PATTERN RECOGNITION AND MACHINE LEARNING

COURSE PROJECT : Stroke Prediction

ABSTRACT :

This paper reports our experience with building a Stroke Prediction classifier. We have a dataset which consists of features as some common reasons for why a person gets a stroke. We use various classification algorithms and compare their results in this report.The task was to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status.

INTRODUCTION :

Stroke is a blood clot in the brain, which can cause permanent damage that has an effect on mobility, cognition, sight or communication. Stroke is the second leading cause of death worldwide and one of the most life- threatening diseases for persons above 65 years. Once a stroke disease occurs, it not only costs huge medical care and permanent disability but can eventually lead to death. Every 4 minutes someone dies of stroke, but up to 80% of stroke can be prevented if we can identify or predict the occurrence of stroke in its early stage. It would be very helpful for society.

DATASET :

The dataset healthcare-dataset-stroke-data.csv is used as the training dataset. The train dataset contains 5110 rows where each row represents a train-person with 12 columns containing :

- 1 id column
- 1 stroke column : Class label - 0 for no stroke and 1 for stroke
- 10 input parameter columns

The dataset has been split into train and test with test size of 0.3

METHODOLOGY :

CONTENTS OF COLAB NOTEBOOK

- Importing libraries
- Importing dataset
- Exploratory Data Analysis
- Categorical Data Distribution
- Numerical Data Distribution
- Data Preprocessing
- Implementing different models
- Feature Selection
- Building Final model

OVERVIEW

There are various classification algorithms present out of which we shall implement the following

- Logistic Regression
- Decision Tree Classification
- Naive Bayes
- KNN

● MLP
● Random Forest Classification
● SVM
We also make use of QDA and LDA for dimensionality reduction and SFS AND SFFS for feature selection.

Exploring the dataset and pre-processing

- We dropped the "id" column, since there is no duplicate data.
- We have shown **target distribution**.
- In **categorical data distribution** we have shown **count plot** for different columns, **count plot by target** for the same, **stacked bar plot** for smoking status attribute.
- The "Other" category in the "gender" attribute appears only once. We replace it with the mode of gender attribute.
- We have four categories in the "smoking_status" attribute. There is an "unknown" category, which is like a null value. It can be confusing for machine learning models, hence we replaced it with the highest probability of getting a stroke or not getting a stroke.
    - ❖ "unknown" with "never smoked" for a patient who is not getting a stroke
    - ❖ "Unknown" with "formerly smoked" for a patient who is getting a stroke.
- For **numerical data distribution**, we have shown **box plot**, **histogram**, **pair plot**, and **heatmap correlation** of numerical columns.
- For **data preprocessing,**
    - ○ we have replaced null values of the "bmi" column with the median.
    - ○ Feature encoding of "gender", "ever_married", "work_type", "residence_type" and "smoking_status" columns.
    - ○ For train test split , we have used stratify to make sure that the proportion of target variable on train and test remains balanced.
    - ○ For feature scaling, we have used standardization.

Implementation of classification algorithms

● *Logistic Regression* : Logistic Regression model is widely used for binary classification and hence is good suited for
classification into stroke vs no stroke.
● *Decision Tree Classifier* : Decision Tree Classifier is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
We calculated the best max depth for the tree and used it for finding accuracy.
● *Naive Bayes* : Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
○ Three types of Naive Bayes were used :
■ Gaussian NB
■ Bernoulli NB
■ Multinomial NB
● *KNN (k - nearest neighbors)* : KNN are supervised algorithms which classify on the basis of distance from similar points. Here k is the number of nearest neighbors to be considered in the majority voting process.
● *Multilayer Perceptron* : MLP is a feedforward Neural Network which uses backpropagation to update weights and improve
the results.
● *Random Forest Classifier :* Random Forest Classifiers use boosting ensemble methods to train upon various decision trees

and produce aggregated results.It is one of the most used machine learning algorithms.
We calculated the best max depth for the tree and used it for finding accuracy.
● *Support Vector Machine :* In SVM , data points are plotted into n-dimensional graphs which are then classified by drawing hyperplanes. We used it with a polynomial kernel.
Dimensionality reduction techniques used : QDA and LDA
Feature Selection Technique : SFS and SFFS

EVALUATION OF MODELS :

IMPLEMENTING MODELS

MODELS		ACCURACY	F1 SCORE	PRECISION	RECALL
Logistic Regression		0.948467	0.97	0.95	1
Decision Tree Classification	default	0.91389	0.95	0.96	0.95
	Max_depth = 2	0.951076	0.97	0.95	1
Naive Bayes Classification	GaussianNB	0.86366	0.93	0.97	0.88
	BernoulliNB	0.9491	0.97	0.95	1
	MultinomialNB	0.7912	0.88	0.97	0.81
KNN		0.9439	0.97	0.95	0.99
Multi layer perceptron		0.9504	0.97	0.95	1
Random forest classification	Max_depth = 2	0.95107	0.97	0.95	1
	Max_depth = 18	0.9517	0.98	0.95	1
Support vector machine (SVM)		0.95107	0.97	0.95	1

DIMENSION REDUCTION

Dimension reduction method	Accuracy
LDA (Linear discriminant analysis)	0.9432
QDA (Quadratic discriminant analysis)	0.87801

FEATURE SELECTION

Feature selection method	k	CV score
Sequential forward selection	4	0.9513

Sequential forward floating selection	4	0.9513

RESULTS AND ANALYSIS :

For final model we have used Logistic Regression

	Precision	Recall	F1-score
0	0.95	1	0.97
1	-	-	0.02
accuracy	0.17	0.01	0.95
Macro avg	0.56	0.5	0.5
Weighted avg	0.91	0.95	0.93

CONFUSION MATRIX

	Not stroke	Stroke
Not stroke	1453	5
Stroke	74	1

CONCLUSION

- We've got around 95% accuracy but it only has ~1% TPR. It will lead to a fatal prediction because we can't identify someone who actually has a stroke.
- Getting high TPR is more preferred in this case.

CONTRIBUTORS :

Nitya Anand Shah (B20CS039)	Niharika Manhar (B20CS038)
• Exploratory data analysis • Categorical data distribution • Implementing different models • Building final model • Report	• Numerical data distribution • Data preprocessing • Implementing different models • Feature selection • Report

REFERENCES :

[1] K Nearest Neighbor | KNN Algorithm | KNN in Python & R (analyticsvidhya.com)

[2] Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith G

[3] Pattern Classification -Book by David G. Stork, Peter E. Hart, and Richard O. Duda

[5] Link for dataset : Kaggle