Stroke Prediction using Machine Learning

Nisarg Doshi	202111033
Utkarsh Pandya	202111026

May 1, 2022

Introduction

- A stroke[1] is a medical condition in which poor blood flow to the brain causes cell death. It cause parts of the brain to stop functioning properly.

- Stroke was the second most frequent cause of death worldwide in 2011, accounting for 6.2 million deaths (11% of the total).

- It is believed that high blood pressure, high cholesterol, smoking, obesity, and diabetes are leading causes of stroke.

- Stroke can be prevented if people change their lifestyle and habits.

[1] Martin G (2009). Palliative Care Nursing: Quality Care to the End of Life, Third Edition. Springer Publishing Company. p. 290. ISBN 978-0-8261-5792-8. Archived from the original on 2017-08-03.

Problem Statement

- Our objective is to detect whether an individual is likely to get a stroke based on parameters like gender, age, bmi, work type etc.

- To pursue this objective we have used 4 classifiers:
  1. k-Nearest Neighbours (kNN)
  2. Decision Tree
  3. Logistic Regression
  4. Support Vector Machine (SVM)

Dataset

- For this objective, we have used “Stroke Prediction dataset”[2] which consists following attributes of individuals from age 0 to 82 years: (5120 entries, 12 columns)

  1) id
  2) gender
  3) age
  4) hypertension
  5) heart_disease
  6) ever_married
  7) work_type
  8) Residence_type
  9) avg_glucose_level
  10) bmi
  11) smoking_status
  12) stroke

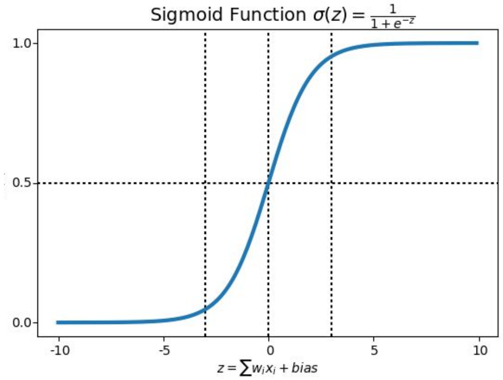
[2] Kaggle.com. 2022. Stroke Prediction Dataset. [online] Available at:
¡https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Data Preprocessing

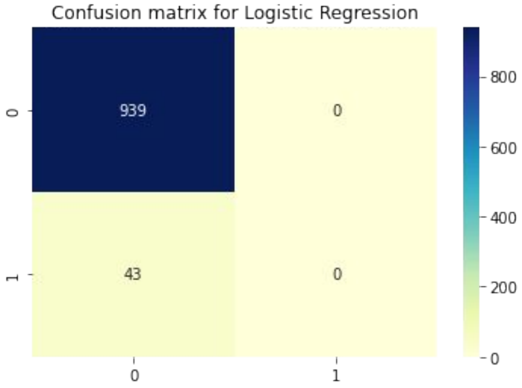
- Missing values in bmi column
- Label Encoding & One Hot Encoding for gender, heart_disease, work_type and other categorical features
- Size: 4909 entries, 20 columns
- Train Test Split: 80% training, 20% testing

Logistic Regression

- Logistic Regression uses the sigmoid function and Confusion matrix for predictions on test dataset



Sigmoid function[3]

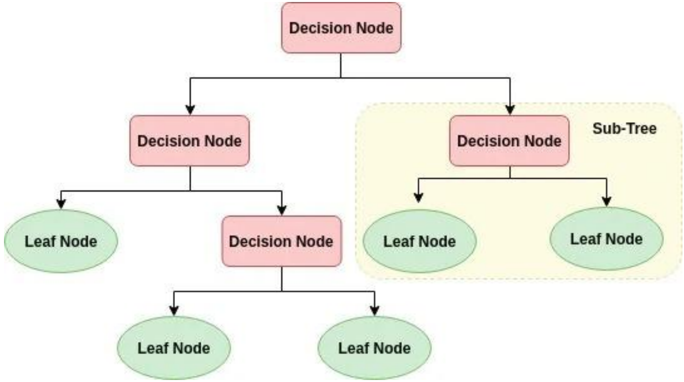


Confusion matrix (result)

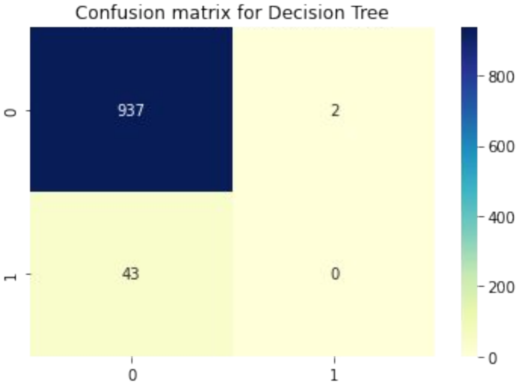
[3] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant.
Applied logistic regression. Vol. 398. John Wiley Sons, 2013

Decision Tree

- Decision tree concept and Confusion matrix for predictions on test dataset



Decision Tree[4]

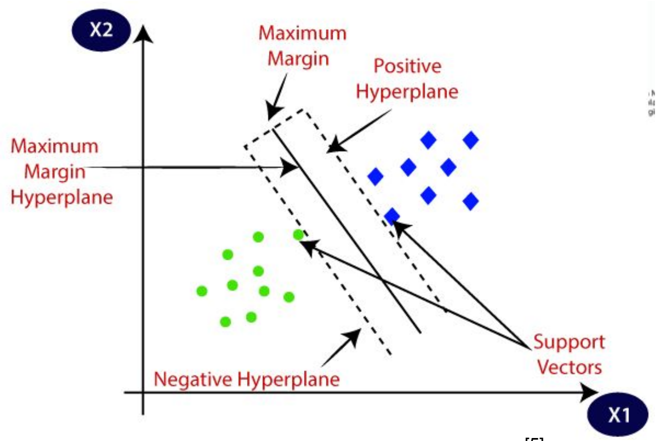


Confusion matrix (result)

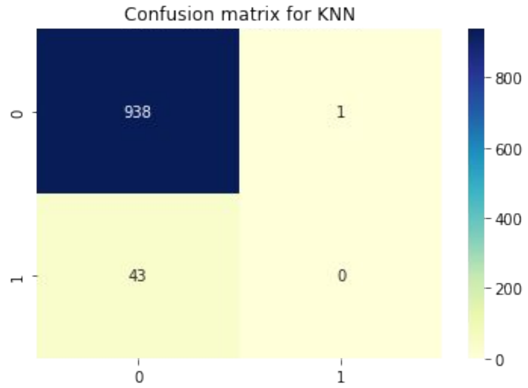
[4] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Support Vector Machine (SVM)

- SVM concept and Confusion matrix for predictions on test dataset



Support Vector Machine[5]

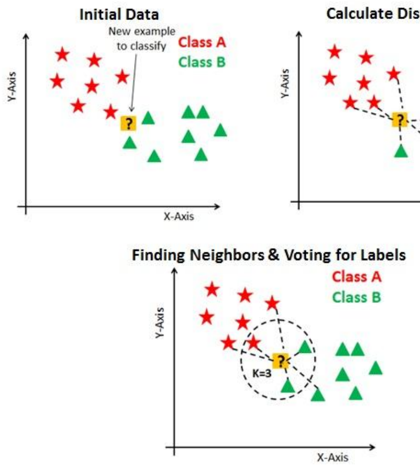


Confusion matrix (result)

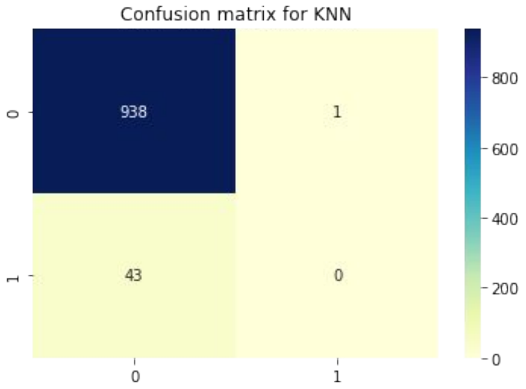
[5] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

k-Nearest Neighbours

- kNN concept and Confusion matrix for predictions on test dataset



KNN Classifier[6]



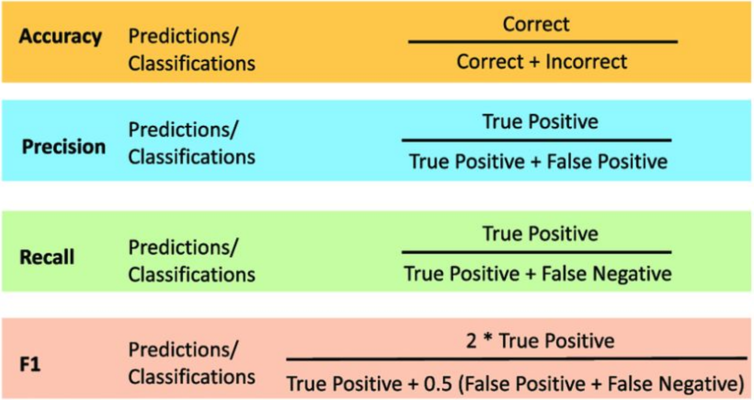
Confusion matrix (result)

[6] https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

Metrics

- The following metrics are widely used in ML applications to check the performance of the models:

- As, this is a medical application and false negatives should be minimum so we are more interested in Recall as a metric.



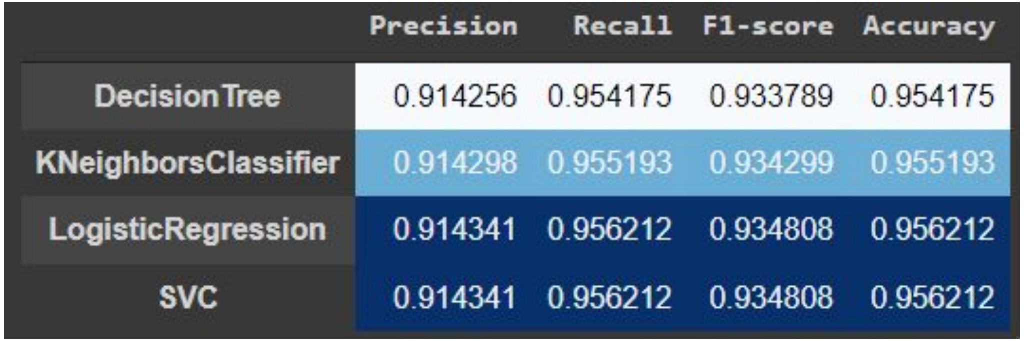
Accuracy	Predictions/ Classifications	$\frac{\text{Correct}}{\text{Correct + Incorrect}}$
Precision	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Positive}}$
Recall	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
F1	Predictions/ Classifications	$\frac{\text{2 * True Positive}}{\text{True Positive + 0.5 (False Positive + False Negative)}}$

Performance Metrics[7]

[7] Ping Shung, K., 2022. Accuracy, Precision, Recall or F1?. [online] Medium. Available at :https://towardsdatascience.com/accuracyprecision-recall-or-f1-331fb37c5cb9

Results

- Logistic regression and SVM are performing better in terms of recall as well as accuracy than other models.



	Precision	Recall	F1-score	Accuracy
DecisionTree	0.914256	0.954175	0.933789	0.954175
KNeighborsClassifier	0.914298	0.955193	0.934299	0.955193
LogisticRegression	0.914341	0.956212	0.934808	0.956212
SVC	0.914341	0.956212	0.934808	0.956212

Scope of Improvement

- Due to limitation of dataset we cannot conclude with surety the effect of certain attributes in prediction.
- Like people who had hypertension is significantly lower than the number of people who didn't.
- Further, we can also apply Principal Component Analysis (PCA) to reduce dimensionality.

Thank You