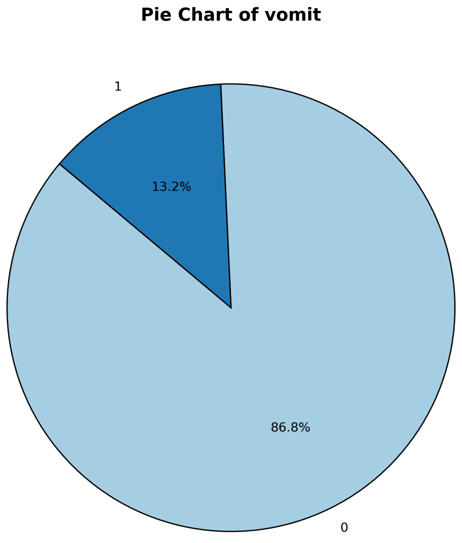
Loss of Consciousness



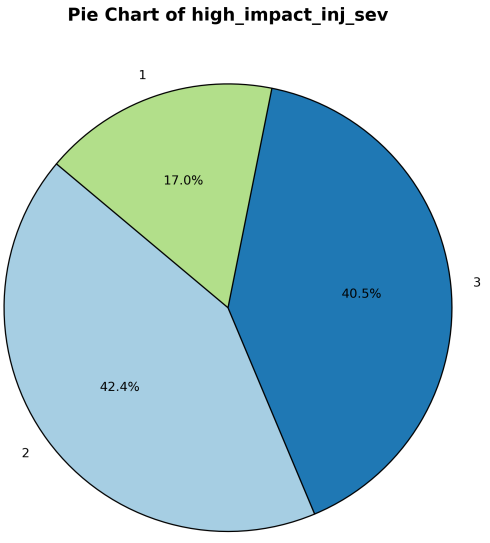
AMS



Headache



Vomit



Severity of Injury Mechanism

3. Findings

1. **First Findings**

   From the grouped bar chart below, AMS appears to be a significant factor in determining whether a CT scan is performed. Specifically, when AMS is 0, the majority of cases did not undergo a CT scan, whereas when AMS is 1, most cases received a CT scan. This suggests that AMS status plays a crucial role in clinical decision-making regarding CT utilization.

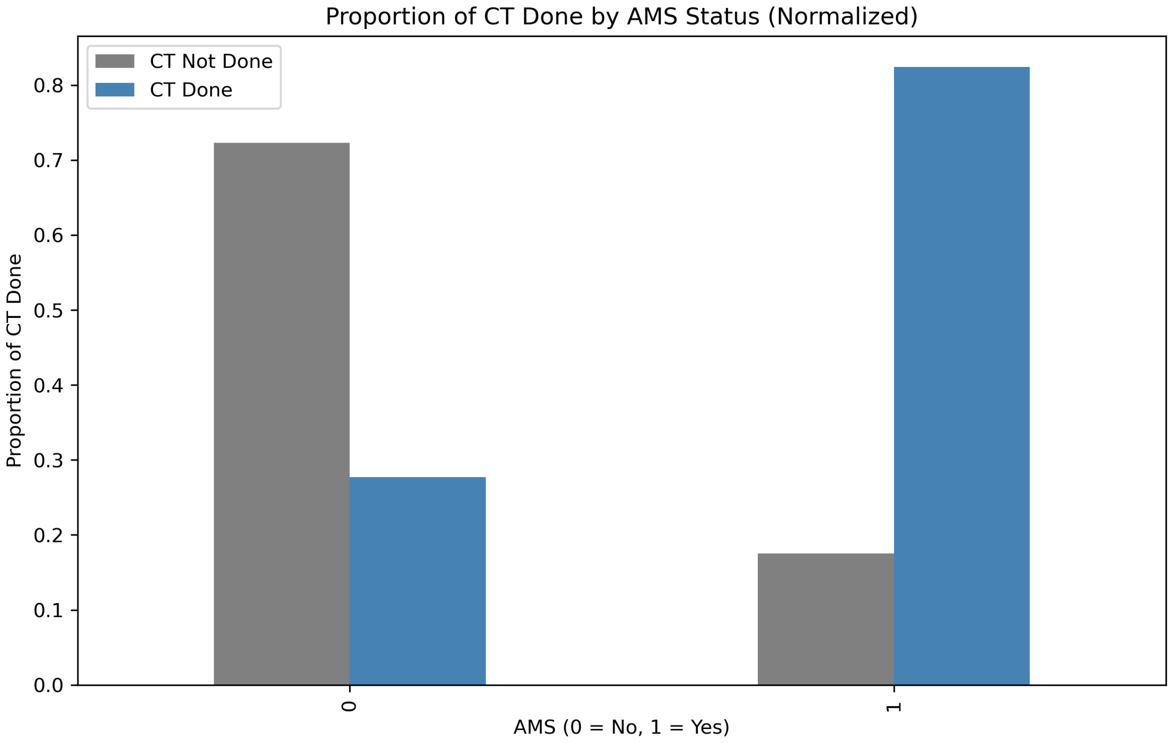
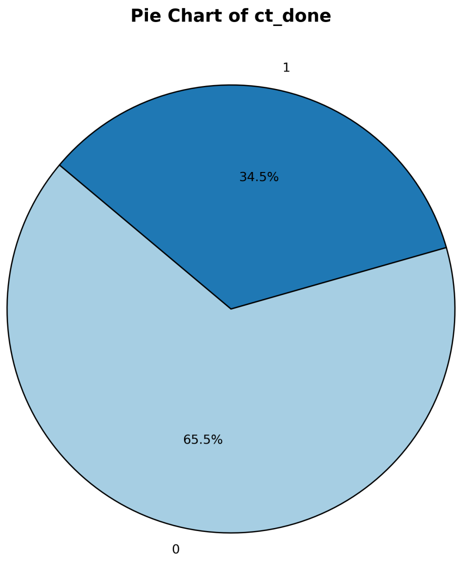


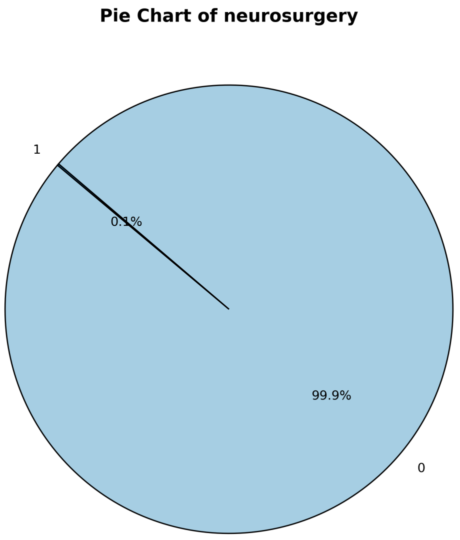
Figure 1: Grouped Bar Chart of AMS and CT Done

2. **Second Findings**

   Although only 0.1% of patients required neurosurgical intervention, 34.5% underwent a CT scan. This discrepancy really suggests a huge overutilization of CT imaging, where a significant proportion of scans may have been performed on patients who ultimately did not require acute intervention.



CT Done



Neurosurgery

3. Third Findings

The figure below indicates a strong relationship between GCS score and ciTBI, suggesting that lower GCS scores are associated with a higher likelihood of clinically important traumatic brain injury.

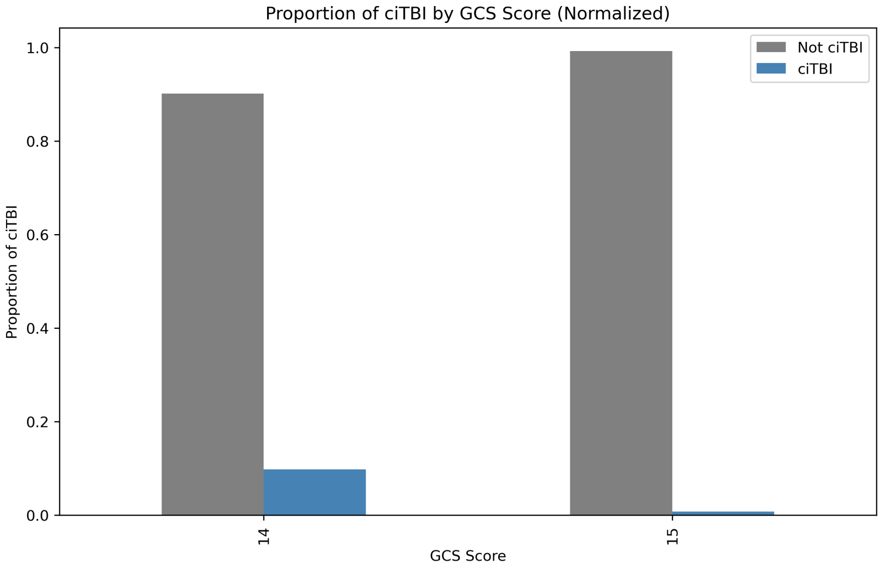
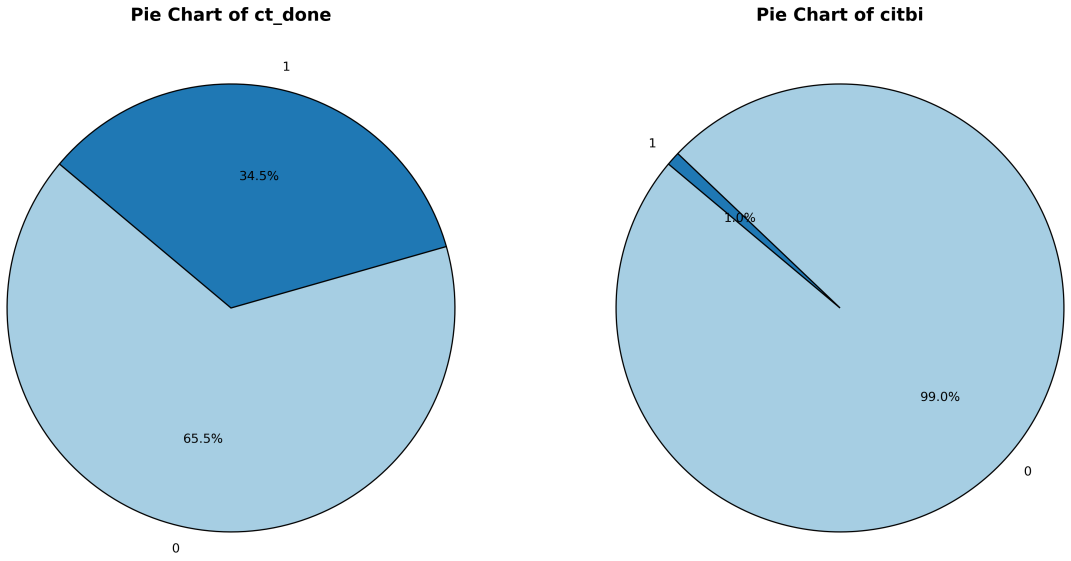


Figure 2: Grouped Bar Chart of GCS Score and ciTBI

3.4 Reality Check

To validate our cleaned dataset, we compare two key variables—ciTBI occurrence rate and CT scan utilization rate—against the findings reported in Kuppermann et al. [1]. According to this study, clinically-important traumatic brain injuries (ciTBI) occurred in approximately 0.9% of cases, while CT scans were performed in 35.3% of patients. These figures serve as external benchmarks to assess the consistency and reliability of our dataset.

After computing the relative frequencies of ciTBI cases and CT scans performed, we compare them against the reference values. Our dataset shows a ciTBI rate close to 0.9% and a CT scan rate around 35.3%, showing that our data passes the reality check, confirming consistency with Kuppermann et al. [1].

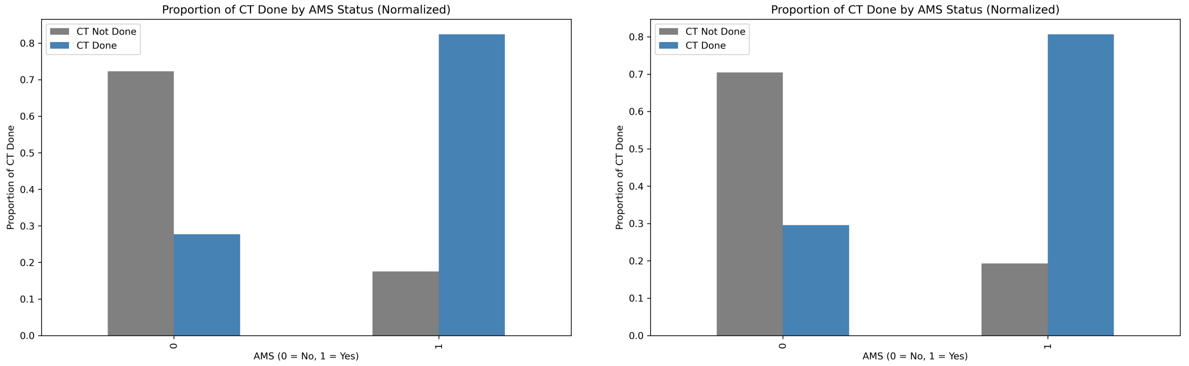


CT Done **ciTBI**

3.5 Stability Check

To assess the stability of our first finding, we compare results using two versions of the dataset. One where missing values were imputed based on domain knowledge, and another where they were imputed in the opposite manner (e.g., values originally filled with 1 were instead filled with 0). The left plot represents the dataset with domain-informed imputation, while the right plot represents the dataset where missing values were left unfilled.

As shown in the comparison, the two distributions exhibit high similarity, suggesting that our imputation strategy does not significantly alter the overall conclusions. This stability check supports the reliability of our findings, indicating that our results are not overly sensitive to the choice of handling missing values.



With Subjective Imputation **Without Subjective Imputation**

4. Modeling

4.1 Implementation

We implemented two Decision Tree Classifiers to predict whether a CT scan is necessary (ciTBI = 1). Decision trees were chosen for their interpretability and ability to handle both categorical and numerical features without scaling.

1. **Algorithms Used**
   - Model 1: Features: *gcs_total, loc_separate, ha, ams*
   - Model 2: Features: *gcs_total, loc_separate, vomit, ha_severity, high_impact_inj_sev*
   - Both models use a **Gini-impurity-based Decision Tree Classifier** with class weights to address data imbalance.
2. **Hyperparameter Selection**
   - *criterion*="gini": Selected for computational efficiency.
   - *class_weight* = {0:1, 1:500}: Increased weight for ciTBI=1 to improve recall.
   - *random_state* = 42: Ensured reproducibility.
   - *max_depth* and *min_samples_split*: Tuned using GridSearchCV with cross-validation (cv=5), optimizing for recall.
3. **Reproducibility**
   - Data was preprocessed by removing missing values and encoding categorical variables.

- Train-test split: 80% training, 20% testing (train_test_split).
- Decision trees were trained using scikit-learn's DecisionTreeClassifier.
- Models were evaluated using accuracy, recall, and confidence intervals.
- Decision paths leading to ciTBI=0 were extracted for interpretability.

4.2 Interpretability

Our decision tree models are highly interpretable, as they follow a step-by-step decision process based on clinical features. Each split represents a simple, logical rule that aligns with clinical reasoning, making the models easy to understand and apply.

Model1:

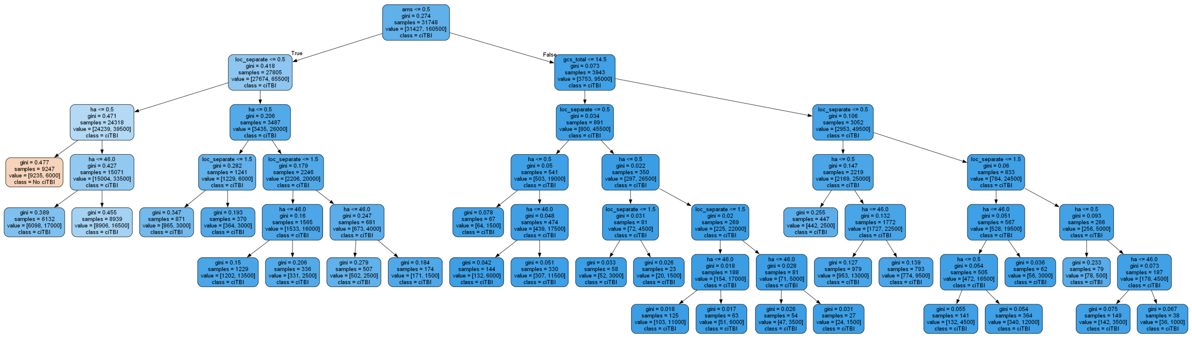

ams <= 0.5 gini = 0.274 samples = 31748 value = [31427, 160500] class = ciTBI
True
False
loc_separate <= 0.5 gini = 0.418 samples = 27805 value = [27674, 65500] class = ciTBI
gcs_total <= 14.5 gini = 0.073 samples = 3943 value = [3753, 95000] class = ciTBI
ha <= 0.5 gini = 0.471 samples = 24318 value = [24239, 39500] class = ciTBI
ha <= 0.5 gini = 0.206 samples = 3487 value = [3435, 26000] class = ciTBI
loc_separate <= 0.5 gini = 0.034 samples = 891 value = [800, 45500] class = ciTBI
loc_separate <= 0.5 gini = 0.106 samples = 3052 value = [2953, 49500] class = ciTBI
gini = 0.477 samples = 9247 value = [9235, 6000] class = No ciTBI
ha <= 46.0 gini = 0.427 samples = 15071 value = [15004, 33500] class = ciTBI
loc_separate <= 1.5 gini = 0.282 samples = 1241 value = [1229, 6000] class = ciTBI
loc_separate <= 1.5 gini = 0.179 samples = 2246 value = [2206, 20000] class = ciTBI
ha <= 0.5 gini = 0.05 samples = 541 value = [503, 19000] class = ciTBI
ha <= 0.5 gini = 0.022 samples = 350 value = [297, 26500] class = ciTBI
ha <= 0.5 gini = 0.147 samples = 2219 value = [2169, 25000] class = ciTBI
loc_separate <= 1.5 gini = 0.06 samples = 833 value = [784, 24500] class = ciTBI
gini = 0.389 samples = 6132 value = [6098, 17000] class = ciTBI
gini = 0.455 samples = 8939 value = [8906, 16500] class = ciTBI
gini = 0.347 samples = 871 value = [865, 3000] class = ciTBI
gini = 0.193 samples = 370 value = [364, 3000] class = ciTBI
ha <= 46.0 gini = 0.16 samples = 1565 value = [1533, 16000] class = ciTBI
ha <= 46.0 gini = 0.247 samples = 681 value = [673, 4000] class = ciTBI
gini = 0.078 samples = 67 value = [64, 1500] class = ciTBI
ha <= 46.0 gini = 0.048 samples = 474 value = [439, 17500] class = ciTBI
loc_separate <= 1.5 gini = 0.031 samples = 81 value = [72, 4500] class = ciTBI
loc_separate <= 1.5 gini = 0.02 samples = 269 value = [225, 22000] class = ciTBI
gini = 0.255 samples = 447 value = [442, 2500] class = ciTBI
ha <= 46.0 gini = 0.132 samples = 1772 value = [1727, 22500] class = ciTBI
ha <= 46.0 gini = 0.051 samples = 567 value = [528, 19500] class = ciTBI
ha <= 0.5 gini = 0.093 samples = 266 value = [256, 5000] class = ciTBI
gini = 0.15 samples = 1229 value = [1202, 13500] class = ciTBI
gini = 0.206 samples = 336 value = [331, 2500] class = ciTBI
gini = 0.279 samples = 507 value = [502, 2500] class = ciTBI
gini = 0.184 samples = 174 value = [171, 1500] class = ciTBI
gini = 0.042 samples = 144 value = [132, 6000] class = ciTBI
gini = 0.051 samples = 330 value = [307, 11500] class = ciTBI
gini = 0.033 samples = 58 value = [52, 3000] class = ciTBI
gini = 0.026 samples = 23 value = [20, 1500] class = ciTBI
ha <= 46.0 gini = 0.018 samples = 188 value = [154, 17000] class = ciTBI
ha <= 46.0 gini = 0.028 samples = 81 value = [71, 5000] class = ciTBI
gini = 0.127 samples = 979 value = [953, 13000] class = ciTBI
gini = 0.139 samples = 793 value = [774, 9500] class = ciTBI
ha <= 0.5 gini = 0.054 samples = 505 value = [472, 16500] class = ciTBI
gini = 0.036 samples = 62 value = [56, 3000] class = ciTBI
gini = 0.233 samples = 79 value = [78, 500] class = ciTBI
ha <= 46.0 gini = 0.073 samples = 187 value = [178, 4500] class = ciTBI
gini = 0.018 samples = 125 value = [103, 11000] class = ciTBI
gini = 0.017 samples = 63 value = [51, 6000] class = ciTBI
gini = 0.026 samples = 54 value = [47, 3500] class = ciTBI
gini = 0.031 samples = 27 value = [24, 1500] class = ciTBI
gini = 0.055 samples = 141 value = [132, 4500] class = ciTBI
gini = 0.054 samples = 364 value = [340, 12000] class = ciTBI
gini = 0.075 samples = 149 value = [142, 3500] class = ciTBI
gini = 0.067 samples = 38 value = [36, 1000] class = ciTBI

Figure 3: Model 1

In **Model 2**, several clinical features were transformed into **binary indicators** to simplify decision-making: **GCS Total Score(gcs_total):** Converted into 0(normal, GCS=15) and 1(impaired, GCS=14) to emphasize neurological deterioration. **Loss of Consciousness(loc_separate):** Transformed into 1(no loss of consciousness) and 0(loss of consciousness). **Vomiting(vomit):** Transformed inoto 1(no vmiting) and 0(vomiting). **Headache Severity(ha_severity):** Transformed inoto 1(mild and moderate) and 0(other) to better distinguish risk levels. **High-Impact Injury(high_impact_inj_sev):** Set to 1(low-impact injury) and 0(0ther).

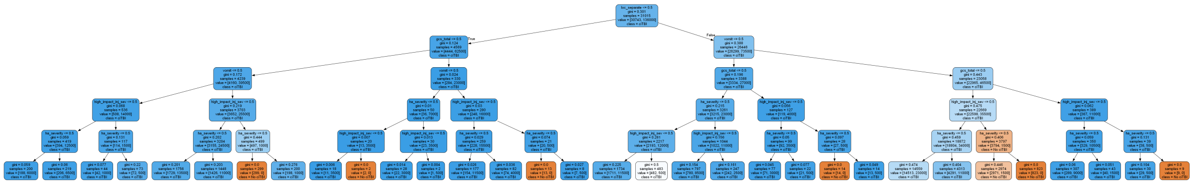

loc_separate <= 0.5 gini = 0.301 samples = 31015 value = [30743, 136000] class = ciTBI
True
False
gcs_total <= 0.5 gini = 0.124 samples = 4569 value = [4444, 62500] class = ciTBI
vomit <= 0.5 gini = 0.388 samples = 26446 value = [26299, 73500] class = ciTBI
vomit <= 0.5 gini = 0.172 samples = 4239 value = [4160, 39500] class = ciTBI
vomit <= 0.5 gini = 0.024 samples = 330 value = [284, 23000] class = ciTBI
gcs_total <= 0.5 gini = 0.196 samples = 3388 value = [3334, 27000] class = ciTBI
gcs_total <= 0.5 gini = 0.443 samples = 23058 value = [22965, 46500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.068 samples = 536 value = [508, 14000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.219 samples = 3703 value = [3652, 25500] class = ciTBI
ha_severity <= 0.5 gini = 0.01 samples = 50 value = [36, 7000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.03 samples = 280 value = [248, 16000] class = ciTBI
ha_severity <= 0.5 gini = 0.215 samples = 3261 value = [3215, 23000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.056 samples = 127 value = [119, 4000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.475 samples = 22669 value = [22598, 35500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.062 samples = 389 value = [367, 11000] class = ciTBI
ha_severity <= 0.5 gini = 0.059 samples = 419 value = [394, 12500] class = ciTBI
ha_severity <= 0.5 gini = 0.131 samples = 117 value = [114, 1500] class = ciTBI
ha_severity <= 0.5 gini = 0.202 samples = 3204 value = [3155, 24500] class = ciTBI
ha_severity <= 0.5 gini = 0.444 samples = 499 value = [497, 1000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.007 samples = 20 value = [13, 3500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.013 samples = 30 value = [23, 3500] class = ciTBI
ha_severity <= 0.5 gini = 0.029 samples = 259 value = [228, 15500] class = ciTBI
ha_severity <= 0.5 gini = 0.074 samples = 21 value = [20, 500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.261 samples = 2217 value = [2193, 12000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.156 samples = 1044 value = [1022, 11000] class = ciTBI
ha_severity <= 0.5 gini = 0.05 samples = 99 value = [92, 3500] class = ciTBI
ha_severity <= 0.5 gini = 0.097 samples = 28 value = [27, 500] class = ciTBI
ha_severity <= 0.5 gini = 0.459 samples = 18872 value = [18804, 34000] class = ciTBI
ha_severity <= 0.5 gini = 0.406 samples = 3797 value = [3794, 1500] class = No ciTBI
ha_severity <= 0.5 gini = 0.059 samples = 350 value = [329, 10500] class = ciTBI
ha_severity <= 0.5 gini = 0.131 samples = 39 value = [38, 500] class = ciTBI
gini = 0.059 samples = 200 value = [188, 6000] class = ciTBI
gini = 0.06 samples = 219 value = [206, 6500] class = ciTBI
gini = 0.077 samples = 44 value = [42, 1000] class = ciTBI
gini = 0.22 samples = 73 value = [72, 500] class = ciTBI
gini = 0.201 samples = 1756 value = [1729, 13500] class = ciTBI
gini = 0.203 samples = 1448 value = [1426, 11000] class = ciTBI
gini = 0.0 samples = 299 value = [299, 0] class = No ciTBI
gini = 0.276 samples = 200 value = [198, 1000] class = ciTBI
gini = 0.006 samples = 18 value = [11, 3500] class = ciTBI
gini = 0.0 samples = 2 value = [2, 0] class = No ciTBI
gini = 0.014 samples = 28 value = [22, 3000] class = ciTBI
gini = 0.004 samples = 2 value = [1, 500] class = ciTBI
gini = 0.026 samples = 177 value = [154, 11500] class = ciTBI
gini = 0.036 samples = 82 value = [74, 4000] class = ciTBI
gini = 0.0 samples = 13 value = [13, 0] class = No ciTBI
gini = 0.027 samples = 8 value = [7, 500] class = ciTBI
gini = 0.225 samples = 1734 value = [1711, 11500] class = ciTBI
gini = 0.5 samples = 483 value = [482, 500] class = ciTBI
gini = 0.154 samples = 797 value = [780, 8500] class = ciTBI
gini = 0.161 samples = 247 value = [242, 2500] class = ciTBI
gini = 0.045 samples = 77 value = [71, 3000] class = ciTBI
gini = 0.077 samples = 22 value = [21, 500] class = ciTBI
gini = 0.0 samples = 14 value = [14, 0] class = No ciTBI
gini = 0.049 samples = 14 value = [13, 500] class = ciTBI
gini = 0.474 samples = 14559 value = [14513, 23000] class = ciTBI
gini = 0.404 samples = 4313 value = [4291, 11000] class = ciTBI
gini = 0.446 samples = 2974 value = [2971, 1500] class = No ciTBI
gini = 0.0 samples = 823 value = [823, 0] class = No ciTBI
gini = 0.06 samples = 307 value = [289, 9000] class = ciTBI
gini = 0.051 samples = 43 value = [40, 1500] class = ciTBI
gini = 0.104 samples = 30 value = [29, 500] class = ciTBI
gini = 0.0 samples = 9 value = [9, 0] class = No ciTBI

Figure 4: Model 2

4.3 Stability

Model 1(origin data set):

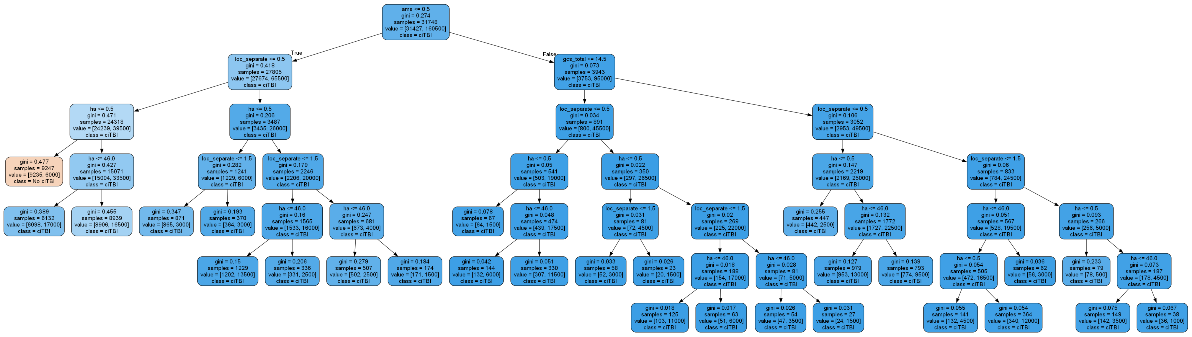

ams <= 0.5 gini = 0.274 samples = 31748 value = [31427, 160500] class = ciTBI
True
False
loc_separate <= 0.5 gini = 0.418 samples = 27805 value = [27674, 65500] class = ciTBI
gcs_total <= 14.5 gini = 0.073 samples = 3943 value = [3753, 95000] class = ciTBI
ha <= 0.5 gini = 0.471 samples = 24318 value = [24239, 39500] class = ciTBI
ha <= 0.5 gini = 0.206 samples = 3487 value = [3435, 26000] class = ciTBI
loc_separate <= 0.5 gini = 0.034 samples = 891 value = [800, 45500] class = ciTBI
loc_separate <= 0.5 gini = 0.106 samples = 3052 value = [2953, 49500] class = ciTBI
gini = 0.477 samples = 9247 value = [9235, 6000] class = No ciTBI
ha <= 46.0 gini = 0.427 samples = 15071 value = [15004, 33500] class = ciTBI
loc_separate <= 1.5 gini = 0.282 samples = 1241 value = [1229, 6000] class = ciTBI
loc_separate <= 1.5 gini = 0.179 samples = 2246 value = [2206, 20000] class = ciTBI
ha <= 0.5 gini = 0.05 samples = 541 value = [503, 19000] class = ciTBI
ha <= 0.5 gini = 0.022 samples = 350 value = [297, 26500] class = ciTBI
ha <= 0.5 gini = 0.147 samples = 2219 value = [2169, 25000] class = ciTBI
loc_separate <= 1.5 gini = 0.06 samples = 833 value = [784, 24500] class = ciTBI
gini = 0.389 samples = 6132 value = [6098, 17000] class = ciTBI
gini = 0.455 samples = 8939 value = [8906, 16500] class = ciTBI
gini = 0.347 samples = 871 value = [865, 3000] class = ciTBI
gini = 0.193 samples = 370 value = [364, 3000] class = ciTBI
ha <= 46.0 gini = 0.16 samples = 1565 value = [1533, 16000] class = ciTBI
ha <= 46.0 gini = 0.247 samples = 681 value = [673, 4000] class = ciTBI
gini = 0.078 samples = 67 value = [64, 1500] class = ciTBI
ha <= 46.0 gini = 0.048 samples = 474 value = [439, 17500] class = ciTBI
loc_separate <= 1.5 gini = 0.031 samples = 81 value = [72, 4500] class = ciTBI
loc_separate <= 1.5 gini = 0.02 samples = 269 value = [225, 22000] class = ciTBI
gini = 0.255 samples = 447 value = [442, 2500] class = ciTBI
ha <= 46.0 gini = 0.132 samples = 1772 value = [1727, 22500] class = ciTBI
ha <= 46.0 gini = 0.051 samples = 567 value = [528, 19500] class = ciTBI
ha <= 0.5 gini = 0.093 samples = 266 value = [256, 5000] class = ciTBI
gini = 0.15 samples = 1229 value = [1202, 13500] class = ciTBI
gini = 0.206 samples = 336 value = [331, 2500] class = ciTBI
gini = 0.279 samples = 507 value = [502, 2500] class = ciTBI
gini = 0.184 samples = 174 value = [171, 1500] class = ciTBI
gini = 0.042 samples = 144 value = [132, 6000] class = ciTBI
gini = 0.051 samples = 330 value = [307, 11500] class = ciTBI
gini = 0.033 samples = 58 value = [52, 3000] class = ciTBI
gini = 0.026 samples = 23 value = [20, 1500] class = ciTBI
ha <= 46.0 gini = 0.018 samples = 188 value = [154, 17000] class = ciTBI
ha <= 46.0 gini = 0.028 samples = 81 value = [71, 5000] class = ciTBI
gini = 0.127 samples = 979 value = [953, 13000] class = ciTBI
gini = 0.139 samples = 793 value = [774, 9500] class = ciTBI
ha <= 0.5 gini = 0.054 samples = 505 value = [472, 16500] class = ciTBI
gini = 0.036 samples = 62 value = [56, 3000] class = ciTBI
gini = 0.233 samples = 79 value = [78, 500] class = ciTBI
ha <= 46.0 gini = 0.073 samples = 187 value = [178, 4500] class = ciTBI
gini = 0.018 samples = 125 value = [103, 11000] class = ciTBI
gini = 0.017 samples = 63 value = [51, 6000] class = ciTBI
gini = 0.026 samples = 54 value = [47, 3500] class = ciTBI
gini = 0.031 samples = 27 value = [24, 1500] class = ciTBI
gini = 0.055 samples = 141 value = [132, 4500] class = ciTBI
gini = 0.054 samples = 364 value = [340, 12000] class = ciTBI
gini = 0.075 samples = 149 value = [142, 3500] class = ciTBI
gini = 0.067 samples = 38 value = [36, 1000] class = ciTBI

Figure 5: Model 1(Original Sata Set)

Model 1(Stability check data set):

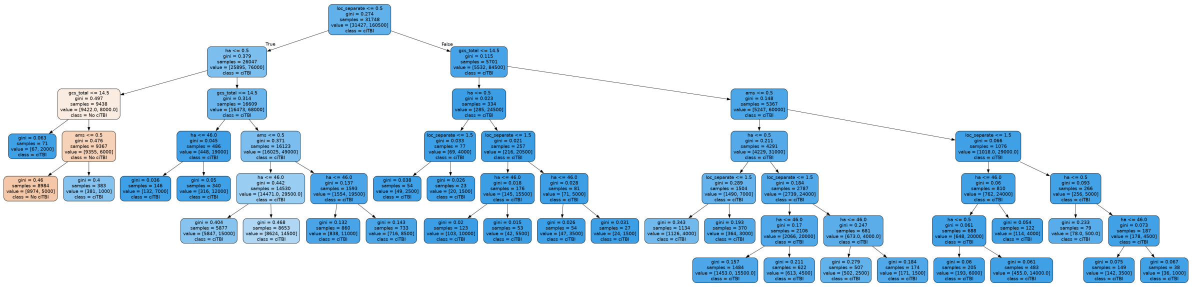

loc_separate <= 0.5 gini = 0.274 samples = 31748 value = [31427, 160500] class = ciTBI
True
False
ha <= 0.5 gini = 0.379 samples = 26047 value = [25895, 76000] class = ciTBI
gcs_total <= 14.5 gini = 0.115 samples = 5701 value = [5532, 84500] class = ciTBI
gcs_total <= 14.5 gini = 0.497 samples = 9438 value = [9422.0, 8000.0] class = No ciTBI
gcs_total <= 14.5 gini = 0.314 samples = 16609 value = [16473, 68000] class = ciTBI
ha <= 0.5 gini = 0.023 samples = 334 value = [285, 24500] class = ciTBI
ams <= 0.5 gini = 0.148 samples = 5367 value = [5247, 60000] class = ciTBI
gini = 0.063 samples = 71 value = [67, 2000] class = ciTBI
ams <= 0.5 gini = 0.476 samples = 9367 value = [9355, 6000] class = No ciTBI
ha <= 46.0 gini = 0.045 samples = 486 value = [448, 19000] class = ciTBI
ams <= 0.5 gini = 0.371 samples = 16123 value = [16025, 49000] class = ciTBI
loc_separate <= 1.5 gini = 0.033 samples = 77 value = [69, 4000] class = ciTBI
loc_separate <= 1.5 gini = 0.021 samples = 257 value = [216, 20500] class = ciTBI
ha <= 0.5 gini = 0.211 samples = 4291 value = [4229, 31000] class = ciTBI
loc_separate <= 1.5 gini = 0.066 samples = 1076 value = [1018.0, 29000.0] class = ciTBI
gini = 0.46 samples = 8984 value = [8974, 5000] class = No ciTBI
gini = 0.4 samples = 383 value = [381, 1000] class = ciTBI
gini = 0.036 samples = 146 value = [132, 7000] class = ciTBI
gini = 0.05 samples = 340 value = [316, 12000] class = ciTBI
ha <= 46.0 gini = 0.442 samples = 14530 value = [14471.0, 29500.0] class = ciTBI
ha <= 46.0 gini = 0.137 samples = 1593 value = [1554, 19500] class = ciTBI
gini = 0.038 samples = 54 value = [49, 2500] class = ciTBI
gini = 0.026 samples = 23 value = [20, 1500] class = ciTBI
ha <= 46.0 gini = 0.018 samples = 176 value = [145, 15500] class = ciTBI
ha <= 46.0 gini = 0.028 samples = 81 value = [71, 5000] class = ciTBI
loc_separate <= 1.5 gini = 0.289 samples = 1504 value = [1490, 7000] class = ciTBI
loc_separate <= 1.5 gini = 0.184 samples = 2787 value = [2739, 24000] class = ciTBI
ha <= 46.0 gini = 0.06 samples = 810 value = [762, 24000] class = ciTBI
ha <= 0.5 gini = 0.093 samples = 266 value = [256, 5000] class = ciTBI
gini = 0.404 samples = 5877 value = [5847, 15000] class = ciTBI
gini = 0.468 samples = 8653 value = [8624, 14500] class = ciTBI
gini = 0.132 samples = 860 value = [838, 11000] class = ciTBI
gini = 0.143 samples = 733 value = [716, 8500] class = ciTBI
gini = 0.02 samples = 123 value = [103, 10000] class = ciTBI
gini = 0.015 samples = 53 value = [42, 5500] class = ciTBI
gini = 0.026 samples = 54 value = [47, 3500] class = ciTBI
gini = 0.031 samples = 27 value = [24, 1500] class = ciTBI
gini = 0.343 samples = 1134 value = [1126, 4000] class = ciTBI
gini = 0.193 samples = 370 value = [364, 3000] class = ciTBI
ha <= 46.0 gini = 0.17 samples = 2106 value = [2066, 20000] class = ciTBI
ha <= 46.0 gini = 0.247 samples = 681 value = [673.0, 4000.0] class = ciTBI
ha <= 0.5 gini = 0.061 samples = 688 value = [648, 20000] class = ciTBI
gini = 0.054 samples = 122 value = [114, 4000] class = ciTBI
gini = 0.233 samples = 79 value = [78.0, 500.0] class = ciTBI
ha <= 46.0 gini = 0.073 samples = 187 value = [178, 4500] class = ciTBI
gini = 0.157 samples = 1484 value = [1453.0, 15500.0] class = ciTBI
gini = 0.211 samples = 622 value = [613, 4500] class = ciTBI
gini = 0.279 samples = 507 value = [502, 2500] class = ciTBI
gini = 0.184 samples = 174 value = [171, 1500] class = ciTBI
gini = 0.06 samples = 205 value = [193, 6000] class = ciTBI
gini = 0.061 samples = 483 value = [455.0, 14000.0] class = ciTBI
gini = 0.075 samples = 149 value = [142, 3500] class = ciTBI
gini = 0.067 samples = 38 value = [36, 1000] class = ciTBI

Figure 6: Model 1(Stability Check Data Set)

Prediction Change Rate: 0.0078

Model 2(origin data set):

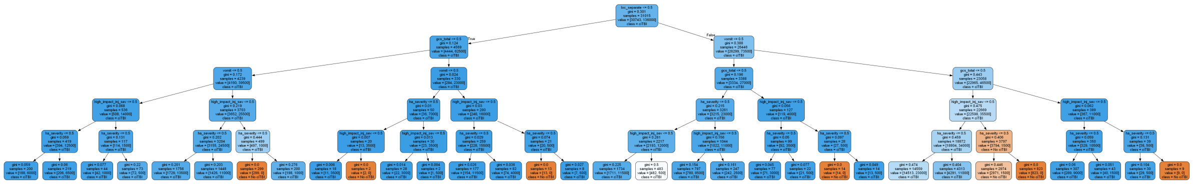

loc_separate <= 0.5 gini = 0.301 samples = 31015 value = [30743, 136000] class = ciTBI
True
False
gcs_total <= 0.5 gini = 0.124 samples = 4569 value = [4444, 62500] class = ciTBI
vomit <= 0.5 gini = 0.388 samples = 26446 value = [26299, 73500] class = ciTBI
vomit <= 0.5 gini = 0.172 samples = 4239 value = [4160, 39500] class = ciTBI
vomit <= 0.5 gini = 0.024 samples = 330 value = [284, 23000] class = ciTBI
gcs_total <= 0.5 gini = 0.198 samples = 3388 value = [3334, 27000] class = ciTBI
gcs_total <= 0.5 gini = 0.443 samples = 23058 value = [22965, 46500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.068 samples = 536 value = [508, 14000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.219 samples = 3703 value = [3652, 25500] class = ciTBI
ha_severity <= 0.5 gini = 0.01 samples = 50 value = [36, 7000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.03 samples = 280 value = [248, 16000] class = ciTBI
ha_severity <= 0.5 gini = 0.215 samples = 3261 value = [3215, 23000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.056 samples = 127 value = [119, 4000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.475 samples = 22669 value = [22598, 35500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.062 samples = 389 value = [367, 11000] class = ciTBI
ha_severity <= 0.5 gini = 0.059 samples = 419 value = [394, 12500] class = ciTBI
ha_severity <= 0.5 gini = 0.131 samples = 117 value = [114, 1500] class = ciTBI
ha_severity <= 0.5 gini = 0.202 samples = 3204 value = [3155, 24500] class = ciTBI
ha_severity <= 0.5 gini = 0.444 samples = 499 value = [497, 1000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.007 samples = 20 value = [13, 3500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.013 samples = 30 value = [23, 3500] class = ciTBI
ha_severity <= 0.5 gini = 0.029 samples = 259 value = [228, 15500] class = ciTBI
ha_severity <= 0.5 gini = 0.074 samples = 21 value = [20, 500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.261 samples = 2217 value = [2193, 12000] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.156 samples = 1044 value = [1022, 11000] class = ciTBI
ha_severity <= 0.5 gini = 0.06 samples = 99 value = [92, 3500] class = ciTBI
ha_severity <= 0.5 gini = 0.097 samples = 28 value = [27, 500] class = ciTBI
ha_severity <= 0.5 gini = 0.459 samples = 18872 value = [18804, 34000] class = ciTBI
ha_severity <= 0.5 gini = 0.406 samples = 3797 value = [3794, 1500] class = No ciTBI
ha_severity <= 0.5 gini = 0.059 samples = 350 value = [329, 10500] class = ciTBI
ha_severity <= 0.5 gini = 0.131 samples = 39 value = [38, 500] class = ciTBI
gini = 0.059 samples = 200 value = [188, 6000] class = ciTBI
gini = 0.06 samples = 219 value = [206, 6500] class = ciTBI
gini = 0.077 samples = 44 value = [42, 1000] class = ciTBI
gini = 0.22 samples = 73 value = [72, 500] class = ciTBI
gini = 0.201 samples = 1756 value = [1729, 13500] class = ciTBI
gini = 0.203 samples = 1448 value = [1426, 11000] class = ciTBI
gini = 0.0 samples = 299 value = [299, 0] class = No ciTBI
gini = 0.276 samples = 200 value = [198, 1000] class = ciTBI
gini = 0.006 samples = 18 value = [11, 3500] class = ciTBI
gini = 0.0 samples = 2 value = [2, 0] class = No ciTBI
gini = 0.014 samples = 28 value = [22, 3000] class = ciTBI
gini = 0.004 samples = 2 value = [1, 500] class = ciTBI
gini = 0.026 samples = 177 value = [154, 11500] class = ciTBI
gini = 0.036 samples = 82 value = [74, 4000] class = ciTBI
gini = 0.0 samples = 13 value = [13, 0] class = No ciTBI
gini = 0.027 samples = 8 value = [7, 500] class = ciTBI
gini = 0.226 samples = 1734 value = [1711, 11500] class = ciTBI
gini = 0.5 samples = 483 value = [482, 500] class = ciTBI
gini = 0.154 samples = 797 value = [780, 8500] class = ciTBI
gini = 0.161 samples = 247 value = [242, 2500] class = ciTBI
gini = 0.045 samples = 77 value = [71, 3000] class = ciTBI
gini = 0.077 samples = 22 value = [21, 500] class = ciTBI
gini = 0.0 samples = 14 value = [14, 0] class = No ciTBI
gini = 0.049 samples = 14 value = [13, 500] class = ciTBI
gini = 0.474 samples = 14559 value = [14513, 23000] class = ciTBI
gini = 0.404 samples = 4313 value = [4291, 11000] class = ciTBI
gini = 0.446 samples = 2974 value = [2971, 1500] class = No ciTBI
gini = 0.0 samples = 823 value = [823, 0] class = No ciTBI
gini = 0.06 samples = 307 value = [289, 9000] class = ciTBI
gini = 0.051 samples = 43 value = [40, 1500] class = ciTBI
gini = 0.104 samples = 30 value = [29, 500] class = ciTBI
gini = 0.0 samples = 9 value = [9, 0] class = No ciTBI

Figure 7: Model 2(Original Data Set)

Model 2(Stability check data set):

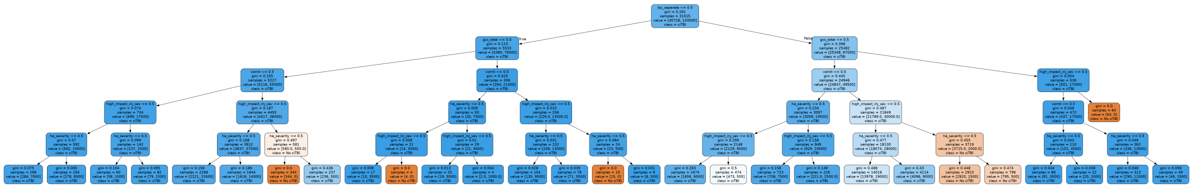

loc_separate <= 0.5 gini = 0.291 samples = 31015 value = [30728, 143500] class = ciTBI
True
False
gcs_total <= 0.5 gini = 0.123 samples = 5533 value = [5380, 76500] class = ciTBI
gcs_total <= 0.5 gini = 0.398 samples = 25482 value = [25348, 67000] class = ciTBI
vomit <= 0.5 gini = 0.155 samples = 5227 value = [5116, 55500] class = ciTBI
vomit <= 0.5 gini = 0.025 samples = 306 value = [264, 21000] class = ciTBI
vomit <= 0.5 gini = 0.445 samples = 24946 value = [24847, 49500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.054 samples = 536 value = [501, 17500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.074 samples = 734 value = [699, 17500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.187 samples = 4493 value = [4417, 38000] class = ciTBI
ha_severity <= 0.5 gini = 0.009 samples = 50 value = [35, 7500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.033 samples = 256 value = [229.0, 13500.0] class = ciTBI
ha_severity <= 0.5 gini = 0.234 samples = 3097 value = [3058, 19500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.487 samples = 21849 value = [21789.0, 30000.0] class = ciTBI
vomit <= 0.5 gini = 0.048 samples = 472 value = [437, 17500] class = ciTBI
gini = 0.0 samples = 64 value = [64, 0] class = No ciTBI
ha_severity <= 0.5 gini = 0.07 samples = 592 value = [562, 15000] class = ciTBI
ha_severity <= 0.5 gini = 0.099 samples = 142 value = [137, 2500] class = ciTBI
ha_severity <= 0.5 gini = 0.168 samples = 3912 value = [3837, 37500] class = ciTBI
ha_severity <= 0.5 gini = 0.497 samples = 581 value = [580.0, 500.0] class = No ciTBI
high_impact_inj_sev <= 0.5 gini = 0.008 samples = 21 value = [14, 3500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.01 samples = 29 value = [21, 4000] class = ciTBI
ha_severity <= 0.5 gini = 0.031 samples = 232 value = [206, 13000] class = ciTBI
ha_severity <= 0.5 gini = 0.084 samples = 24 value = [23, 500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.299 samples = 2148 value = [2129, 9500] class = ciTBI
high_impact_inj_sev <= 0.5 gini = 0.156 samples = 949 value = [929, 10000] class = ciTBI
ha_severity <= 0.5 gini = 0.477 samples = 18130 value = [18074, 28000] class = ciTBI
ha_severity <= 0.5 gini = 0.455 samples = 3719 value = [3715.0, 2000.0] class = No ciTBI
ha_severity <= 0.5 gini = 0.043 samples = 110 value = [101, 4500] class = ciTBI
ha_severity <= 0.5 gini = 0.049 samples = 362 value = [336, 13000] class = ciTBI
gini = 0.075 samples = 298 value = [284, 7000] class = ciTBI
gini = 0.065 samples = 294 value = [278, 8000] class = ciTBI
gini = 0.104 samples = 60 value = [58, 1000] class = ciTBI
gini = 0.095 samples = 82 value = [79, 1500] class = ciTBI
gini = 0.158 samples = 2268 value = [2221, 23500] class = ciTBI
gini = 0.186 samples = 1644 value = [1616, 14000] class = ciTBI
gini = 0.0 samples = 344 value = [344, 0] class = No ciTBI
gini = 0.436 samples = 237 value = [236, 500] class = ciTBI
gini = 0.006 samples = 17 value = [10, 3500] class = ciTBI
gini = 0.0 samples = 4 value = [4, 0] class = No ciTBI
gini = 0.013 samples = 25 value = [19, 3000] class = ciTBI
gini = 0.004 samples = 4 value = [2.0, 1000.0] class = ciTBI
gini = 0.028 samples = 154 value = [135, 9500] class = ciTBI
gini = 0.039 samples = 78 value = [71, 3500] class = ciTBI
gini = 0.0 samples = 15 value = [15, 0] class = No ciTBI
gini = 0.031 samples = 9 value = [8, 500] class = ciTBI
gini = 0.263 samples = 1674 value = [1656, 9000] class = ciTBI
gini = 0.5 samples = 474 value = [473, 500] class = ciTBI
gini = 0.158 samples = 723 value = [708, 7500] class = ciTBI
gini = 0.149 samples = 226 value = [221.0, 2500.0] class = ciTBI
gini = 0.488 samples = 14016 value = [13978, 19000] class = ciTBI
gini = 0.43 samples = 4114 value = [4096, 9000] class = ciTBI
gini = 0.448 samples = 2923 value = [2920, 1500] class = No ciTBI
gini = 0.474 samples = 796 value = [795, 500] class = No ciTBI
gini = 0.044 samples = 88 value = [81, 3500] class = ciTBI
gini = 0.038 samples = 22 value = [20, 1000] class = ciTBI
gini = 0.048 samples = 313 value = [290, 11500] class = ciTBI
gini = 0.058 samples = 49 value = [46, 1500] class = ciTBI

Figure 8: Model 2(Stability Check Data Set)

Prediction Change Rate: 0.0006

Based on the prediction change rate and the decision tree model showed above, we conclude that both models are stable.

5. Discussion

The dataset size did not impose major restrictions, but the limited number of ciTBI=1 cases posed challenges in training a balanced model. To address this, class weighting was applied to ensure the model did not overly favor ciTBI=0. Additionally, maintaining consistency between the original and perturbed datasets was essential for conducting a reliable stability analysis.

This study encompasses the three realms of data science: data and reality, algorithms and models, and future data and reality. While the dataset captures clinical characteristics, it does not perfectly reflect reality due to potential biases, measurement errors, and missing values. The decision tree models were chosen for their interpretability, and hyperparameter tuning, along with stability checks, ensured robustness. However, for these models to be effective in real-world settings, further validation on unseen patient data is necessary to assess their generalizability.

Although the dataset is derived from real clinical observations, it does not provide a one-to-one correspondence with reality. Certain factors influencing clinical decision-making may not be fully captured in structured data. Data visualization plays a crucial role in making findings more interpretable, but it can also oversimplify complex relationships in real-world clinical decision-making. Thus, while the models offer a structured approach to predicting ciTBI, clinical expertise remains essential in ensuring their practical applicability.

6. Conclusion

This study explored the PECARN TBI dataset to improve CT scan decision-making in pediatric head trauma. Decision tree models identified key predictors such as altered mental status, GCS score, and injury severity, confirming their clinical relevance. Stability checks showed that the models remained robust under data perturbation. While the models demonstrated strong interpretability and reliability, further validation on external datasets is needed to ensure broader applicability. Ultimately, this work supports reducing unnecessary CT scans while maintaining high sensitivity for serious TBIs through data-driven decision support.

7. Academic honesty statement

I pledge that the work presented in this report is my own, and that all sources I have used, including discussions with classmates, have been properly cited. I understand that academic integrity is essential to maintaining the credibility and reliability of research.

Academic research honesty is necessary because it ensures the validity of knowledge, fosters trust within the academic community, and upholds ethical standards. Misrepresentation of work not only undermines the learning process but also distorts the integrity of scientific inquiry. By adhering to principles of honesty, we contribute to a research environment where ideas are built upon accurately and responsibly.

8. Collaborators

I done all the works by myslef in this lab.

9. Bibliography

1. Kuppermann, Nathan, et al. “Identification of Children at Very Low Risk of Clinically-Important Brain Injuries after Head Trauma: A Prospective Cohort Study.” *The Lancet*, vol. 374, no. 9696, 2009, pp. 1160-1170. https://doi.org/10.1016/S0140-6736(09)61558-0.