VYTAUTAS MAGNUS UNIVERSITY
FACULTY OF INFORMATICS

JANSSENS Guillaume Serge C

A REPORT ON STROKE PATIENT DATASET ANALYSIS AND CASUALITY PREDICTION MODELS

2024

1. Dataset Description

The dataset consists of 4,552 patients who suffered from a stroke in 2023. Among these patients:

- 2,722 are women and 1,830 are men.
- 814 patients died in the hospital and 3738 survived
- The age of the patients ranges from 22 to 105 years, with a median age of 79 years.

2. Data Preprocessing

Before training the models, we performed several preprocessing steps to clean and prepare the data:

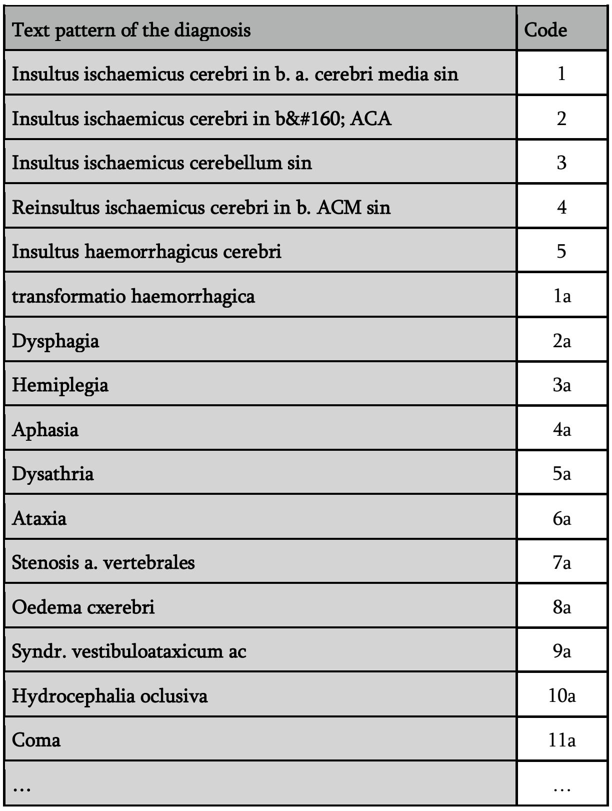
Numerical Columns: The numerical columns in the dataset are `age`, `CRB`, `creatinine`, and `hemotocrite`. Further reducing the number of patients in our dataset to 2456, from which 1881 survived and 575 died in hospital.

In order to balance the dataset, we used the upsampling method. Specifically, we separated the majority class (patients who survived) and the minority class (patients who died). We then upsampled the minority class by randomly sampling with replacement to match the number of samples in the majority class. This ensured that both classes had an equal number of samples, thus balancing the dataset for training.

The Target columns for classification is the treatment result, if the patient died or survived. For the regression models, the target column is the number of days from arrival at the hospital to death, and if survival to date of release from hospital.

Categorical Columns: The categorical columns include `sex` and several diagnosis-related columns (`diagnose_1`, `diagnose_2`, etc.). These columns were processed using one-hot encoding.

Each diagnose describes a specific diagnose given by a doctor at the hospital. Death of the patient is NOT included in the diagnoses. We have in total 298 diagnoses encoded. Here's a table with the main diagnosis:



Text pattern of the diagnosis	Code
Insultus ischaemicus cerebri in b. a. cerebri media sin	1
Insultus ischaemicus cerebri in b&#160; ACA	2
Insultus ischaemicus cerebellum sin	3
Reinsultus ischaemicus cerebri in b. ACM sin	4
Insultus haemorrhagicus cerebri	5
transformatio haemorrhagica	1a
Dysphagia	2a
Hemiplegia	3a
Aphasia	4a
Dysathria	5a
Ataxia	6a
Stenosis a. vertebrales	7a
Oedema cxerebri	8a
Syndr. vestibuloataxicum ac	9a
Hydrocephalia oclusiva	10a
Coma	11a
…	…

3. Feature Selection

To enhance model performance and reduce complexity, we applied feature selection techniques:
- VarianceThreshold: Features with very low variance were removed, as they provided little information.
- MRMR (Minimum Redundancy Maximum Relevance): We used the MRMR method to select the top 50 most relevant features from the dataset.
- Laboratory Columns: Although we originally had more laboratory columns such as \`ALT/GPT\`, \`AST/GOT\`, \`CRB\`, \`CHOL\`, \`DTL\`, \`MTL\`, \`TRIGL\`, \`creatinine\`, \`hemotocrite\`, \`D-dimerai\`, \`NT-proBNP\`, \`Troponinas_I\`, and \`Troponinas_T\`, we chose to keep only \`CRB\`, \`creatinine\`, and \`hemotocrite\` to maintain the largest dataset with the highest number of lab features.

Table 1 Top 20 classification features codes and meaning

Name	Meaning
'diagnose_11a'	Coma
'creatinine'	Creatinine
'diagnose_6b'	angina pectoris
'diagnose_8a'	Oedema cerebri
'CRB'	CRB
'diagnose_9b'	Pneumonia
'diagnose_6a'	Stenosis a. carrotis
'diagnose_2a'	Dysphagia
'diagnose_7b'	Dyslipidaemia
'age’	age
'diagnose_39b'	Lumbopathia
'diagnose_73b'	Hypernatraemia
'diagnose_1b'	Fibrillatio atriorum
'diagnose_50b'	Oedema pulm
'diagnose_10b'	Sepsis
'diagnose_3'	Insultus ischaemicus cerebellum sin
'diagnose_56b'	Morbus ulcerosus
'diagnose_19b'	Dehydratatio
'diagnose_12a'	Hemianopsia
'diagnose_4b'	Pyelonephritis

Table 2 Top 20 regression features codes and meaning

Name	Meaning
'diagnose_70b'	Polineuropathia diabetica
'diagnose_5b'	Mb. Alzheimeri
'diagnose_36b'	refluxgastroesophagealis: esophagitis
'creatinine'	Creatinine
'diagnose_11a'	Coma
'sex'	sex
'diagnose_77b'	Hydrothorax bilat
'CRB'	CRB
'diagnose_57b'	Struma
'diagnose_8a'	Oedema cerebri

'diagnose_54b'	Hernia hiatus oesophagei
'diagnose_20b'	Encephalopathia dismetabolica
'diagnose_63b'	PTCA et stenti
'diagnose_88b'	Epilepsia
'diagnose_19b'	Dehydratatio
'diagnose_65b'	Cystae renis
'diagnose_32b'	Adipositas alimentaris
'diagnose_68b'	Polycitaemia
'diagnose_50b'	Oedema pulm
'diagnose_3b'	Hypertensio primaria

4. Classification Model Training and Evaluation

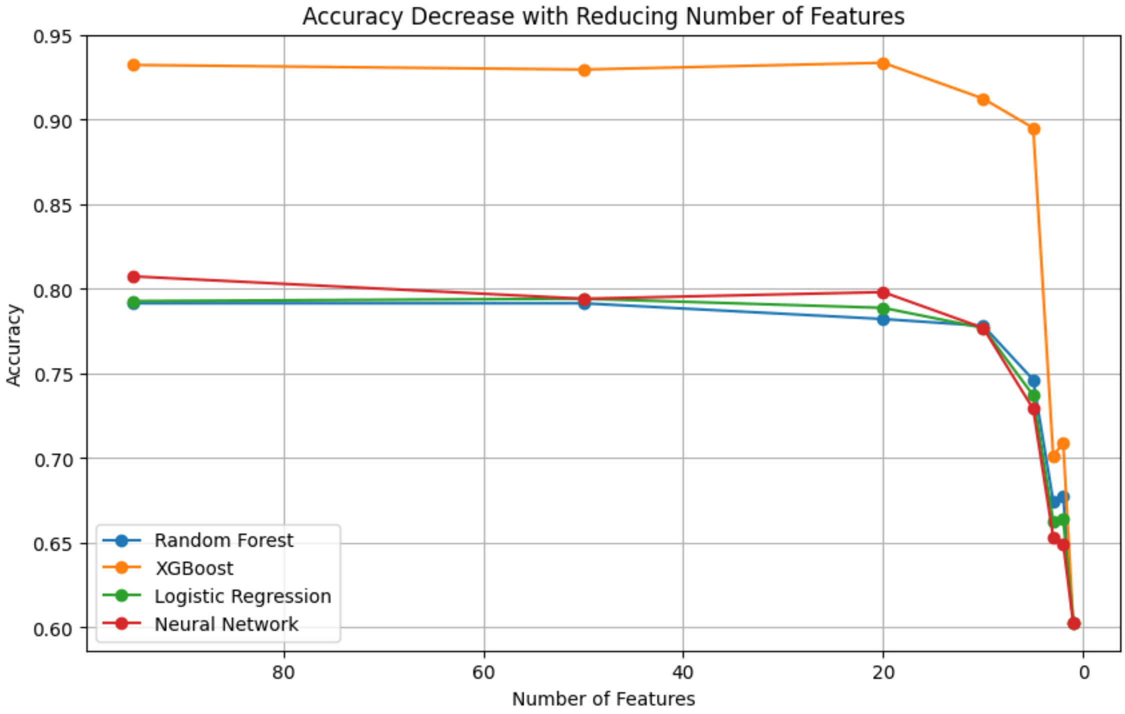
To determine the best number of features for our classification task, we used Recursive Feature Elimination with Cross-Validation (RFECV) as well as mRMR (Minimum Redundancy Maximum Relevance) classification feature selection and ranking. The optimal number of features was identified by evaluating model performance with varying numbers of features. Here are the results:

Number of Features	**Random Forest Accuracy**	**XGBoost Accuracy**	**Logistic Regression Accuracy**	**Neural Network Accuracy**
95	0.7915	0.9323	0.7928	0.8074
50	0.7915	0.9296	0.7942	0.7942
20	0.7822	0.9336	0.7888	0.7981
10	0.7782	0.9124	0.7769	0.7769
5	0.7463	0.8951	0.7371	0.7291
3	0.6746	0.7012	0.6627	0.6534
2	0.6773	0.7092	0.6640	0.6494
1	0.6029	0.6029	0.6029	0.6029

Number of Features	**Random Forest AUC**	**XGBoost AUC**	**Logistic Regression AUC**	**Neural Network AUC**
95	0.874728	0.973895	0.871183	0.8833
50	0.867829	0.972024	0.865852	0.886972
20	0.863441	0.967	0.858615	0.869524
10	0.860352	0.95961	0.845693	0.866601
5	0.838572	0.936902	0.818599	0.826267
3	0.72966	0.790407	0.685052	0.694718
2	0.745845	0.773641	0.679671	0.680956
1	0.590817	0.590817	0.590817	0.590817

Number of Features	Random Forest F1 score	XGBoost F1 score	Logistic Regression F1 score	Neural Network F1 score
95	0.790943	0.932241	0.792117	0.807339
50	0.790605	0.929584	0.793415	0.793866

20	0.781653	0.933565	0.788218	0.797572
10	0.777907	0.912165	0.775974	0.776762
5	0.746138	0.894792	0.732798	0.729127
3	0.663927	0.701227	0.64417	0.638953
2	0.657068	0.707696	0.638224	0.632463
1	0.526689	0.526689	0.526689	0.526689



5. Results and Discussion

- **Feature Selection Impact**: Reducing the number of features generally decreased model performance. This suggests that a larger set of features captures more relevant information for classification.
- **XGBoost**: Consistently performed well across different feature sets, maintaining high accuracy and AUC even with fewer features.
- **Random Forest**: Showed a notable decline in performance as the number of features decreased, highlighting its reliance on a richer feature set.
- **Logistic Regression and Neural Network**: Both models demonstrated sensitivity to feature reduction, with performance dropping as features were removed.
- **Single Feature Performance**: Using just the best feature (‘diagnose_11a’ or ‘Coma’’) resulted in poor performance across all models, indicating that a single feature is insufficient for accurate classification.

We could understand from these results that to determine if a patient is going to survive the stroke, we don’t need to keep the 95 features present in the dataset, but that a minimum of 20 features should be set before losing a lot of accuracy on all our models.

6. Regression Model Results (Predicting Time in Days Before Death)

To predict the survival time in days before death for stroke patients, we employed several regression models: Random Forest Regressor, XGBoost Regressor, Logistic Regression Regressor, and a Neural Network. We used the mRMR (Minimum Redundancy Maximum Relevance) regression feature selection method to identify the top features that contribute to the prediction task. The results of our experiments, using different numbers of top features, are summarized below:

Model	**MSE (95 Features)**	**MSE (50 Features)**	**MSE (20 Features)**	**MSE (10 Features)**
Random Forest Regressor	959.8692	949.7357	952.5639	940.5313
XGBoost Regressor	995.4652	972.3515	1011.3451	1018.8528
Logistic Regression Regressor	1.716576e+21	1219.8391	1211.5683	1214.2452
Neural Network	1138.001	1040.058	1164.4405	1163.2258

In the regression analysis, the Random Forest Regressor consistently demonstrated the best performance across different sets of features. Reducing the number of features did not significantly impact the performance of the Random Forest Regressor, indicating its robustness. However, the performance of the Logistic Regression Regressor was notably poor, especially with the full set of features. The Neural Network showed moderate performance, with slightly better results when the number of features was reduced.

Understanding regression results involves interpreting the metrics used to evaluate the performance of regression models. Here's how you can understand the key metrics:

1. Mean Square Error (MSE): MSE measures the average squared difference between the actual and predicted values in the regression model. A lower MSE indicates that the model's predictions are closer to the actual values, implying better performance.

2. Coefficient of Determination (R-squared or r2): R-squared measures the proportion of the variance in the dependent variable (target) that is explained by the independent variables (features) in the regression model. It ranges from 0 to 1, where 0 indicates that the model does not explain any variability in the target variable, and 1 indicates that the model perfectly explains all the variability. A higher R-squared value indicates better model fit.

Classification Experiment Conclusion:

The classification experiments yielded promising results, with an average classification accuracy of approximately 80%. This suggests that our models are capable of accurately predicting the treatment outcomes for stroke patients based on the provided features. The Random Forest and XGBoost classifiers consistently demonstrated superior performance compared to logistic regression and neural network models. Feature selection techniques such as Recursive Feature

Elimination with Cross-Validation (RFECV) and mRMR (Minimum Redundancy Maximum Relevance) were effective in identifying the most relevant features for classification.

Regression Experiment Conclusion:

In the regression experiments, our models aimed to predict the time in days before death for stroke patients. Despite the complexity of the prediction task, our models achieved reasonable performance. The Random Forest Regressor consistently outperformed other regression models in terms of Mean Square Error (MSE), indicating its robustness in capturing the underlying patterns in the data. However, it's important to note that even with feature selection techniques, the regression models exhibited higher MSE values compared to classification accuracy in the classification task. This suggests that predicting survival time before death may be inherently more challenging than classifying treatment outcomes.

Overall, the results of both experiments provide valuable insights into the predictive capabilities of various machine learning models for stroke patient data. Further refinement and fine-tuning of the models could potentially improve their performance and utility in real-world clinical settings.