Stroke and Diabetes Prediction using Machine Learning Algorithms

Nehal Mostafa, Aya Ehab, Radwa Abd El-Hakeem
AAST, Egypt, nonamostafa767@gmail.com, ayaaehabb111@gmail.com, eadwahakeem57@gmail.com

Supervisor: Nashwa El-Bendary, Dr Professor
AAST, Egypt, Nashwa.elbendary@aast.edu

ABSTRACT-

Diabetes is a disease that has no permanent cure; hence early detection is required. It is a dreadful disease identified by escalated levels of glucose in the blood. Machine learning algorithms help in identification and prediction of diabetes at an early stage. The main objective of this study is to predict diabetes mellitus with better accuracy using an ensemble of machine learning algorithms. machine learning (ML) algorithms, and K-fold Cross Validation; Accuracy are used in Predicting Diabetes (PD) dataset in our research, collected from the Kaggle Machine Learning. The dataset contains information about 768 patients and their corresponding nine unique attributes and has been considered for experimentation, which gathers details of patients with and without having diabetes. The proposed ensemble soft voting classifier gives binary classification and uses the ensemble of three machine learning algorithms. random forest, K-Nearest Neighbors (KNN), and Naive Bayes for the classification. Empirical evaluation of the proposed methodology has been conducted with state-of-the-art methodologies and base classifiers such as K-Nearest Neighbors (KNN). by taking accuracy, precision, recall and specificity as the evaluation criteria. The proposed ensemble approach gives the highest accuracy, precision, recall and specificity value with 77.922%, 83.006%, 83,552% and 67.088% respectively on the Prediction Diabetes (PD) dataset. Further, the efficiency of the proposed methodology has also been compared and analyzed with Stroke Prediction dataset. The proposed ensemble soft voting classifier has given accuracy, precision, recall and specificity value with93.83%,92.59%,96.12% and 91.91% on Stroke Prediction dataset using Random Forest Algorithm.

Keywords:
Diabetes, Prediction, Stroke, KNN, random forest

1.INTRODUCTION

To extract hidden patterns and relationships from large data bases, data mining merges statistical analysis, machine learning and database technology. Diabetes is a chronic disease which causes serious health complications including heart disease, kidney failure and blindness. It is a major risk factor for cardiovascular disease (disease of the heart and circulatory system). Diabetes also increases the risk of micro-vascular damage and macro vascular complications. Thus, diabetes is found to be one of the leading causes of global death by disease. Around 366 million people have diabetes worldwide according to statistics taken in the year 2011. Also, it has been projected that the people with diabetes will increase to around 552 million by the year 2030, diabetes also can play a role in increasing the risk of a stroke. A stroke is a cerebrovascular disease in which arteries carrying oxygen and nutrients to the brain gets ruptured and there is no blood supply to the parts of the brain. This result in complete damage of blood cells in the brain. A fairly many people are losing lives, especially in developing countries. According to the reports of the American Heart Association, the mortality rate for 2017 was 37.6 in every 100,000 stroke cases. According to the World Health Organization, stroke has been classified as a non-communicable disease. The reports of 2012 say that Stroke was the main cause of death due to non-communicable disease, causing 17.5 million deaths. It is also the fourth major cause of death in India. It is also the fifth major cause of death in the United States. Nearly 800,000 people have a stroke per year which equates to one person every 40 s. Diabetes is a well-established risk factor for stroke. It can cause pathologic changes in blood vessels at various locations and can lead to stroke if cerebral vessels are directly affected. Additionally, mortality is higher and poststroke outcomes are poorer in patients with stroke with uncontrolled glucose levels.

2. METHODOLOGY

The methodology for predicting stroke and diabetes diseases was done by using two algorithms, one for the diabetes and the other for the stroke and the results of both are shown in figure describes the architectural diagram for predicting both stroke and diabetes.

1. random forest
2. KNN

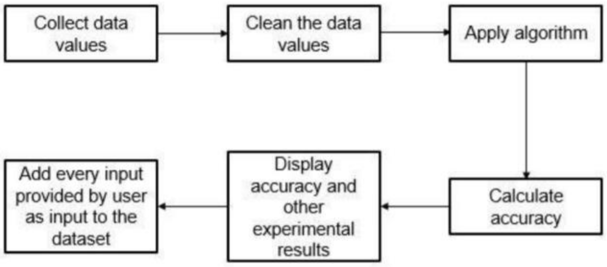

Collect data values
Clean the data values
Apply algorithm
Calculate accuracy
Display accuracy and other experimental results
Add every input provided by user as input to the dataset

Figure 1 Methodology to predict stroke and diabetes

Random Forest Algorithm

A. Random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms, the general idea of this algorithm is that a combination of learning models increases the overall results, it builds multiple decision trees and then they get merged together to choose the most accurate prediction

1. Dataset collection and pre-processing

   The dataset that was used in predicting stroke is healthcare-dataset-stroke-data obtained from Kaggle.

   The dataset has 12 attributes

2. Implementation

   The implementation of random forest works as follows:

   a. Load the stroke disease dataset
   b. After preprocess, split the heart disease into train and test with proportion of 70:30 using Random Forest Classifier function
   c. Then K-Fold validation is used
   d. Train model using train set
   e. Make prediction using on the test fold
   f. Map predictions to outcomes (0 or 1)
   g. Accuracy is calculated

Accuracy = (TP+TN)/(TP+FP+FN+TN)

Where:

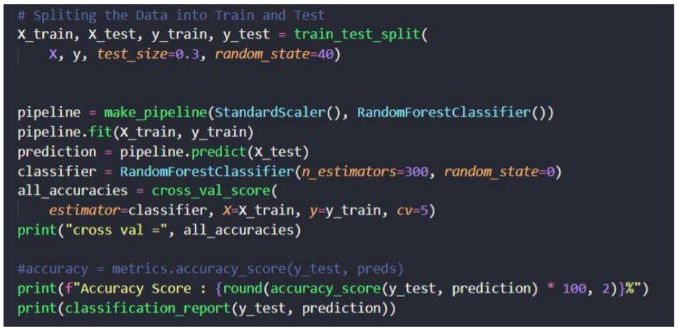
TP- True Positive (predicted having disease, actually does).

TN-True Negative (prediction is no, actually don’t have the disease.)

FP-False Positive (predicted yes, but don’t actually have the disease).

FN-False Negative (predicted no, but actually have the disease.

The accuracy of the prediction by using random forest algorithm is 93.83%

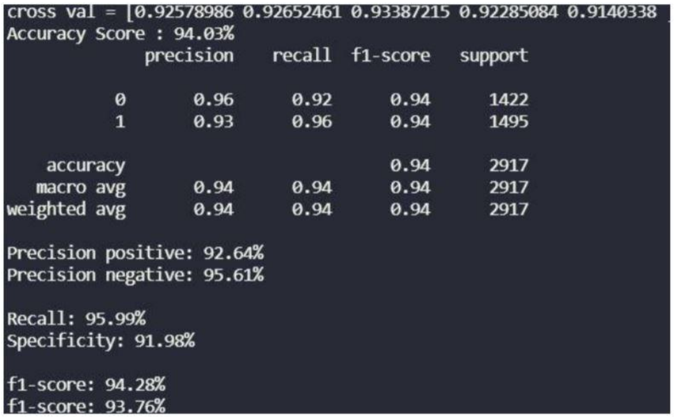


```
# Spliting the Data into Train and Test
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=40)


pipeline = make_pipeline(StandardScaler(), RandomForestClassifier())
pipeline.fit(X_train, y_train)
prediction = pipeline.predict(X_test)
classifier = RandomForestClassifier(n_estimators=300, random_state=0)
all_accuracies = cross_val_score(
    estimator=classifier, X=X_train, y=y_train, cv=5)
print("cross val =", all_accuracies)

#accuracy = metrics.accuracy_score(y_test, preds)
print(f"Accuracy Score : {round(accuracy_score(y_test, prediction) * 100, 2)}%")
print(classification_report(y_test, prediction))
```

Figure 2 Random Forest sample



```
cross val = [0.92578986 0.92652461 0.93387215 0.92285084 0.9140338
Accuracy Score : 94.03%
              precision    recall  f1-score   support

           0       0.96      0.92      0.94      1422
           1       0.93      0.96      0.94      1495

    accuracy                           0.94      2917
   macro avg       0.94      0.94      0.94      2917
weighted avg       0.94      0.94      0.94      2917

Precision positive: 92.64%
Precision negative: 95.61%

Recall: 95.99%
Specificity: 91.98%

f1-score: 94.28%
f1-score: 93.76%
```

Figure 3 Experimental Results

B. KNN Algorithm

KNN is a supervised learning algorithm that produces output by storing all available cases and classifies new cases according to the similarity measures, we used this algorithm in predicting whether the patient has diabetes or not

1. Implementation

   The implementation of KNN is as follows:

   a. Calculate Euclidean distance between the new point and existing points
   b. distance = √ [(x2 – x1)2 + (y2 – y1)2].
   c. Choose value of K and select K neighbors closest to the new point
   d. Count the votes of all the k neighbors
   e. Predict the class
   f. Map predictions to outcome
   g. Calculate accuracy

3-RESULT

Results from Random Forest and KNN algorithms used in predicting stroke and diabetes

ALGORITHM	ACCURACY
RANDOM FOREST	93.83%
KNN	80.08%

4-CONCLUSION AND FUTURE SCOPE

we predict if the patient has stroke or not by using random forest algorithm (according to our features in the data set). And also, we predict if the patient has diabetes

or not by using K-Nearest neighbor algorithm (according to features in the data set). So, by using these algorithms we can predict the result.

5-REFERENCES

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0998-2

https://www.researchgate.net/publication/342437236_Prediction_of_Stroke_Using_Machine_Learning

https://link.springer.com/chapter/10.1007/978-3-030-65621-8_15

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5298897/#:~:text=Diabetes%20is%20a%20well%2Destablished,stroke%20with%20uncontrolled%20glucose%20levels.

https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/

https://builtin.com/data-science/random-forest-algorithm