Stat214 Lab1

2025/02/21

1 Introduction

The premise of this report was based on Kuppermann (2009). I investigated whether various clinical indicators in the mild trauma group - such as consciousness, GCS, dizziness, vomiting, headache, and scalp haematoma - contribute to the prediction of traumatic brain injury (clinically-important traumatic brain injury: ciTBI). I will assess the risk of ciTBI through EDA and modeling of head injuries in children with minor trauma, and consider the possibility of reducing CT scans for unnecessary children. ciTBI is one of the major causes of death and sequelae in children, and while rapid diagnosis by CT scan is required, but there is concern about the increased risk of disease due to radiation exposure for children. In recent years, as the frequency of use has increased due to the increase in the number of CT scanners, but the number of cases of ciTBI even in cases of trauma is very small, so the aim is to avoid unnecessary exposue risks by not conducting tests in cases of low ciTBI risk. In this report, I first confirmed the outline and generation process of the collected data, and then investigated all the variables. Furthermore, I used domain knowledge to preprocess the variables in order to reduce missing values, reduce variables that do not contribute to the prediction of ciTBI, and create variables with more information.

In the EDA, I investigated variables that might contribute to predictions, sample bias, and bias in the ratio of 0s and 1s. In the modeling, I used decision trees and logistic regression to predict ciTBI.

2 Data

The data used in this analysis is a cohort of 57,030 children with head injuries collected from 25 North American pediatric emergency departments between June 2004 and March 2006. The target population was patients under the age of 18, and of these, 43,399 patients were evaluated. In Kuppermann (2009), the data was narrowed down to patients with mild trauma with a GCS score of 14-15, and there were 42,412 cases, but in this report, this narrowing

down was not done because there is a possibility that even patients with a poor GCS score have a low risk of ciTBI. In addition to basic patient information (age, gender, race, etc.), the data also includes details of the mechanism of injury, level of consciousness, seizure, vomiting, headache, scalp haematoma, and CT scan results and ciTBI. The purpose of collecting this data was to create rules for diagnosing ciTBI, and it is hoped that the data obtained will contribute to proposing new solutions to the discovery of ciTBI and avoiding CT scans for patients at low risk.

2.1 Data Collection

This data was collected in the form of records made by doctors prior to making clinical decisions when the target pediatric patients visited the emergency department with head injuries. The feature of this data set is that it was collected at each facility according to consistent evaluation criteria using a standardized data form. Quality control methods such as double entry, random triple checking, and facility monitoring were implemented for data entry at each facility. Re-evaluation by specialists and other doctors was also carried out. The evaluation methods for each variable were appropriately adjusted according to the age of the patients and the severity of their injuries. In addition, thorough follow-up was carried out after the patients were discharged, and in addition to follow-up checks by telephone, the medical records were rechecked and the morgue was checked to determine whether there was any ciTBI that may have been missed in the initial diagnosis. In short, the data was created using these multiple data collection methods to increase the reliability of the data and improve the accuracy of the analysis results.

2.2 Data Preprocessing

With the exception of GCS and age, all of the data is categorical. Another characteristic of the data is that there are many cases where the percentage of missing values is high. This is shown in Figure 1. This is thought to be because the results of each test are often given as Yes/No, and if the test is not carried out, it becomes a missing value. There are four types of missing values: those where data is not entered, and those where the numbers 90, 91, and 92 are entered. Therefore, we use domain knowledge to combine variables and reduce missing values.

Figure 1 shows the results of extracting representative columns and visualizing the missing value rate. The missing value rate for each column was calculated, and 11 columns were extracted based on the missing value rate rankings, which are based on the 10% increments of the percentiles. A bar graph was created for the extracted columns. Looking at this graph, we can see that more than half of the columns have a missing value ratio of over 50%, and that the missing value ratio for data of 20% is around 90% or more.

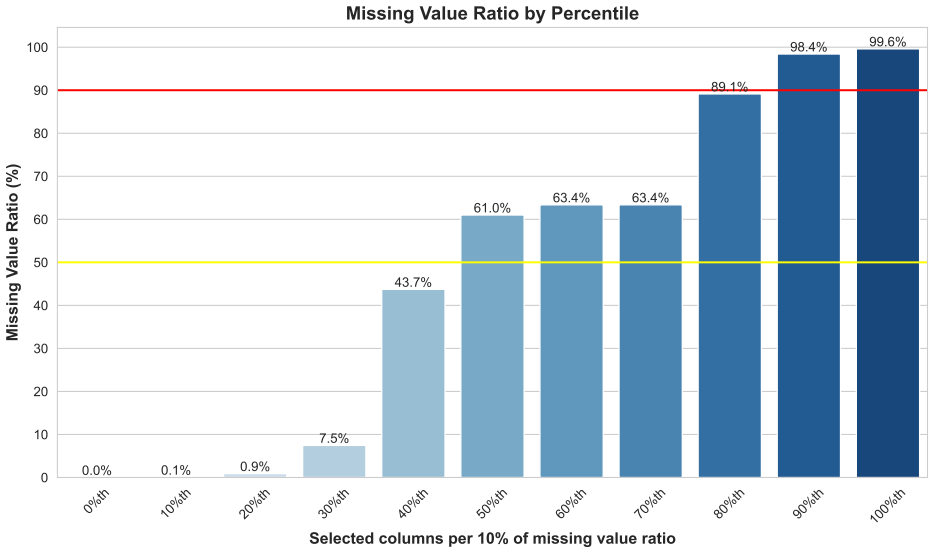

Figure 1: Missing Value Ratio

In addition, because the ciTBI prediction includes data that is irrelevant or unusable, such as CT results, we exclude such data. Furthermore, because some data has overlapping meanings, I retained data with more information.

I have defined a function called 'combine_category' to combine category variables based on domain knowledge. This function combines a basic category (e.g. Yes/No) and its detailed information (e.g. specific numerical values or time periods) into a single value. Specifically, if the basic value is missing, it returns a missing value (np.nan), if it is 0, it returns 0 as it is, if it is 1 and detailed values exist, it returns the value, and if detailed values are missing, it returns the value obtained by adding 1 to the maximum value of detailed data (as a worst case). Following this, I dropped the original Yes/No judgement. By doing this, I reduced missing values and variables while maintaining the amount of information.

Next, the main preprocessing function 'preprocess_data' formats the data in the following steps.

1. The data is read from the CSV file while empty characters are converted to np.nan.
2. Columns such as 'EmplType' and 'Certification' that are judged to be not directly related to the judgment of ciTBI are deleted.
3. I made corrections based on domain knowledge for each variable as the following table.
4. In addition, we have deleted items that are not necessary for analysis as the following table.
5. For variables that have both a Yes/No judgement and detailed numerical information, we use the combine_category function mentioned above to combine the original Yes/No judgement and detailed information.

6. For Clav, NeuroD, OSI, the definitions of the other (ClavOth, NeuroDOth, OSIOth) have been changed based on the relationship between the detailed findings and the original values.
7. I created a target variable for the clinical outcome of interest, ciTBI, and deleted the original variable.
8. After converting all columns to Int64 type, the data is divided into training and testing data at a ratio of 80% and 20%.

Detailed Preprocessing is shown by each variable below:

Variable / Group	Processing
EmplType, Certification	- Dropped since they do not affect the ciTBI outcome.
InjuryMech	- Missing values and entries with value 90 are replaced with 13 as others.
Amnesia_verb	- Values equal to 91 are replaced with nan because whether preverbal or not is not related to ciTBI.
LOCSeparate & LocLen	- In LOCSeparate, value 2 is replaced with 1 (treating "suspected" as Yes). - In LocLen, value 92 is replaced with NaN. - Combined using the combine_category function with LOCSeparate as the origin.
Seiz Group (Seiz, SeizOccur, SeizLen)	- Replace value 92 with NaN in SeizOccur and SeizLen. - Combined using combine_category with Seiz as the origin.
HA_verb Group (HA_verb, HASeverity, HAStart)	- Replace 91 with NaN in HA_verb because whether preverbal or not is not related to ciTBI. - Replace 92 with NaN in HASeverity and HAStart. - Combined using combine_category with HA_verb as the origin.
Vomit Group (Vomit, VomitNbr, VomitStart, VomitLast)	- Replace value 92 with NaN in VomitNbr, VomitStart, and VomitLast. - Combined using combine_category with Vomit as the origin.
GCSGroup	- Dropped because GCSTotal provides the same information.
AMS Variables (AMSAgitated, AMSSleep, AMSSlow, AMSRepeat, AMSOth)	- Replace value 92 with NaN in all these columns.
SFxPalp Group (SFxPalp, SFxPalpDepress)	- Replace value 2 with 1 in SFxPalp (treating unclear cases as Yes). - Replace value 92 with NaN in SFxPalpDepress, then increment by 1. - Combined using combine_category with SFxPalp as the origin.
SFxBas Group (SFxBasHem, SFxBasOto, SFxBasPer, SFxBasRet, SFxBasRhi)	- For each detailed variable, replace 92 with NaN and add 1. - Combined using combine_category with SFxBasHem as the origin.
Hema Group (HemaLoc, HemaSize)	- Replace value 92 with NaN in both columns; combined using combine_category with HemaLoc as the origin.

Variable / Group	Processing
Clav Group (Clav, ClavFace, ClavNeck, ClavFro, ClavOcc, ClavPar, ClavTem, ClavOth)	- Replace value 92 with NaN in ClavFace, ClavNeck, ClavFro, ClavOcc, ClavPar, and ClavTem. - Make ClavOth as the difference between Clav and the sum of detailed items (set to 1 if positive, otherwise 0). - Drop the original Clav column.
NeuroD Group (NeuroD, NeuroDMotor, NeuroDSensory, NeuroDCranial, NeuroDReflex, NeuroDOth)	- Replace value 92 with NaN in NeuroDMotor, NeuroDSensory, NeuroDCranial, and NeuroDReflex. - Make NeuroDOth based on the difference between NeuroD and the sum of detailed items (set to 1 if positive, otherwise 0) and any existing NeuroDOth (set to 1 if exists, otherwise 0). - Drop the original NeuroD column.
OSI Group (OSI, OSIExtremity, OSICut, OSICspine, OSIFlank, OSIAbdomen, OSIPelvis, OSIOth)	- Replace value 92 with NaN in OSIExtremity, OSICut, OSICspine, OSIFlank, OSIAbdomen, and OSIPelvis. - Make OSIOth based on the difference between OSI and the sum of detailed items (set to 1 if positive, otherwise 0) and any existing OSIOth (set to 1 if exists, otherwise 0). - Drop the original OSI column.
CT-Related Columns	- Dropped all columns related to CT imaging (CTForm1, IndAge, IndAmnesia, IndAMS, IndClinSFx, IndHA, IndHema, IndLOC, IndMech, IndNeuroD, IndRqstMD, IndRqstParent, IndRqstTrauma, IndSeiz, IndVomit, IndXraySFx, IndOth, CTSed, CTSedAgitate, CTSedAge, CTSedRqst, CTSedOth) because our goal isto prevent use of CT data.
AgeinYears, AgeTwoPlus	- Dropped as duplicate age information.
Race	- Replace value 90 with 6 as others.
Additional CT-Related Columns	- Dropped further CT-related columns (Observed, EDDisposition, CTDone, EDCT, PosCT, Finding1–14, Finding20–23) because our goal isto prevent use of CT data.
HospHead	- Dropped because it is not related to the ciTBI outcome.
PosIntFinal	- Fullfilled as 0/1 if all of DeathTBI, HospHeadPosCT, Intub24Head, and Neurosurgery are 0 or any of them equals 1 when PosIntFinal is missing.

2.3 Data Exploration

From here on, I will carry out the analysis using the processed training data. First, the final number of variables after processing is 60. As mentioned in the previous chapter, the data contains many missing values, and even after preprocessing, the rate of missing values in the entire data set is 30.1%.

Next, after excluding InjuryMech and Race, which have more than three categories but have no rankings, I checked the Spearman's rank correlation between the predicted target, PosIntFinal, and each variable. Figure 2 shows the results. The largest rank correlation was a little below 0.4, so it does not seem that there are any very effective factors for prediction. In addition,

the majority of variables were skewed towards 0~0.1, so it is highly likely that there are many variables that have almost no effect on prediction. Also, some of the variables had negative correlations (eg. GCS), so taking care is needed when interpreting the model.

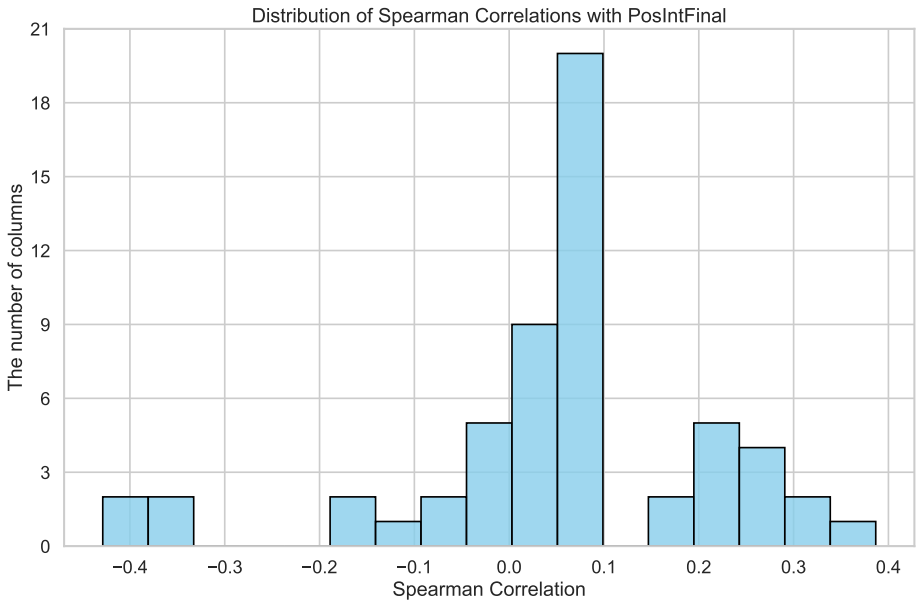

Figure 2: Correlation with PosIntFinal

3 Findings

In this section, I will investigate the variables that are useful for predicting ciTBI. After that, I will investigate the bias in the data collection and the ratio of 0s and 1s.

3.1 Data useful to ciTBI prediction

I visualized the importance of each category (AgeInMonth and GCS are also treated as categories) contributing to the prediction of ciTBI. For each category in each column (including np.nan), I calculated the percentage PosIntFinal is 1. Then, for each column, I calculated the Importance, which is the number obtained by subtracting the smallest from the largest of the percentages; the Importance is only the attribute of the category alone to the prediction of ciTBI, and does not include cross-influence, but it is useful for estimating the contribution of the variable. In Figure 3, we can see multiple rows ranging from quite close to 1 to 0.5, although most of the columns are close to 0. Also, when I check the names of the Top 5

columns, GCS and Intubated are highly influential. From this, it can be inferred that ciTBI can be predicted by using these columns.

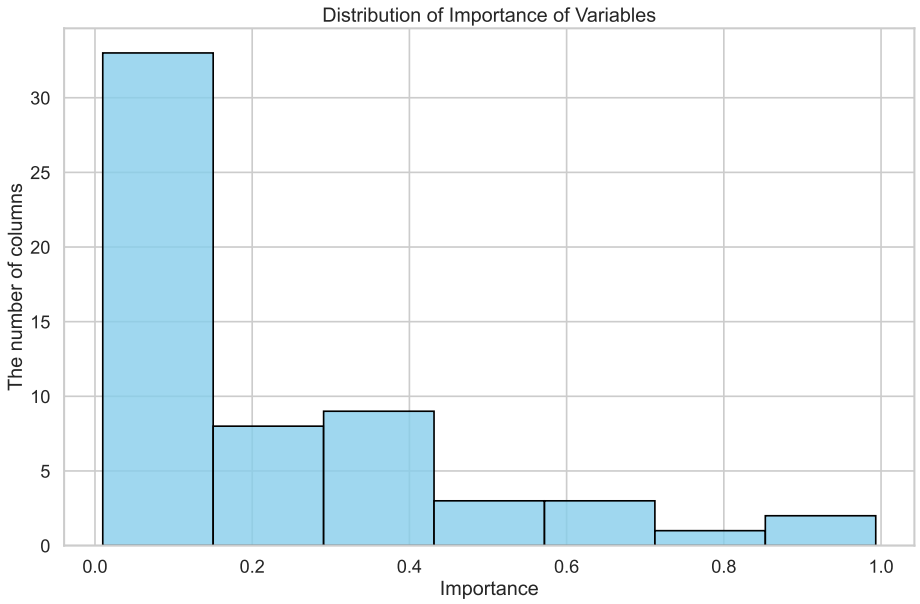

Figure 3: Importance of Variables

Top 5 important columns are shown below:

Column	Importance
GCSTotal	0.993
GCSMotor	0.990
Intubated	0.713
GCSEye	0.702
Paralyzed	0.642

3.2 Data colllected form biased sample

Kuppermann (2009) said that data was collected from those aged 18 and under, but as Figure 4 shows, there is a bias towards a higher number of users aged 0 to 4, and after that the numbers are more or less even. Also, the male-female ratio differs for each age group, with slightly more females in the younger age groups, but from around the age of 6, the number of girls per month is about double that of boys. According to Figure 5, the data for white children accounts for around 50%, and the data for black children accounts for around 40%, so there is also a bias in the racial composition of the data source. As a result, this analysis was conducted using a biased sample.

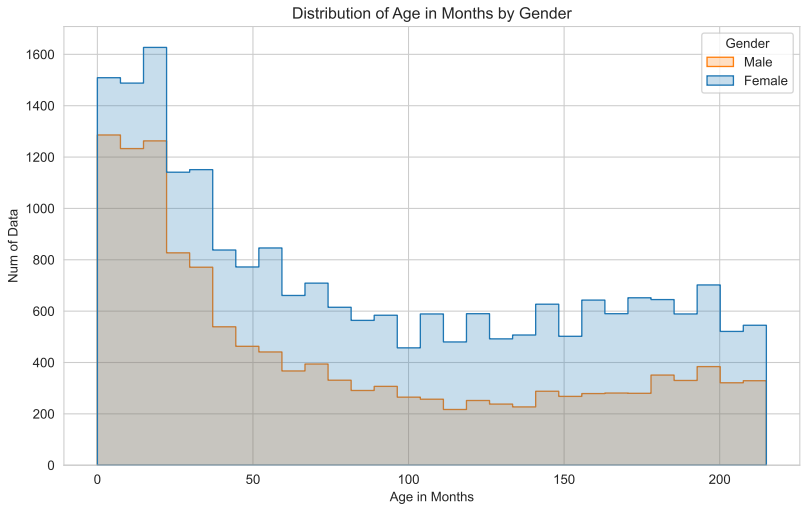

Figure 4: Age Distribution

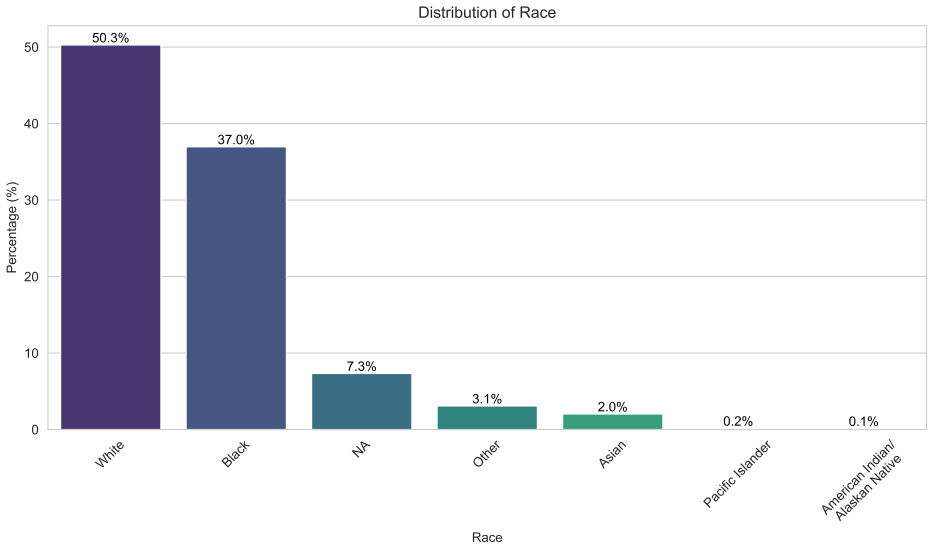

Figure 5: Race Distribution

3.3 Bias in the Ratio of 0s and 1s

Figure 6 was made to see the bias in the ratio of 0s and 1s. First, the ratio of 0s and 1s was calculated for data consisting only of 0s and 1s, and the absolute value of the difference was calculated and a histogram was created. From Figure 6, it can be seen that there are many cases where there is a significant bias in the ratio. This is thought to be happening because the data is basically not applicable to things such as test results and symptoms. Even for the target variable of this analysis, PosIntFinal, the number of ciTBI is only 1.8%. Therefore, analysis based on the premise that there is a bias between 0s and 1s is necessary. In other words, some of the variables are highly correlated.

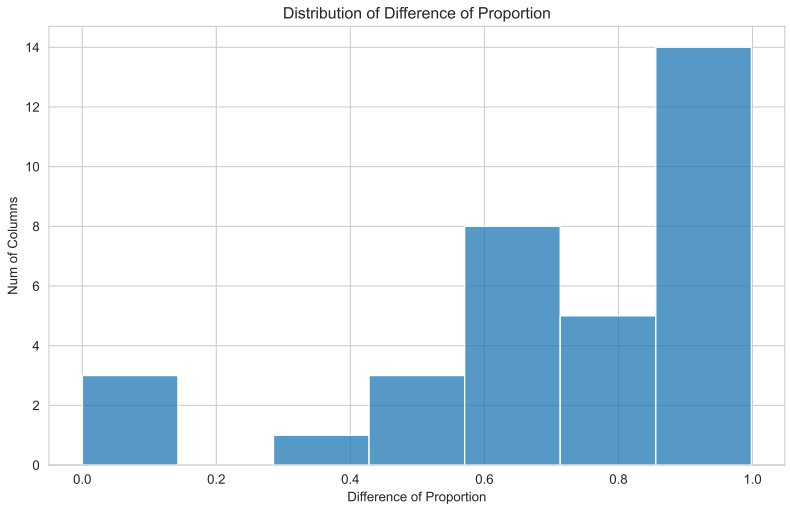

Figure 6: Ratio of 0s and 1s

3.4 Reality Check

In the reality check, I would like to confirm the extent to which the data reflects the actual situation, taking into account the actual emergency medical care situation. First, the data is collected under strict quality control, so it is thought to accurately reflect the target clinical population. In addition, the incidence of ciTBI is very low at around 1.8%, which is consistent with the low incidence seen in actual clinical settings as described by Kuppermann (2009). Therefore, it is thought that this is a realistic analysis for hospitals that accept patients similar to the target clinical population. However, there is a possibility of bias in the selection of the target clinical population due to the fact that there is an excessive number of children aged 0 to 4 years included, and there is an imbalance in the distribution of gender and race. Therefore,

when making decisions using this data, it is necessary to confirm whether the department using it accepts patients similar to the target population.

3.5 Stability Check

I created a bootstrap sample and compared the bootstrap sample with the original Importance. As you can see in Figure 7, there is no big difference between the two, so we can assume that the analysis of Importance is stable.

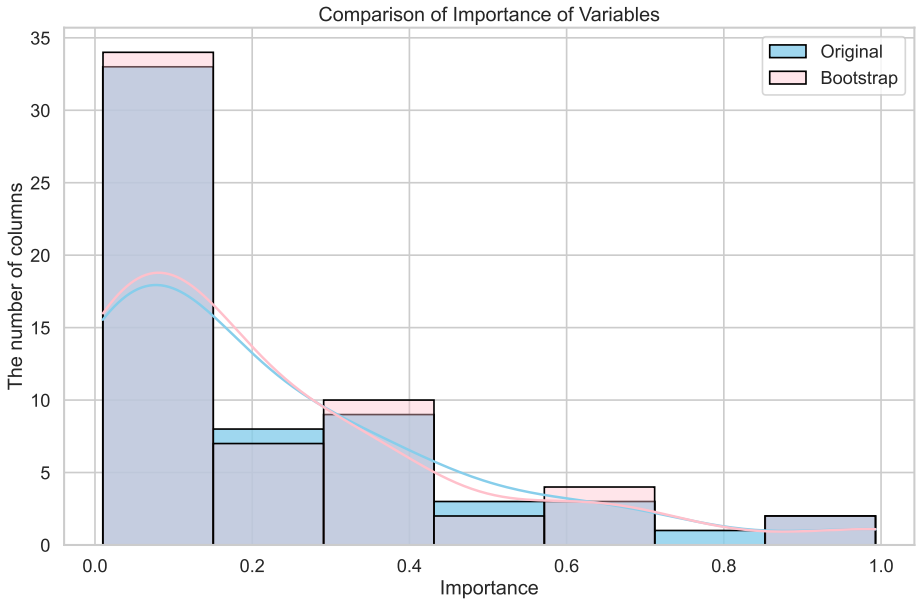

Figure 7: Effectiveness of Bootstrap

4 Modeling

4.1 Inplemention

I chose highly interpretable decision trees and logistic regression to facilitate decision making by physicians. Also, by not separating models by age, physicians do not need to use tow models, and it is also possible to apply the model even if the patient's age is unknown. For the decision tree, I restricted its depth and its number of nodes to avoid complexity. For logistic regression, it was constrained by the l1 norm to reduce the number of variables used.

For variable processing, data with missing PosIntFinal were not used in both models. For the decision tree, the missing values were converted to a form that could also handle missing values as large outliers, 99999, and fed into the model.

For logistic regression, data with missing values in AgeInMonth and GCSTotal, which are numerical data, were deleted and then the remaining data were standardized. For categorical data, one-hot encoding was performed, including missing values (missing values were also recognized as one category).

The following hyperparameters were used as candidates, and the one with the smallest evaluation function was selected using optuna and Cross Validation. As in Kuppermann (2009), the evaluation function was obtained by multiplying false negatives by 500 and adding false positives.

Hyperparameters for decision tree are shown below:

Hyperparameter	Value/Range
criterion	“gini”, “entropy”
splitter	“best”, “random”
max_depth	1 to 8
min_samples_split	2 to 20
min_samples_leaf	2 to 20
max_features	None, “sqrt”, “log2”
max_leaf_nodes	2 to 20
class_weight	None, “balanced”

Hyperparameters for Logistic Regression are shown below:

Hyperparameter	Value/Range
C	1e-3 to 1e3 (log-uniform scale)
class_weight	None, “balanced”
penalty	“l1”
solver	“saga”
max_iter	1000

The Precision Matrix for the train data is as follows. Following the model, the CT should be applied to only about 20% of patients in the decision tree and about 9% in the logistic regression, although it will miss about 60 people (0.2%) in the decision tree and 92 people (0.3%) in the logistic regression.

Decision tree train data results are shown below:

Train data	Predict	Negative	Positive
Negative		27552	6539
Positive		60	568

Logistic Regression train data results are shown below:

Train data	Predict	Negative	Positive
Negative		31562	2529
Positive		92	536

4.2 Interpretability

Decision tree can be used to easily and simply evaluate patients, as shown in Figure 8 below. In addition, since most of the logistic regression is categorical, it can be evaluated by adding up the values. However, the result of the logistic regression was different from what I expected, and almost all of the regression coefficients were not 0, so it seemed difficult to use.

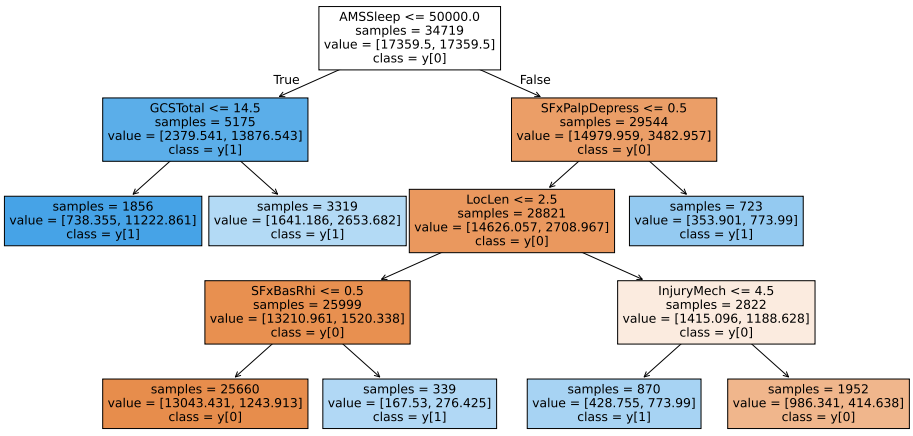

Figure 8: Decision Tree

4.3 Stability

Here, we will evaluate the results using the test data. The Precision Matrix for the test data is as follows. There were 14 patients in the decision tree and 27 in the logistic regression who actually had a ciTBI but predicted that they would not. Also, 1681 patients in the decision tree and 649 patients in the logistic regression were classified as having a ciTBI but were predicted not to have a ciTBI. This was only slightly worse than the ratio in the training data. Therefore, it is thought that the stability is high.

Decision tree test data results are shown below:

Test data	Predict	Negative	Positive
Negative		6867	1681
Positive		14	118

Logistic Regression test data results are shown below:

Test data	Predict	Negative	Positive
Negative		7899	649
Positive		27	105

5 Discussion

There were many different types of variables and a large number of data points. In addition, most of the data was categorical data, and there were many missing values, so the pre-processing and one-hot calculation costs were high, and it took a long time to process. Furthermore, due to lack of computational power, the hyperparameters did not seem to converge to the appropriate ones in the logistic regression, and the explanatory variables were not reduced well.

This analysis covers all of the three steps from checking the data collection process to applying the algorithm and evaluating the results using test data. However, as noted in the EDA, the samples collected are biased, so it is thought that the data may not correspond perfectly on a one-to-one basis. In addition, although visualization is also carried out, it is thought to be important to check whether the visuals are consistent with the physicians' sense in order to check whether the data reflects the reality.

6 Conclusion

This analysis was carried with using Kuppermann (2009). In the pre-processing of the data, variables were created with domain knowledge to reduce missing values and excessive information. In the EDA part, I identified variables that might contribute to prediction and checked for bias in the data. In the modeling part, I used decision trees and logistic regression, which are easy to interpret, and the results of the decision tree was also good in the test data and considerd that it has sufficient stability.

7 Academic honesty statement

I declare that this work was done by myself and I have not copied and pasted from other articles without citation. I followed LLM usage rules.

8 Collaborators

I did not collaborate with anyone.

9 Bibliography

Kuppermann, N., Holmes, J. F., Dayan, P. S., Hoyle, J. D., Atabaki, S. M., Holubkov, R., ... & Wootton-Gorges, S. L. (2009). Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. The Lancet, 374(9696), 1160-1170.