Stroke Predictive Model

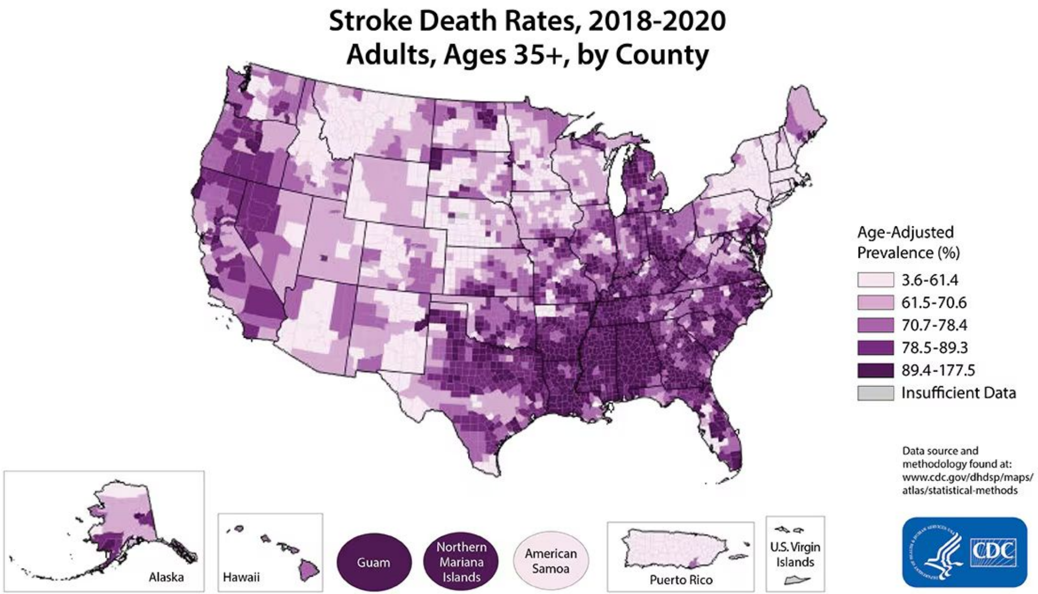
Leveraging Machine Learning to Predict Stroke Risk

By: Kanishk Sivanandam and Chetan Maviti

Project Objective

Goal: To develop a predictive model that determines whether or not someone is at risk of getting a stroke using patient health and lifestyle data.

Impact: Enhancing healthcare by enabling early intervention and personalized patient care for strokes.



Dataset Overview

Dataset Size:

- 12,760 instances
- 27 dimensions.

Some Features:

- Patient age
- Cholesterol levels (HDL,LDL)
- Blood Pressure (Systolic/Diastolic)
- Physical Activity (Low, Moderate, High)

Class: **Diagnosis (Stroke, No Stroke)**

Factors that you can control account for 82% to 90% of all strokes:

- High blood pressure
- Obesity
- Physical inactivity
- Poor diet
- Smoking

Other risk factors are based on lifestyle, genetics, and environment.

- **Age** is a risk factor, too. A stroke can occur at any age, but the risk is higher for babies under the age of 1 and for adults as they grow older.
- **Anxiety, depression, and high stress levels,** as well as working long hours and not having much contact with family, friends, or others outside the home, may raise your risk for stroke.
- **Family history and genetics** play a role as well. Your risk of having a stroke is higher if a parent or other family member has had a stroke, particularly at a younger age. Certain genes affect your stroke risk, including those that determine your blood type. People with blood type AB (which is not common) have a higher risk.
- **Living or working in areas with air pollution** can also contribute to stroke risk.
- **Other medical conditions,** such as sleep apnea, kidney disease, and migraine headaches, are also factors.
- **Other unhealthy lifestyle habits,** including drinking too much alcohol, getting too much sleep (more than 9 hours), and using illegal drugs such as cocaine, may raise stroke risk.

(From https://www.nhlbi.nih.gov/)

Preprocessing

1. **Symptoms Attribute Clean Up**
   List of 10 Symptoms (Ex. Blurred Vision), split into 10 binary attributes. Removed missings.

2. **Clean Up Unnecessary Columns**
   Patient name, Patient ID, Record date, "Cholesterol_levels", "Blood_Pressure_Levels"

3. **Clean Up Missing Values**
   Remove empty blood pressure data. ReplaceMissingValues filter (mean & mode). MathExpression filter + floor function

4. **Label Encoding**
   Convert non-binary nominal attributes to numeric (Ex. Keto Diet to 1). LabelEncoder method Sklearn

5. **Data Normalization**
   Ranges from 0-1 to 100-200. Normalized using Normalize filter on 0-1 scale. Rounded to thousandths using pandas.

Methods of Attribute Selection

OneR

Simple, but effective, OneR builds a set of rules for each attribute to predict the class variable (Diagnosis), then selected the single set of rules that yields the lowest error rate.

CorrelationAttributeEval

Measures the Pearson correlation between each numeric attribute and the class variable. Attributes with high correlation scores are prioritized, helping identify the most relevant features for the model.

Methods of Attribute Selection cont.

PrincipalComponents (PCA)

A technique that transforms a set of possible correlated variables into principal components consisting of various ratios of the original features. These component vectors are used as new attributes, and they are ranked by how much variance in the dataset each one captures.

ReliefF

Determine how relevant attributes are by comparing how well an attribute can distinguish two similar (nearby) instances. It does this by sampling instances and checking if nearby instances of different classes have similar values for the attribute, therefore identifying features that are useful for class separation.

Methods of Attribute Selection cont.

Personal Selection

Removed patient gender, dietary habits, work type of patients, metabolic equivalent of task score, marital status, alcohol intake, residence type, and all Symptoms ("Blurred Vision," "Seizures," "Difficulty Speaking," "Weakness," "Confusion," "Headache," "Dizziness," "Severe Fatigue," "Loss of Balance," and "Numbness")

OneR Attribute Selection

Chosen Cut-Off Value: **50.55**

Remaining Attributes:

LDL_Cholesterol, Family_History_of_Stroke, Patient_Age, Severe Fatigue, Stroke_History, Blurred Vision, and Weakness

```
Ranked attributes:
51.10639     4 LDL_Cholesterol
50.89381    16 Family_History_of_Stroke
50.84549     1 Patient_Age
50.71988    29 Severe Fatigue
50.61359    21 Stroke_History
50.5846     22 Blurred Vision
50.55561    25 Weakness
50.54595    17 Residence_Type
50.51696    15 Hypertension
50.401      27 Headache
50.39134    23 Seizures
50.35269    12 Body_Mass_Index
50.33337     7 Marital_Status
50.27539    30 Loss of Balance
50.27539     2 Patient_Gender
50.18842    13 Alcohol_Intake
```

CorrelationAtributeEval Attribute Selection

Chosen Cut-Off Value: **0.01**

Remaining Attributes:

Average_Glucose_Level, Smoking_Status, Residence_Type, Systolic_BP, Hypertension, Stress_Levels, Weakness, Stroke_History, Blurred Vision, Severe Fatigue, HDL_Cholesterol, Family_History_of_Stroke

```
        Correlation Ranking Filter
Ranked attributes:
 0.017893    16 Family_History_of_Stroke
 0.0161245   14 HDL_Cholesterol
 0.0161027   29 Severe Fatigue
 0.0136457   22 Blurred Vision
 0.012273    21 Stroke_History
 0.0122426   25 Weakness
 0.0117823    9 Stress_Levels
 0.0115932   15 Hypertension
 0.0111576   18 Systolic_BP
 0.010922    17 Residence_Type
 0.0104333   19 Smoking_Status
 0.010073    10 Average_Glucose_Level
 0.0099422    3 Dietary_Habits
 0.0098338   20 Diastolic_BP
 0.0097591    7 Marital_Status
 0.0094259   27 Headache
 0.0092256   23 Seizures
 0.0058731   30 Loss of Balance
```

Principal Components Attribute Selection

```
Ranked attributes:
 0.9394    1 -0.705Metabolic_Equivalent_of_Task_Score+0.329Stress_Levels+0.299Systolic_BP+0.294Hypert
 0.9035    2 0.399Seizures-0.383Difficulty Speaking-0.365Family_History_of_Stroke=No-0.325Patient_Gen
 0.868     3 0.486Numbness-0.317Average_Glucose_Level+0.283Stroke_History-0.244Difficulty Speaking+0.
 0.833     4 0.461Weakness-0.362Work_Type_of_patient+0.341Physical_Activity-0.301Seizures-0.256Hypert
 0.7981    5 0.558Headache+0.324Systolic_BP-0.291Seizures-0.245Blurred Vision-0.2Body_Mass_Index...
 0.7635    6 0.517Severe Fatigue-0.444Loss of Balance-0.388Dizziness+0.279Body_Mass_Index+0.219Blurre
 0.7291    7 0.509Marital_Status-0.405Heart_Disease+0.342Diastolic_BP+0.327Loss of Balance-0.295Avera
 0.695     8 0.401Confusion-0.341Blurred Vision-0.304LDL_Cholesterol+0.259Residence_Type=Urban-0.247N
 0.6613    9 0.465Confusion+0.313Physical_Activity+0.286Dietary_Habits+0.266Stroke_History+0.232Diast
 0.6278   10 0.457Weakness+0.339Difficulty Speaking-0.328Dizziness-0.296Family_History_of_Stroke=No-0
 0.5945   11 0.393Residence_Type=Urban-0.317Blurred Vision-0.305Severe Fatigue-0.299Patient_Gender=Fe
 0.5614   12 -0.395Work_Type_of_patient-0.372Patient_Age-0.334Residence_Type=Urban-0.331HDL_Cholester
 0.5284   13 -0.537Alcohol_Intake+0.393Blurred Vision+0.313Systolic_BP+0.287Residence_Type=Urban+0.25
```

Chosen Cut-Off Value: **0.8** (80% variance captured)

Remaining Attributes: First 4 Principal Component Vectors used as attributes

ReliefF Attribute Selection

Chosen Cut-Off Value: **0.001**

Remaining Attributes:

Residence_Type, Body_Mass_Index, Work_Type_of_patient, Loss of Balance, Average_Glucose_Level

```
Ranked attributes:
 0.00155019   10 Average_Glucose_Level
 0.00135279   30 Loss of Balance
 0.00134046    5 Work_Type_of_patient
 0.00130641   12 Body_Mass_Index
 0.00107257   17 Residence_Type
 0.00092763   21 Stroke_History
 0.000831     25 Weakness
 0.00078498    9 Stress_Levels
 0.0007292     1 Patient_Age
 0.00069572    2 Patient_Gender
 0.00066834   18 Systolic_BP
 0.00066154   13 Alcohol_Intake
 0.00052662    7 Marital_Status
 0.00040584   29 Severe Fatigue
 0.00035752   26 Confusion
 0.00029085   20 Diastolic_BP
 0.00018359   27 Headache
 0.00000271   14 HDL_Cholesterol
-0.00052179   24 Difficulty Speaking
-0.00054584    6 Metabolic_Equivalent_of_Task_Score
```

Classification Models

RandomForest

Classifier that builds multiple decision trees during training and outputs the class that is predicted by majority of the individual trees.

J48

Decision tree algorithm that classifies data by identifying the most informative data from the training set. It can handle missing data and prunes the trees to avoid overfitting, ensuring the model remains applicable to new data.

Classification Models cont.

NaiveBayes

A probabilistic classifier based on Bayes' theorem with the assumption of independence between features. Naive Bayes is particularly suited for large datasets and datasets where the assumption holds up reasonably well.

Decision Table

Compiles data into a table format, similar to a simplified rule-based symptoms, and makes predictions based on matching cases. Model is very simply to understand, just follow the table.

Train/Test/Validation Split

- Performed after Attribute selection.
- Created 5 different datasets
    - Split into 70%/15%/15%
- 7244 train instances, 1553 test instances, 1552 validation instances

```python
import sklearn
from google.colab import files
uploaded = files.upload()

import pandas as pd
from sklearn.model_selection import train_test_split

attribute = "ReliefF"

df = pd.read_csv(f'{attribute}.csv')

train_set, temp_set = train_test_split(df, test_size=0.3, random_state=42)
val_set, test_set = train_test_split(temp_set, test_size=0.5, random_state=42)

print("Training set size:", len(train_set))
print("Validation set size:", len(val_set))
print("Test set size:", len(test_set))

train_set.to_csv(f'{attribute}train.csv', index=False)
val_set.to_csv(f'{attribute}val.csv', index=False)
test_set.to_csv(f'{attribute}test.csv', index=False)

files.download(f'{attribute}train.csv')
files.download(f'{attribute}val.csv')
files.download(f'{attribute}test.csv')
```

Results

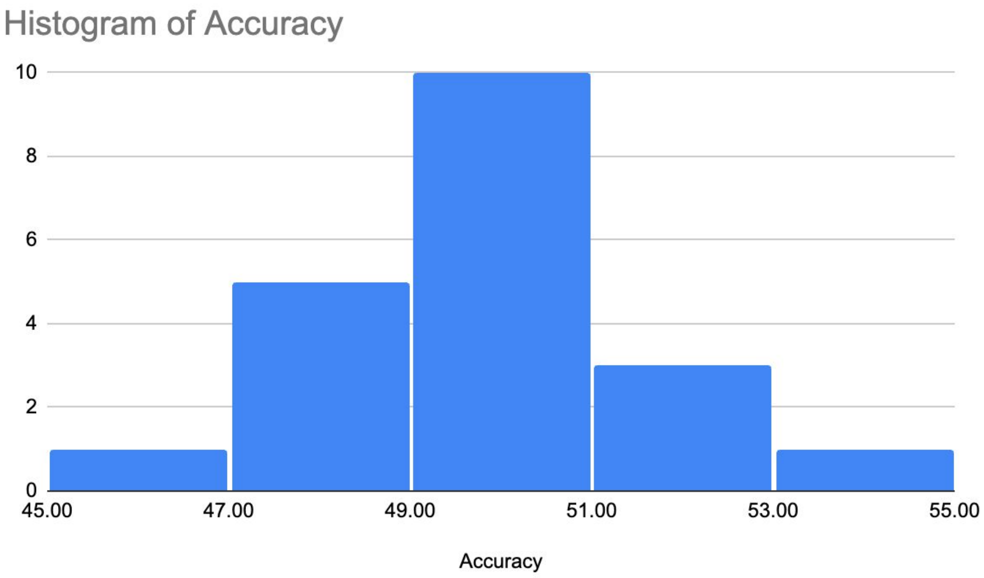
1. CorrelationAttributeEval with RandomForest – 53.51%, 0.535, 0.461, 0.522

2. CorrelationAttributeEval with J48 – 51.45%, 0.514, 0.480, 0.515

3. CorrelationAttributeEval with NaiveBayes – 51.26%, 0.513, 0.474, 0.530

4. Personal Selection with NaiveBayes – 51.10%, 0.511, 0.478, 0.508

5. Personal Selection with RandomForest – 50.87%, 0.509, 0.488, 0.506



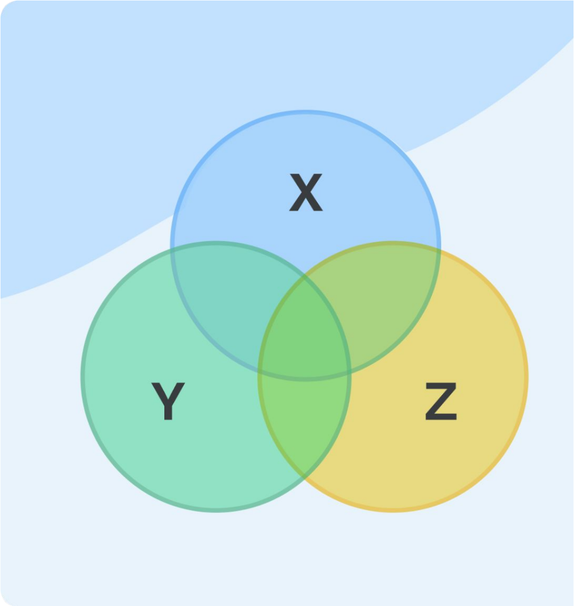
Results Continued

Poor results due to complexity of problem

1. Subjective attributes are self-reported
    - Stress Level, Physical Activity

2. Symptoms attributes created overlapping patterns

3. Multicollinearity – attributes are correlated in different combinations.
    - Overfitting vs Losing valuable data



Conclusion and Future Work

- CorrelationAttributeEval with RandomForest classifier most successful
- Datasets void of largely subjective data
- Find methods to deal with multicollinearity