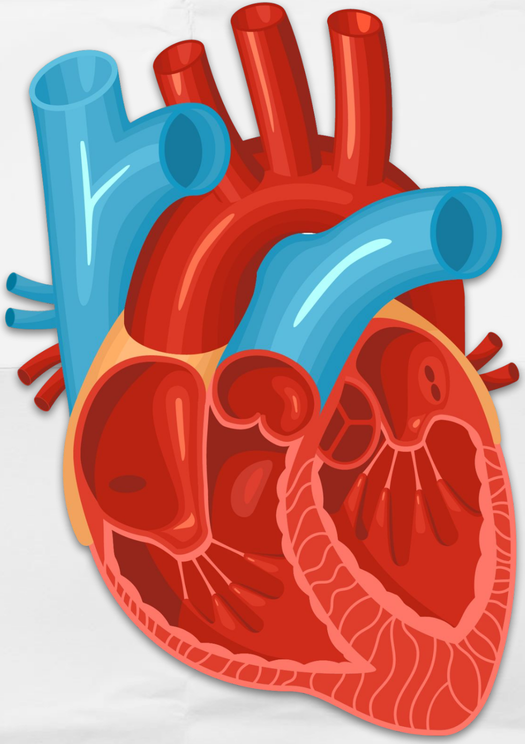
Stroke Prediction

Devanshi, Judy, Meghal, Zohair, Christina

Used - Flask API, JS, Matplotlib, Pandas, Machine Learning



INTRODUCTION

Do you know the risk of stroke doubles every 10 yrs after age 55? Working in high stress environment, bad eating and smoking habits can increase your stroke risk exponentially. In this project we utilize the Stroke Prediction Dataset, which contains various categories of information, including age, gender, marital status, smoking habits, bmi, heart disease, and average glucose levels to create a Stroke Prediction Machine!

Workflow

1

Data Cleaning and EDA

Pandas
Drop Null Values & Duplicate Values
Performed Exploratory Data Analysis to gain insights of the distribution of the dataset and explore potential relationships between stroke outcome and features

2

Data Preprocessing

Resample unbalanced dataset:
1) Oversampling with SMOTE (oversampled minority class)
2) RandomOverSampler
3) Get dummies
4) StandardScaler

3

Machine Learning Models

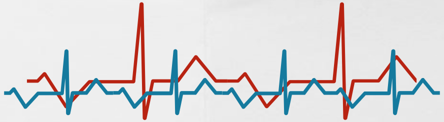
Built and trained machine learning models:

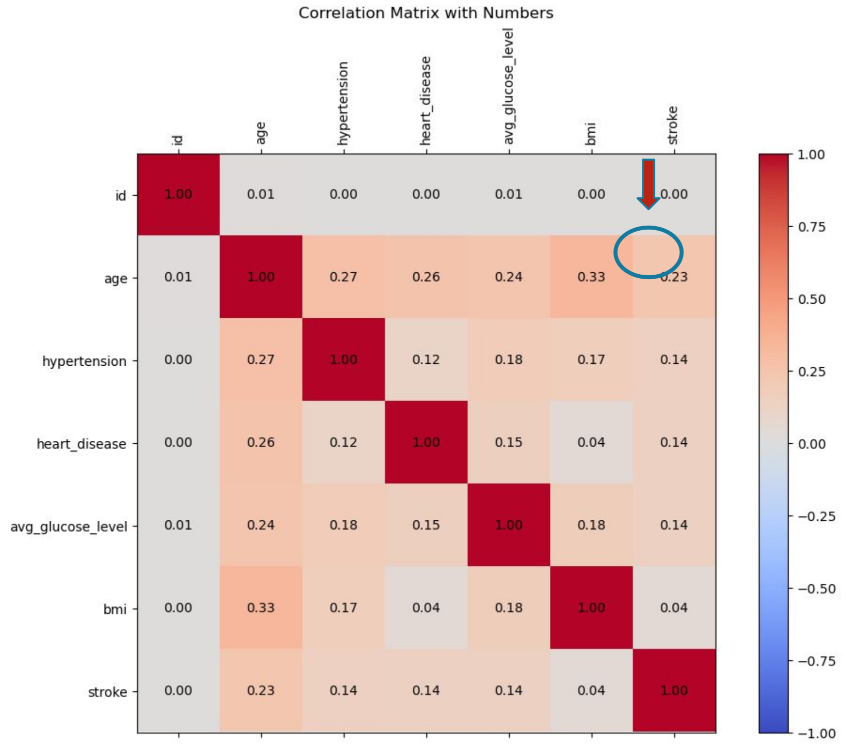
- Decision Tree
- Random Forest
- Deep Learning

4

Prediction via Flask

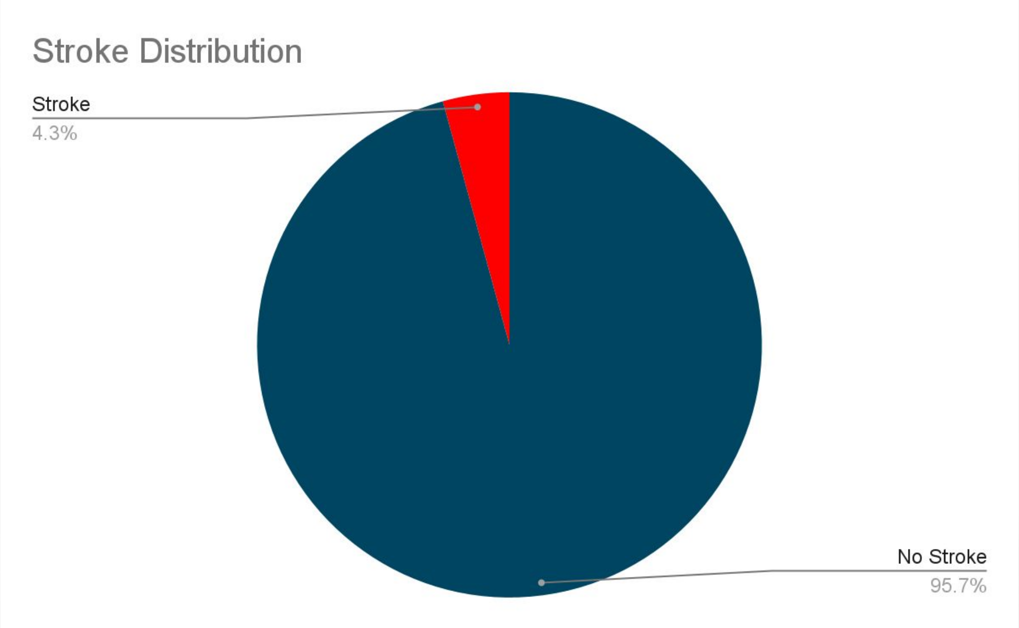
Interactive flask app showcasing the findings and predicts the outcome of stroke by entering in your data





Correlation Matrix with Numbers

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
id	1.00	0.01	0.00	0.00	0.01	0.00	0.00
age	0.01	1.00	0.27	0.26	0.24	0.33	0.23
hypertension	0.00	0.27	1.00	0.12	0.18	0.17	0.14
heart_disease	0.00	0.26	0.12	1.00	0.15	0.04	0.14
avg_glucose_level	0.01	0.24	0.18	0.15	1.00	0.18	0.14
bmi	0.00	0.33	0.17	0.04	0.18	1.00	0.04
stroke	0.00	0.23	0.14	0.14	0.14	0.04	1.00



Correlation Test

- Age has the highest correlation at 0.23
- Hypertension, heart disease, and average glucose level are all relatively the same at 0.14
- BMI has the lowest correlation at 0.04

Stroke Distribution

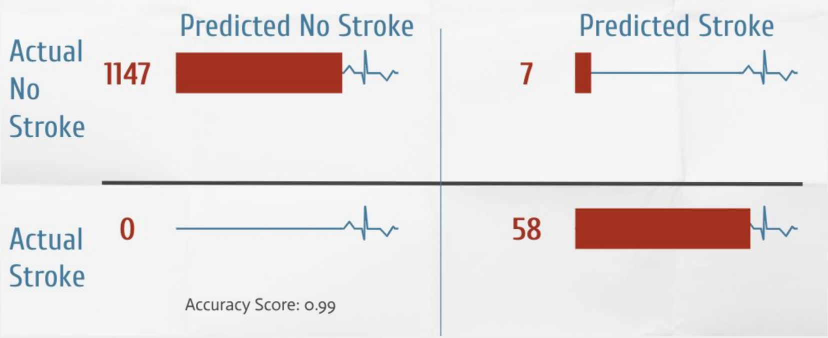
- 4700 - had no strokes
- 209 - had strokes
- 12 columns of data

Why Random Forest and Decision Tree?

Stroke Prediction is a Classification Problem !

- SMOTE
  - Synthetic Minority Oversampling Technique
  - Oversampling minority class for balancing distribution
  - Imbalanced data
- Features Importances:
  - 1) Age = 0.36
  - 2) Avg Glucose Level = 0.20
  - 3) BMI = 0.16

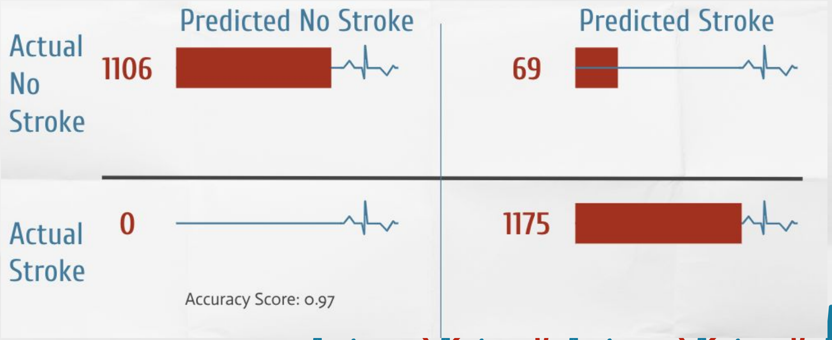
Random Forest



	Predicted No Stroke	Predicted Stroke
Actual No Stroke	1147	7
Actual Stroke	0	58

Accuracy Score: 0.99

Decision Tree



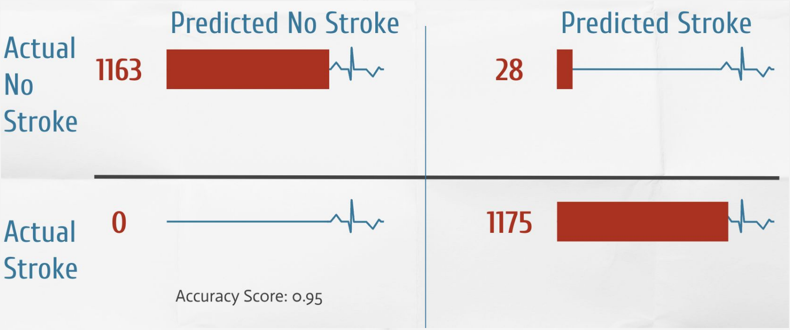
	Predicted No Stroke	Predicted Stroke
Actual No Stroke	1106	69
Actual Stroke	0	1175

Accuracy Score: 0.97

Deep Learning Model

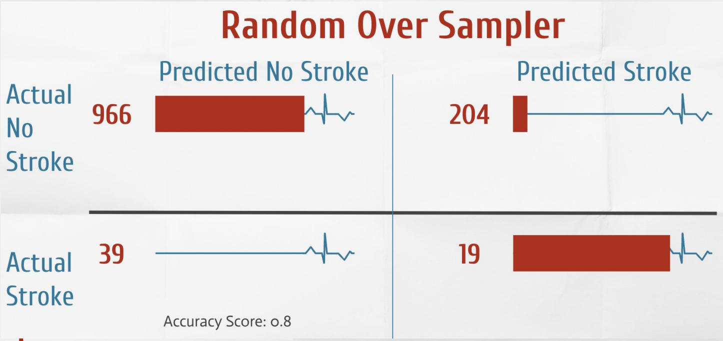
1) **Model**- epochs:200
   Input layer: Relu, 21 neurons (input dimension - 21)
   First hidden layer: Relu, 21 neurons
   Second hidden layer: Sigmoid, 1 neuron
2) Round off predicted output between 0-1 to find accuracy
3) Random Over Sampling caused overfitting
4) Observations:

- More hidden layers can identify complex, nonlinear relationships.
- More neurons provides the network with a greater capacity to learn complex patterns and speeds up the process.
- Tanh has a smooth gradient compared to ReLU, it makes it easier for the optimization algorithm (e.g., gradient descent) to find the global minimum of the loss function, leading to faster convergence during training even with fewer epochs.



	Predicted No Stroke	Predicted Stroke
Actual No Stroke	1163	28
Actual Stroke	0	1175

Accuracy Score: 0.95

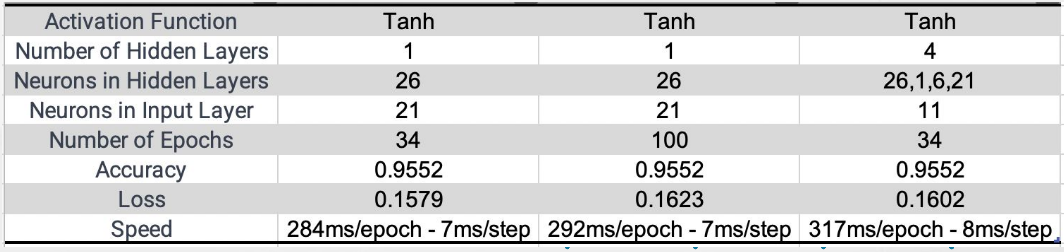


Random Over Sampler

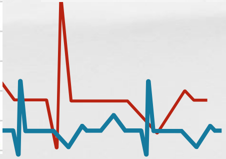
	Predicted No Stroke	Predicted Stroke
Actual No Stroke	966	204
Actual Stroke	39	19

Accuracy Score: 0.8

Hyperparameter Tuning



Activation Function	Tanh	Tanh	Tanh
Number of Hidden Layers	1	1	4
Neurons in Hidden Layers	26	26	26,1,6,21
Neurons in Input Layer	21	21	11
Number of Epochs	34	100	34
Accuracy	0.9552	0.9552	0.9552
Loss	0.1579	0.1623	0.1602
Speed	284ms/epoch - 7ms/step	292ms/epoch - 7ms/step	317ms/epoch - 8ms/step



FLASK PRESENTATION

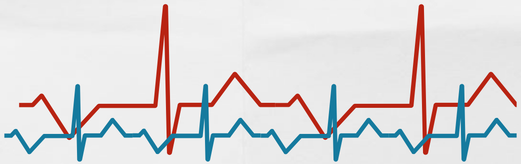
CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon** and infographics & images by **Freepik**

CONCLUSIONS

- The Random Forest Model had the highest accuracy out of all models at **99%**
- Different optimizers for balancing data produced different outcomes on the efficacy of the machine learning models
- **Age** had the **highest correlation** and **bmi** had the **lowest correlation** with stroke. Contrarily, the features importance test for the random forest model ranked bmi as the third most important feature in predicting stroke.
- ML Model Specific Strengths and Limitations for stroke prediction dataset:
- Random Forest:
  - Strength: Combines multiple decision trees to improve prediction accuracy. Reduces overfitting. Works well on imbalanced data.
  - Weakness: Requires more computational resource and has less interpretability
- Decision Trees:
  - Strength: Works well on small datasets, has high interpretability and simplicity
  - Weakness: Does not work well on imbalanced data, it tends to favour majority class
- Deep Learning Model:
  - Strength: Captures complex patterns
  - Weakness: Does not work well with relatively small dataset and class imbalance. Computationally expensive to train and tune

Limitations

- Limited data availability
  - Our stroke data was limited, having only 209 people from 4909 people that actually had a stroke
  - Data was imbalanced
- Overfitting
  - Because of the lack of information, we had to overfit the data into the machine learning models which results in less reliability
- Recommendations:
  - Include weights of each category like bmi, age, smoking habits etc in stroke prediction, to spread awareness of developing better lifestyle habits in early stages to reduce the chances of stroke



THANKS

DO YOU HAVE ANY QUESTIONS?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**