HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING

DATA ANALYSIS AND VISUALIZATION

Final report - Group 10

Thi-Thuy-Linh Tran 20213574
Mac-Quan Phung 20213584
Quang-Dai Tran 20210146

Instructor: Assoc. Prof. Thi-Lan Le
Msc Thi-Thom Tran

Hanoi, 07/2024

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CV	cross-validation
CRS-R	Coma Recovery Scale-Revised
DRS	Disability Rating Scale
DT	Decision Tree
ERBI	Early Rehabilitation Barthel Index
FT	Feature selection
GOS-E	Glasgow Outcome Scale-Extended
KNN	k-Nearest Neighbors
LOOCV	Leave-one-out cross-validation
MRMR	Minimum Redundancy Maximum Relevance
NB	Naïve Bayes
RLAS	Rancho Los Amigos Levels of Cognitive Functioning Scale
SBS	Sequential Backward Selection
SFS	Sequential Forward Selection
SVM	Support Vector Machine
TBI	Traumatic Brain Injury
χ^2	Chi-square

LIST OF FIGURES

LIST OF TABLES

DUTY ROSTER

No.	Name	Job
1	Tran Thi Thuy Linh	• Implement SFS and SBS • Make slides
2	Phung Mac Quan	• Main coder, implement the methods, visualization and create the interface
3	Tran Quang Dai	• Data preprocessing • Write report • Help with the code

CHAPTER 1. Introduction

Traumatic brain injury (TBI) is a serious medical condition with a wide range of potential outcomes. It occurs when a forceful blow or jolt to the head disrupts the normal function of the brain. This disruption can have lasting consequences, making TBI a leading cause of disability and death in adults. According to a 2018 study by Dewan et al. [1], an estimated 69 million people suffer from TBI worldwide each year, with a variety of causes contributing to this staggering number. The regions of Southeast Asia and the Western Pacific are disproportionately affected.

Given the prevalence and profound impact of TBI on individuals and society, the development of effective diagnostic methods is crucial. This report delves into the complexities of TBI, we explored various machine learning algorithms to identify the most suitable models for predicting TBI outcomes.

Our experiments included the most common machine learning algorithms: Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, and linear regression model. For feature selection, we experimented on Minimum Redundancy Maximum Relevance (MRMR) algorithm, Chi-Square and Pearson correlation methods. The goal was to determine which algorithm best predicts patient prognosis, thereby aiding in the development of more accurate diagnostic tools and treatment plans. Through this research, we aim to enhance the understanding of TBI outcomes and contribute to the broader efforts in mitigating the impact of this condition on patients and healthcare systems worldwide.

CHAPTER 2: Literature review

In the research of applying machine learning for TBI outcome prediction, there are two main directions. The first focuses on collecting data to create TBI databases. One of the largest is the IMPACT Database [2], which includes data from 9,205 patients with severe and moderate brain injuries. Other databases focus on specific groups, such as the work by Amorim et al. [3], which studied TBI in low- to middle-income countries and included patients aged 14 years and older with intracranial abnormalities on initial head CT scans. Another example is the work by Chong et al. [4], which focused on TBI in children under 16. The second direction involves finding suitable algorithms and models for predicting TBI outcomes, particularly evaluating whether machine learning outperforms traditional regression models. Research by Bruschetta et al. [5] and Cerasa et al. [6] concluded that ML algorithms do not perform better than more traditional regression models.

CHAPTER 3: Methodology

3.1 Models

- **Support Vector Machine (SVM)**: SVM maps data into a higher-dimensional space using kernel functions to separate classes and identifies the optimal hyperplane for classification. For Radial Basis Function (RBF) kernel:

$$K(x_j, x_k) = e^{-\frac{\|x_j - x_k\|^2}{2\sigma^2}}$$

  where x_j and x_k are the feature vectors of the j-th and k-th observations.

- **k-Nearest Neighbors (k-NN)**: Assigns an object to the most common class among its k nearest neighbors using majority voting.

- **Naïve Bayes (NB)**: Based on Bayes' Theorem, this technique estimates densities and assigns observations to the most probable class, assuming predictor independence given the class. Probabilities were calculated using a Gaussian distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}$$

- **Decision Tree (DT) [17]**: Uses a tree structure where each node represents a test on an attribute, branches represent outcomes, and leaf nodes represent class labels. Classification involves cascading tests from the root to a leaf node. We set the maximum number of decision splits to 10 and used the CART algorithm to select the best split predictor by maximizing Gain, defined as the difference between the information required to classify an object (I) and the residual information after knowing the attribute A (Ires).

3.2 Feature selection

3.2.1 MRMR

MRMR selects a subset of features by prioritizing those that are highly correlated with the target variable (maximum relevance) while minimizing redundancy between selected features.

3.2.2 Chi-square Test

The Chi-square test is a statistical method used to determine the association between features and the target variable. It compares the observed frequency distribution

of the data to the expected distribution if there were no relation between the features. Features with a high Chi-square score are considered more relevant for the target variable.

3.2.3 Pearson correlation

Pearson correlation measures the linear relationship between two continuous variables. In feature selection, it is used to identify and remove highly correlated features to reduce multicollinearity.

3.2.4 Wrapper methods

- **Sequential Forward Selection (SFS):** Starts with an empty set of features and iteratively adds one feature at a time, choosing the one that maximally improves model performance until a stopping criterion is met.
- **Sequential Backward Selection (SBS):** Begins with the full set of features and removes one feature at each iteration based on its impact on model performance.

CHAPTER 4: Experiments

4.1 Dataset

The dataset consisted of 102 patients. Data collection included demographic information and clinical assessments performed at three points: admission (T0), three months post-injury (T1), and discharge (T2), which typically occurred 6-9 months after injury.

4.1.1 Predictors' Selection:

The dataset contains following factors:

- **Demographics**: Age and Sex
- **Imaging**: Marshall Classification (T0). This widely used system assesses the severity of TBI based on a CT scan taken upon admission (T0). It considers factors like brain swelling, hemorrhages, and their location.
- **Admission Status**: Entry Diagnosis (T0). This categorical variable classifies patients' condition at admission (T0) as Vegetative State (VS), Minimally Conscious State (MCS), or Emersion from MCS.
- **Coma Recovery Scale-Revised (CRS-R)**: It assesses the level of consciousness in patients with disorders of consciousness (DOC) throughout their recovery. Scores range from 0 to 23, with higher scores indicating better function.
- **Cognitive Function**: Rancho Los Amigos Levels of Cognitive Functioning Scale (RLAS). This scale evaluates patients' cognitive abilities. It categorizes them into eight levels, ranging from No Response (lowest) to Purposeful - Appropriate (highest).
- **Disability**: Disability Rating Scale (DRS). This scale measures overall functional changes during recovery. It considers factors like arousal, awareness, cognitive abilities for self-care, dependence on others, and social adaptation. Scores range from 0 (no disability) to 29 (extreme vegetative state).
- **Early Rehabilitation**: Early Rehabilitation Barthel Index (ERBI) A and B. These extended versions of the Barthel Index assess early neurological rehabilitation progress. They cover aspects like mechanical ventilation, tracheostomy, and swallowing difficulties. Scores range from -325 (most dependent) to 100 (most independent).

4.1.2 Outcome Measure:

The Glasgow Outcome Scale-Extended (GOS-E) is used to assess the outcome of patients with TBI. It consists of 8 categories:

- **Good Recovery (Upper & Lower)**: These categories indicate minimal to no lasting problems from the TBI. Patients can resume most or all of their pre-injury activities.
- **Moderate Disability (Upper & Lower)**: These categories signify limitations in work capacity and social activities compared to pre-injury levels.
- **Severe Disability (Upper & Lower)**: Patients in these categories require varying degrees of assistance with daily activities and may be unable to work or live independently.
- **Persistent Vegetative State**: Patients are unresponsive and lack awareness of their surroundings.
- **Death**: This category indicates fatality due to the TBI.

Depending on the method, the GOS-E can be used for:

- **Binary Classification**: Split in half to have to classes: Positive Outcome and Negative Outcome
- **Multi-Class Classification**: Join the Upper and Lower categories to have 4 classes: Good Recovery, Moderate Disability, Severe Disability and Vegetative/Death.

4.2 Data preprocessing

There are no missing data in the dataset, except for the 12 death cases, which lack records of clinical assessments. We transformed the categorical fields into numerical values for easier processing in our machine learning models. Specifically, we mapped the 'Sex' field so that 'M' (Male) is represented by 1 and 'F' (Female) is represented by 0. For the 'Marshall (t0)' field, the values were mapped as follows: 'I' to 1, 'II' to 2, 'III' to 3, 'IV' to 4, 'V' to 5, and 'VI' to 6. The 'Entry Diagnosis (t0)' field was mapped so that 'VS' (Vegetative State) is 1, 'MCS' (Minimally Conscious State) is 2, and 'EMERSION' is 3. Based on the GOS-E (t2) score, we created the target variable y for both binary classification (improved vs. not improved) and multi-class classification (4 different outcome classes).

The distribution of the target variables is presented in Figure 4.1 for binary classification and Figure 4.2 for multi-class classification. We can see that the dataset is not balanced, so for training and evaluation, we used the methods of 10-fold cross-validation and leave-one-out cross-validation.

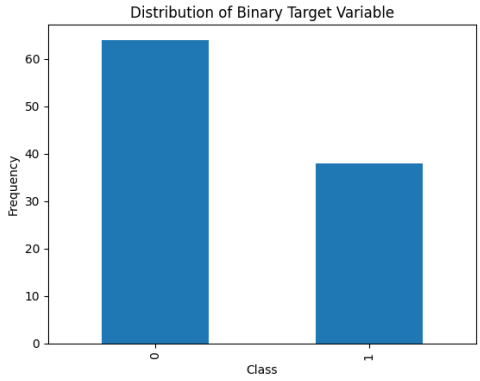

Figure 4.1 Distribution of Binary Target Variable.

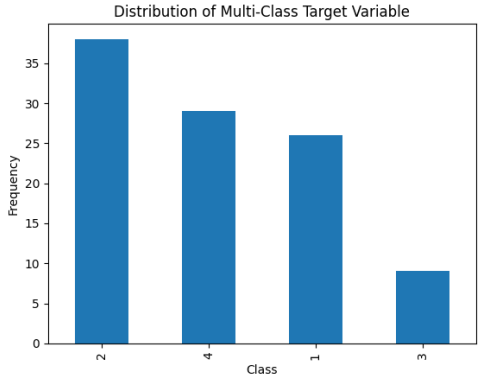

Figure 4.2 Distribution of Multi-Class Target Variable.

4.3 Evaluation metric

For evaluating the performance of the models, we used 10-fold cross-validation (CV) and leave-one-out cross-validation (LOOCV). These methods provide robust estimates of model accuracy by dividing the data into multiple subsets to ensure that each data point is used for both the training and validation processes. We measured key evaluation metrics such as Accuracy, Precision, Recall, and F1-Score for both binary and multiclass (4 classes) classification tasks.

4.4 Coding environment

In this project, we used several libraries and packages essential for data analysis and machine learning. For visualization, we used `matplotlib` library. The machine learning algorithms implemented cross validation methods and metrics are from `sklearn`. `mrmr` package and `SelectKBest` are used for feature selection. We created a simple interface where users can input diagnosis to get the prediction for the GOS-E outcome (Fig. 4.3).

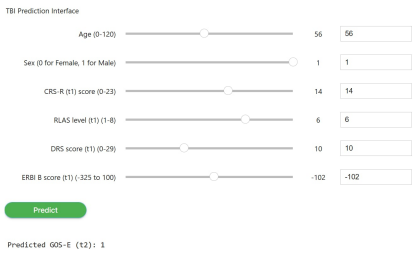

Figure 4.3 Simple user interface.

4.5 Analysis of results

The highest accuracy using 10-fold CV for 2 classes is 86.09% (KNN), 4 classes is 74.36% (SVM). When using LOOCV, highest accuracy for 2 classes is 87.25% (SVM), for 4 classes is 73.53% (SVM). Further analysis will be presented in the following parts.

4.5.1 Important features

The importance of features determined by Chi-square and MRMR are illusatrated in Figure 4.4, 4.5, 4.6 and 4.7. We can see that the important features chosen by the two methods are not completely similar, and also different for 2 and 4 classes of outcome. The common important features by two methods are Entry Diagnosis (t0), CRS-R (t1), Age, RLAS (t1) and Marshall (t0).

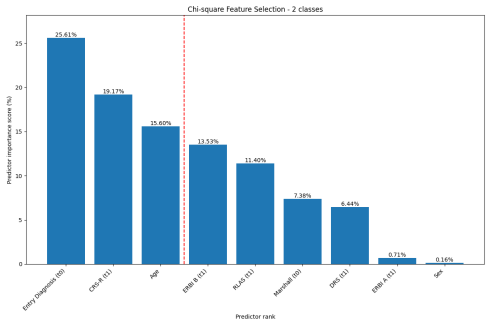

Figure 4.4 Future's importance for 2 class using Chi-Square method.

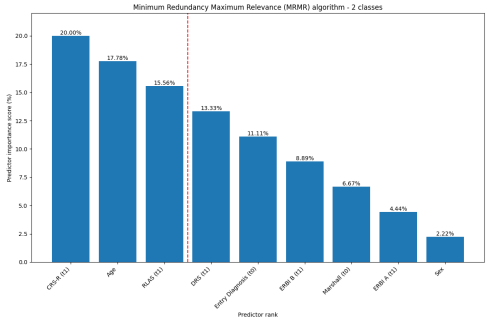

Figure 4.5 Future's importance for 2 class using MRMR method.

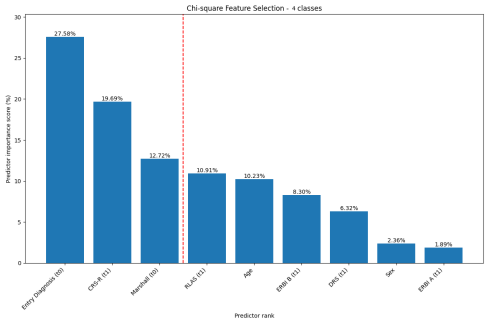

Figure 4.6 Future's importance for 4 class using Chi-Square method.

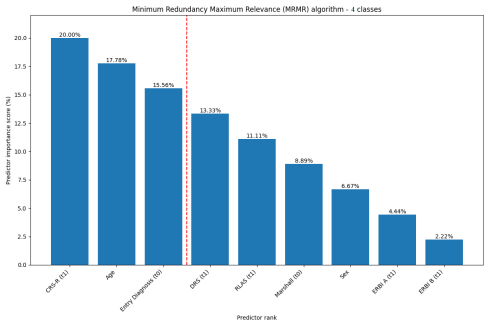

Figure 4.7 Future's importance for 4 class using MRMR method.

4.5.2 Feature selection improvements

The performances (by accuracy) of the chosen models with and without applying feature selection methods are presented in Table 4.1, 4.2, 4.3 and 4.4 We can observe that across all models and both cross-validation methods, feature selection consistently improves the accuracy. This suggests that removing irrelevant or redundant features allows the models to perform better by focusing on the most informative features.

Table 4.1 Accuracy for 2 classes, 10-fold Cross Validation

Models	without FT	MRMR	χ^2 Test	Pearson	SFS	SBS
SVM (RBF)	0.8326	0.8518	0.8427	0.8236	0.8518	**0.8527**
KNN (k=5)	0.8145	0.8518	0.8518	0.8245	0.8527	**0.8609**
Naive Bayes	0.7636	0.8118	0.8118	0.7836	**0.8136**	0.8127
Decision Tree	0.7636	0.8277	0.8127	0.7655	0.8227	**0.8318**
Logistic Regression	0.8036	**0.8518**	0.8527	0.8036	**0.8518**	**0.8518**

Table 4.2 Accuracy for 2 classes, LOO Cross Validation

Models	Without FT	MRMR	χ^2 Test	Pearson	SFS	SBS
SVM (RBF)	0.8137	0.8529	0.8431	0.8235	0.8431	**0.8725**
KNN (k=5)	0.8431	0.8529	0.8529	0.8431	0.8529	**0.8627**
Naive Bayes	0.7647	0.8235	0.8431	0.7745	0.8235	**0.8431**
Decision Tree	0.7451	**0.8137**	0.7941	0.7549	0.7941	0.7843
Logistic Regression	0.7843	0.8529	0.8333	0.7843	**0.8627**	**0.8627**

Table 4.3 Accuracy for 4 classes, 10-fold Cross Validation

Models	Without FT	MRMR	χ^2 Test	Pearson	SFS	SBS
SVM (RBF)	0.7154	0.7155	0.7155	0.7155	**0.7436**	**0.7436**
KNN (k=5)	0.7154	0.7255	0.7164	0.7155	**0.7355**	**0.7355**
Naive Bayes	0.5100	0.6955	0.6273	0.5227	**0.6964**	**0.6964**
Decision Tree	0.5690	0.6091	0.6000	0.6582	0.6555	**0.6745**
Logistic Regression	0.6481	0.6873	0.6964	0.6482	0.6964	**0.6982**

Table 4.4 Accuracy for 4 classes, LOO Cross Validation

Models	Without FT	MRMR	χ^2 Test	Pearson	SFS	SBS
SVM (RBF)	0.7059	0.7059	0.7059	0.7059	**0.7353**	**0.7353**
KNN (k=5)	0.6961	0.6961	0.7059	0.6961	**0.7255**	**0.7255**
Naive Bayes	0.5686	0.6569	0.6471	0.5686	0.6961	**0.7059**
Decision Tree	0.5882	0.5980	0.6078	0.5980	0.5980	**0.6275**
Logistic Regression	0.6275	0.6765	0.6667	0.6275	0.6863	**0.6961**

CHAPTER 5: Conclusion

In the project, we experimented with common machine learning algorithms and various feature selection methods. The results achieved are reasonable and highlight the importance of feature selection in improving model accuracy. After the project, we gained more experience and knowledge, getting to practice the lessons learned. to practice the lessons learned. For future work, we plan to explore advanced machine learning techniques and deep learning models to further enhance the prediction accuracy for TBI outcomes. Additionally, we aim to investigate the impact of different types of feature engineering and data augmentation methods on model performance.

REFERENCES

[1] M. C. Dewan, A. Rattani, S. Gupta, R. E. Baticulon, Y.-C. Hung, M. Punchak, A. Agrawal, A. O. Adeleye, M. G. Shrime, A. M. Rubiano *et al.*, "Estimating the global incidence of traumatic brain injury," *Journal of neurosurgery*, vol. 130, no. 4, pp. 1080–1097, 2018.

[2] A. Marmarou, J. Lu, I. Butcher, G. S. McHugh, N. A. Mushkudiani, G. D. Murray, E. W. Steyerberg, and A. I. Maas, "Impact database of traumatic brain injury: Design and description," *Journal of Neurotrauma*, vol. 24, no. 2, pp. 239–250, 2007, pMID: 17375988.

[3] R. L. Amorim, L. M. Oliveira, L. M. Malbouisson, M. M. Nagumo, M. Simoes, L. Miranda, E. Bor-Seng-Shu, A. Beer-Furlan, A. F. De Andrade, A. M. Rubiano *et al.*, "Prediction of early tbi mortality using a machine learning approach in a lmic population," *Frontiers in neurology*, vol. 10, p. 1366, 2020.

[4] S.-L. Chong, N. Liu, S. Barbier, and M. E. H. Ong, "Predictive modeling in pediatric traumatic brain injury using machine learning," *BMC medical research methodology*, vol. 15, pp. 1–9, 2015.

[5] R. Bruschetta, G. Tartarisco, L. F. Lucca, E. Leto, M. Ursino, P. Tonin, G. Pioggia, and A. Cerasa, "Predicting outcome of traumatic brain injury: is machine learning the best way?" *Biomedicines*, vol. 10, no. 3, p. 686, 2022.

[6] A. Cerasa, G. Tartarisco, R. Bruschetta, I. Ciancarelli, G. Morone, R. S. Calabrò, G. Pioggia, P. Tonin, and M. Iosa, "Predicting outcome in patients with brain injury: differences between machine learning versus conventional statistics," *Biomedicines*, vol. 10, no. 9, p. 2267, 2022.