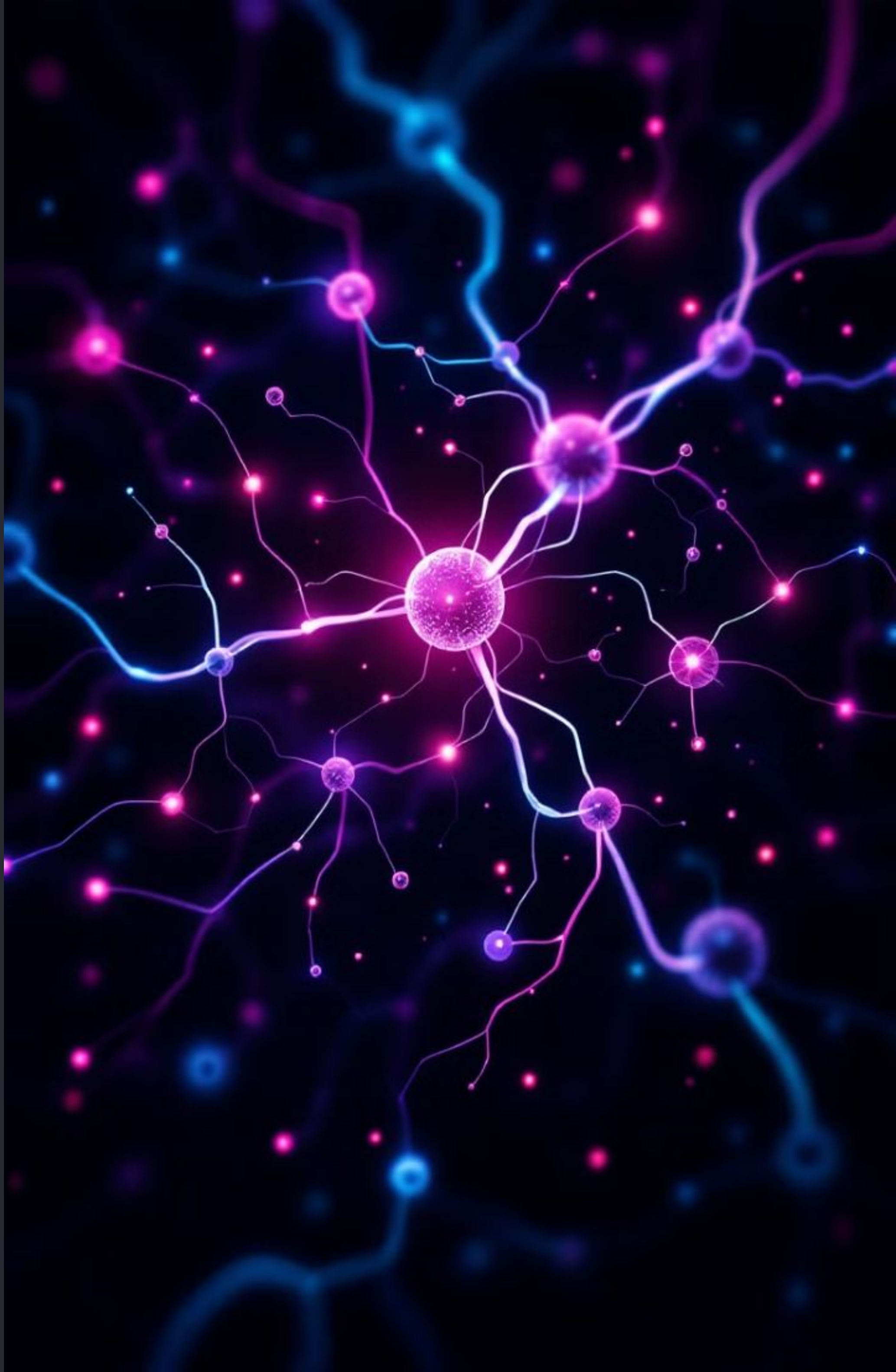
The Strands of Health

Explore how machine learning can uncover insights in health data, forecast outcomes, and guide interventions.

By: Chenwei Pan



INTRODUCTION

Objectives

- Investigate different factors that may contribute to stroke and diabetes
- Develop a Machine Learning-Based Prediction Model
- Create a Streamlit Application for Risk Assessment

Problem

- Stroke risk increases with age, but can occur at any age. Early intervention is crucial.
- Around 830 million people worldwide have diabetes, with over half not receiving treatment [1,2].
- Hypertension may be a primary factor contributing to stroke, diabetes, and other chronic diseases. [3] In Canada, 22.6% of adults have hypertension, with over 70% over 80 [4].
- On a personal level, I have a long family history of hypertension, so I developed a machine-learning model to understand if hypertension increases stroke and diabetes risks. The model analyzes patient data to predict these risks, contributing to a broader understanding of hypertension's impact on overall health.

BACKGROUND INFO.

Stroke: A Race Against Time

1

Occurrence

Stroke happens when blood flow changes in the brain, depriving cells of oxygen. [5]

2

Impact

Brain cells die without oxygen, leading to potential difficulties in speaking, thinking, or walking. [5]

3

Risk Factors

Age increases risk, but strokes can occur at any age. Early action is crucial. [1]

Types of Stroke

- Ischemic strokes make up about 87% of cases, occur when blood flow to the brain is blocked, either by a clot that forms in the brain (thrombotic stroke) or travels from elsewhere in the body (embolic stroke).
- Hemorrhagic strokes account for 13% of cases, and result from bleeding in or around the brain. There are two subtypes: intracerebral hemorrhage and subarachnoid hemorrhage, each with different causes such as high blood pressure or aneurysms. Hemorrhagic strokes require immediate medical intervention to manage bleeding and reduce risks. [6]

What is Diabetes?

Diabetes is a health problem that affects many people. When you have diabetes, your body has trouble using the sugar (glucose) in your blood. This sugar can't get into your cells to give you energy, so you may feel tired.

Insulin is a hormone that helps your body use sugar. When you have diabetes, your body doesn't make enough insulin or can't use it properly. Without insulin, the sugar stays in your blood instead of going into your cells. [7]

Main Types of Diabetes

Type 1 Diabetes – An autoimmune condition where the body produces little or no insulin because the immune system attacks insulin-producing cells. It is usually diagnosed in children or young adults and requires lifelong insulin therapy.

Type 2 Diabetes – The most common type, where the body doesn’t use insulin properly or doesn’t produce enough. It is often linked to obesity, lifestyle factors, and family history. Can be prevented or delayed with a healthy lifestyle.

Gestational Diabetes – Develops during pregnancy and usually disappears after birth. However, it increases the risk of developing type 2 diabetes later in life.

Others - Prediabetes, monogenic (gene mutation), pancreas damage, etc.

[7]

Diabetes: A Global Health Challenge

Global Prevalence

About 830 million people worldwide have diabetes. [2]

Treatment Gap

More than half of people with diabetes are not receiving treatment. [2]

Health Risks

Diabetes raises the risk of eye, kidney, nerve, and heart damage. [7]

Hypertension: A Silent Threat

22.6%

Adult Prevalence

Hypertension affects 22.6% of Canadian adults.

20%

Prehypertension

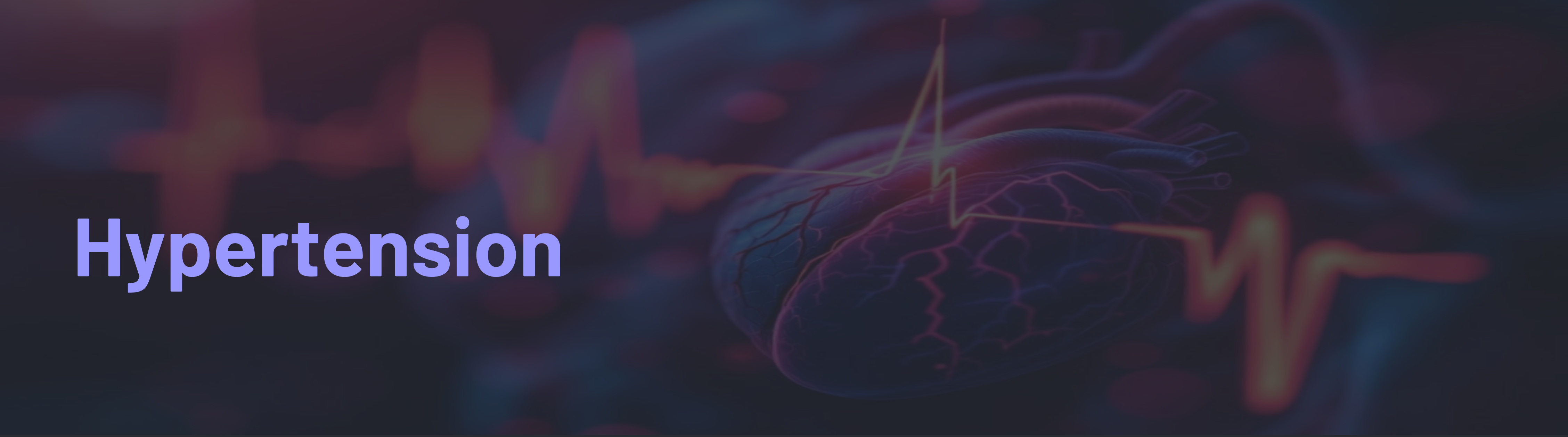
An additional 20% have prehypertension.

90%

Lifetime Risk

90% of Canadians may develop hypertension in their lifetime.

[4]



Hypertension

Definition

High blood pressure straining blood vessels. [9]

Types

Primary: A gradual onset, it develops slowly over time

Secondary: Caused by other underlying medical conditions. Tends to be more sudden and severe. [3]

Risks

Heart disease, stroke, kidney damage. Often symptomless, severe cases may have headaches or nosebleeds. [3]

Machine Learning in Health Prediction

1

1. Data Collection

Gather health datasets from public sources (CDC).

2

2. Model Development

Train machine learning models for prediction (code).

3

3. Performance Assessment

Evaluate model accuracy and reliability. (code)

4

4. Application

Create streamlit app for real-life scenarios.

Common Types of Models/Algorithms Used In Machine Learning

Regression

Predicts a continuous quantity. Uses existing models and labeled data for training (supervised learning).

Predictive Models

Predictive models uses big data analytics and deep learning to examine historical data, patterns, and trends. The more data provided to the machine learning algorithms, the better the predictions.

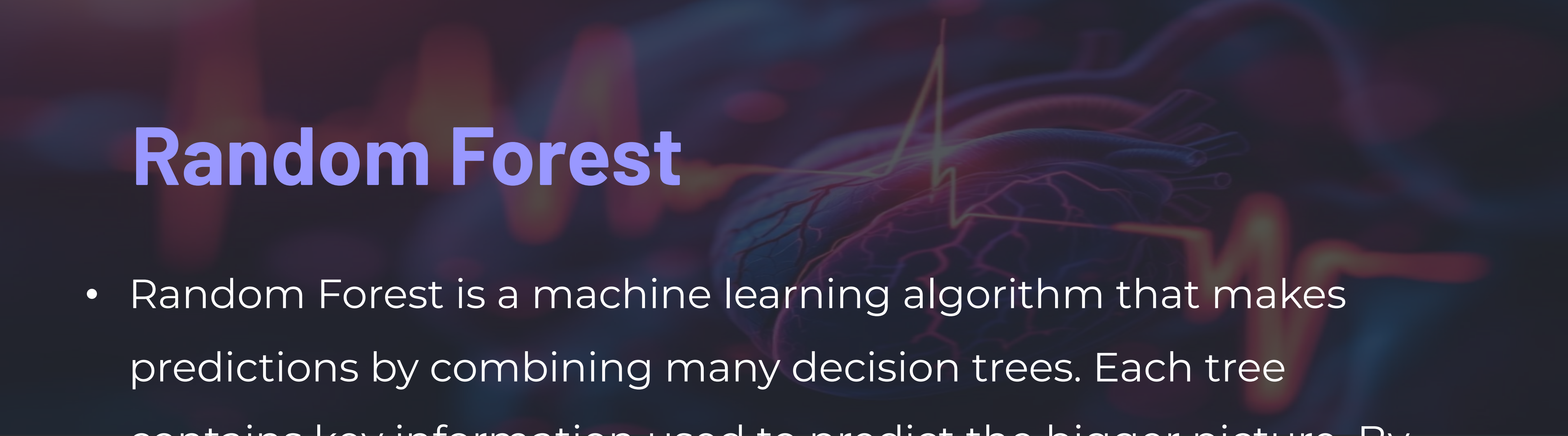
Classification

Predicts discrete class labels. Uses existing models and labeled data for training (supervised learning). A simpler example is logistic regression, which predicts accuracy based on a linear relationship.

[11]

Logistic Regression

- Logistic regression is a statistical method designed to predict binary outcomes, such as yes/no or 0/1.
- Unlike linear regression, which predicts continuous values, logistic regression computes probabilities and classifies outcomes based on a predefined threshold (e.g., probabilities over 50% classified as "yes").
- Logistic regression is well-suited for classification tasks, like classifying emails
- Logistic regression works well with large datasets [12]



Random Forest

- Random Forest is a machine learning algorithm that makes predictions by combining many decision trees. Each tree contains key information used to predict the bigger picture. By combining the answers from all the trees, the Random Forest reduces errors and improves accuracy.
- This works well for tasks like predicting if someone might have a disease or what product someone might buy.
- It’s powerful because it handles large datasets and avoids overfitting. [13]

MY PROJECT

My Project

- I gathered datasets from CDC containing health information, including hypertension, blood sugar, and other relevant factors, from publicly available sources. Using this data, I developed and trained a machine-learning model to predict the likelihood of stroke and diabetes with high accuracy.
- There are two datasets, one I used to predict stroke and one I used to predict diabetes.

My Dataset Source

- The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based health survey on CDC designed to collect data on health-related risk behaviors, chronic conditions, and preventive service use among U.S. adults.
- It uses telephone surveys, including both landline and cell phones, to gather information from a representative sample of adults.
- The BRFSS consists of core questions, optional modules, and state-specific questions, covering topics like health status, access to care, chronic diseases, and health behaviors. Its primary purpose is to monitor public health trends, guide policy development, and support research to improve health outcomes.

What My Code Contains:

1. Import all necessary libraries (Numpy, pandas, etc.)
2. Upload and Prepare Data
3. Process data
4. Find missing values and replace with median
5. Turn everything into numerical data
6. Split data into training and testing
7. Create and train the model pipeline (smote, classifier, preprocessor)
8. Perform grid search for the best pseudo-parameters
9. Cross-validate and print results
10. Generate predictions and evaluate
11. Generate confusion matrix (true positive, true negative, etc)
12. Plot feature importance
13. Calculate and print metrics (accuracy, precision, etc)

Information Collected in My Datasets

Stroke

- Sex
- Age
- Hypertension
- Heart disease
- Ever married
- Work type
- Residence type
- Average glucose level
- BMI
- Smoking status
- Stroke (yes or no)

Diabetes

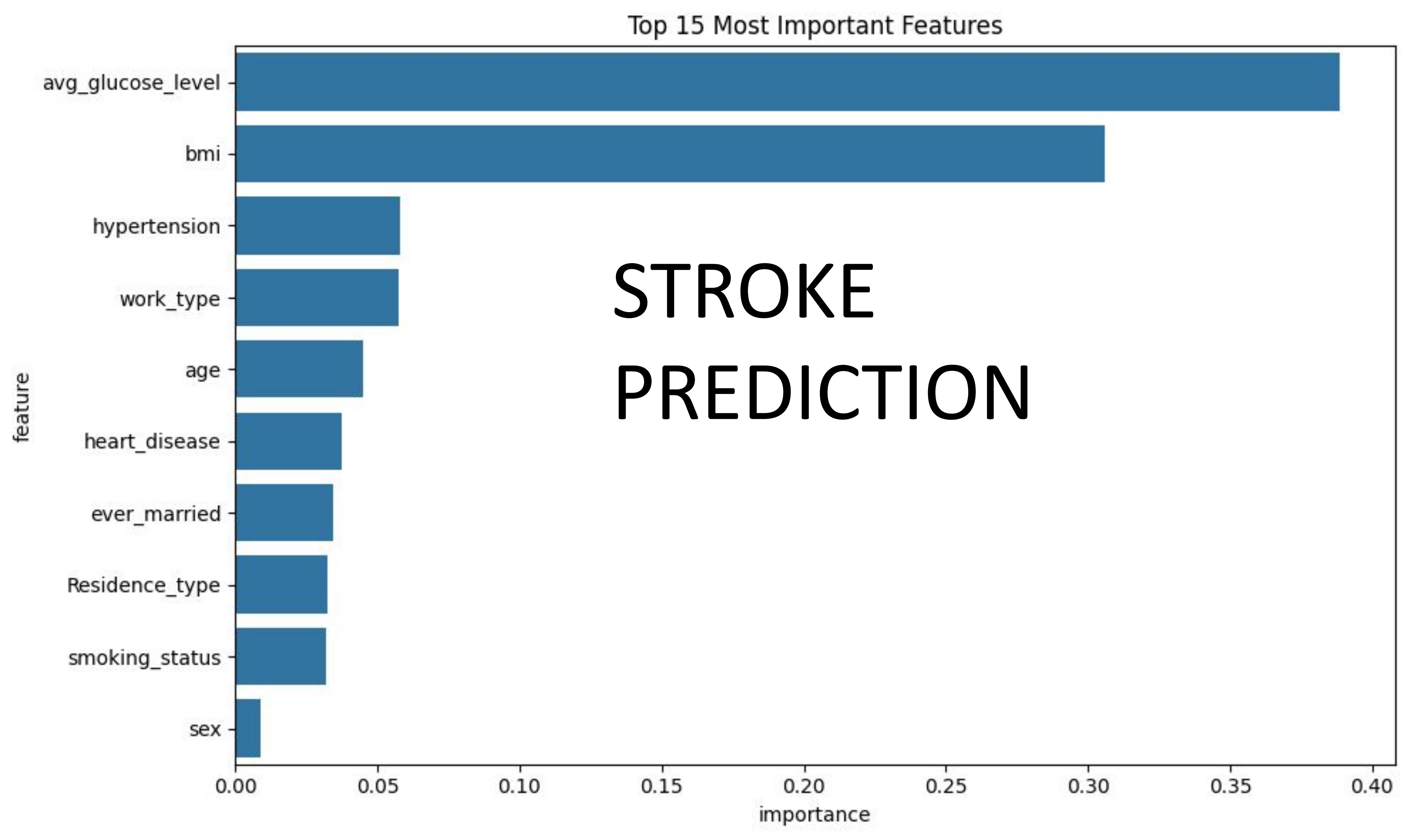
- Age
- Sex
- High cholesterol
- Cholesterol check
- BMI
- Smoker
- Heart disease or attack
- Physical activity
- Fruits
- Vegetables
- Heavy alcohol consumption
- General health
- Mental health
- Physical health
- Difficulty walking
- Stroke
- High blood pressure
- Diabetes (yes or no)

RESULTS & ANALYSIS

TABLE 1

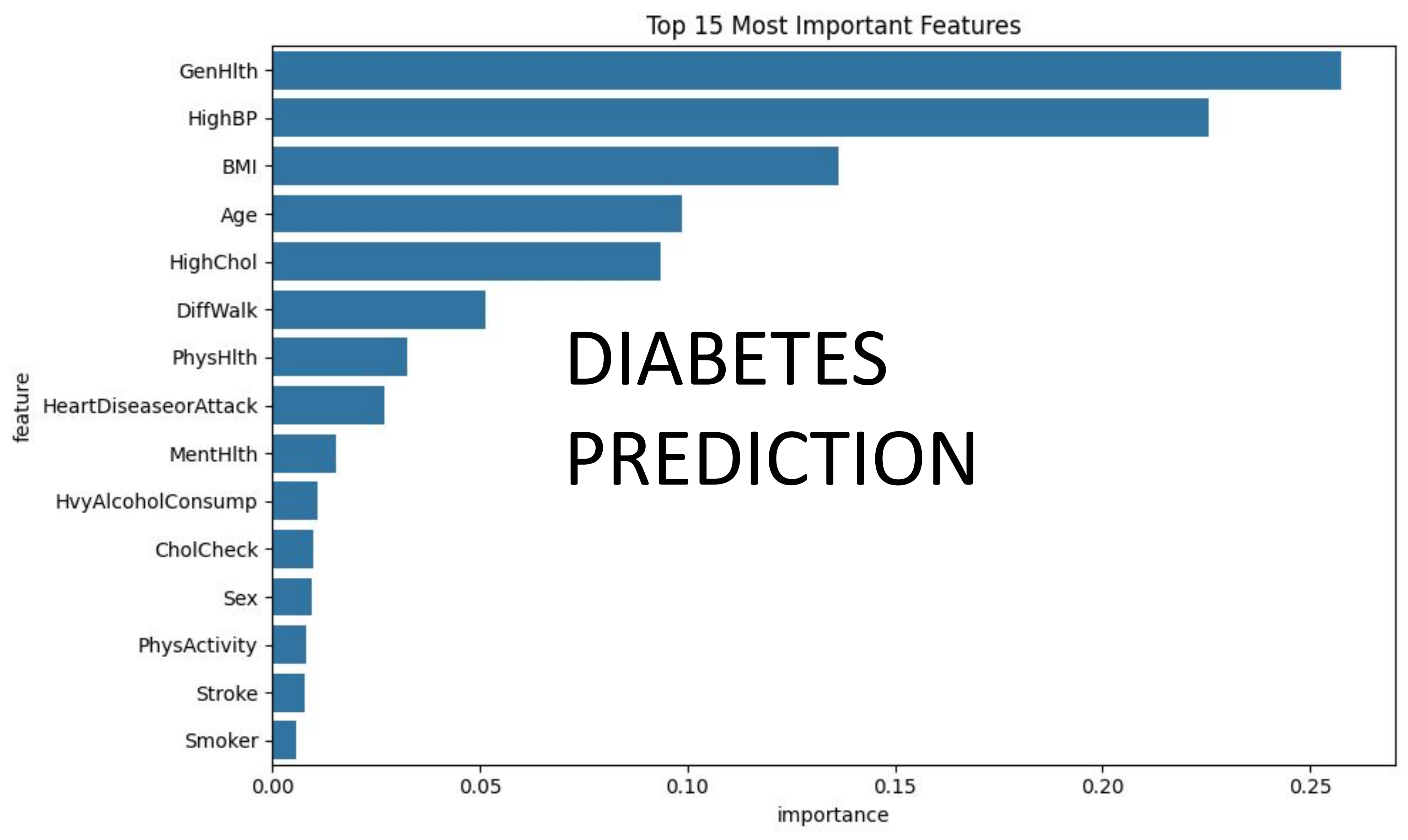
Dataset	Accuracy with Logistic Regression	Accuracy With Random Forest
Stroke	68%	99%
Diabetes	74%	75%

FIGURE 1



STROKE PREDICTION

FIGURE 2



DIABETES PREDICTION

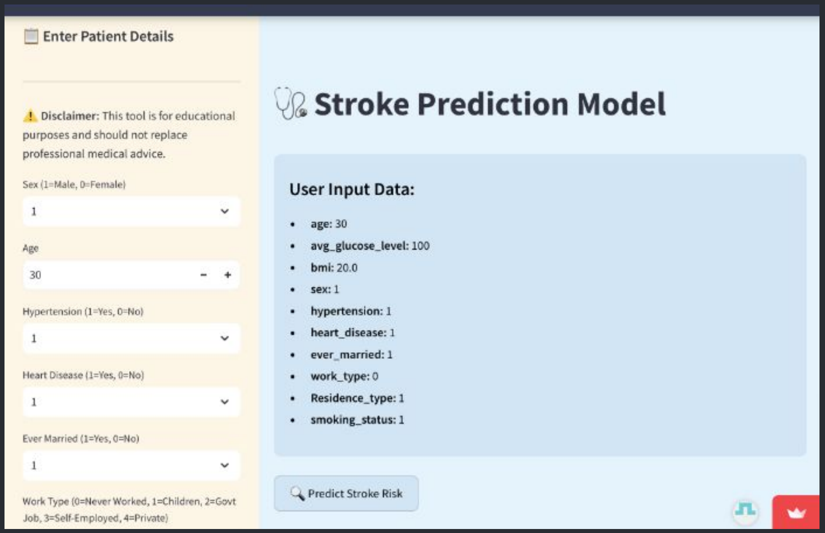
Analysis

- The study found that glucose level and BMI are crucial factors for stroke prediction, with high blood sugar levels (hyperglycemia) potentially increasing stroke risks.
- General health and BMI also impact the likelihood of developing chronic illnesses.
- Hypertension is the third most important factor for stroke and the second most important factor for diabetes, potentially due to medication use.
- The study suggests that machine learning can guide preventive healthcare, highlighting the importance of awareness and lifestyle changes in reducing stroke and diabetes risks.

MY APP

Streamlit

- Streamlit is an open-source Python library that allows you to quickly build interactive web applications.
- I used streamlit to help build my stroke prediction app.



📋 Enter Patient Details

⚠️ **Disclaimer:** This tool is for educational purposes and should not replace professional medical advice.

Sex (1=Male, 0=Female): 1

Age: 30

Hypertension (1=Yes, 0=No): 1

Heart Disease (1=Yes, 0=No): 1

Ever Married (1=Yes, 0=No): 1

Work Type (0=Never Worked, 1=Children, 2=Govt Job, 3=Self-Employed, 4=Private)

🩺 Stroke Prediction Model

User Input Data:

- **age:** 30
- **avg_glucose_level:** 100
- **bmi:** 20.0
- **sex:** 1
- **hypertension:** 1
- **heart_disease:** 1
- **ever_married:** 1
- **work_type:** 0
- **Residence_type:** 1
- **smoking_status:** 1

Parts of My App

My app is a stroke prediction model, because my stroke model was the most accurate.

- It has a section for inputting patient details
- It has a section to double-check your information
- It has a health tips section
- It has a link for more information
- It has a user reviews section

Impact

- This project explores the predictive power of machine learning in healthcare, highlighting its potential for early detection and intervention.
- The app could be used for patient assessment surveys, guiding individuals in assessing stroke risk. It could save time and money by reducing unnecessary doctor visits.
- However, the app should not replace professional medical advice. Similar technologies have already been developed, but this project aims to make them more user-friendly and practical for early health assessments.

Conclusion

- The machine-learning model highlights the importance of maintaining a healthy lifestyle to prevent serious health conditions. Hypertension can lead to serious medical implications.
- I created a Streamlit app using a machine-learning model to identify key risk factors for stroke and diabetes.
- The random forest algorithm performed better than logistic regression, highlighting AI's potential in medical prediction and prevention.

Limitations

- I emailed a multitude of other health organizations (AHS, CIHI, communityhealth, infostats, etc.) and hospitals but the data I needed either required an ethics approved study (which I applied for but no response came) or they didn’t have that data.
- There was limited data available, and I could only find public source data from the U.S.. I hope more people will use my app so that I will be able to collect more local data to improve my model.

Improvements

While my model provides valuable insights, there are several ways it could be improved:

- Expanding the dataset to include a wider range of demographics and medical histories
- Refining features to include additional health metrics like sleep patterns
- Optimizing the model by testing other machine-learning algorithms
- Improving the user experience by enhancing the Streamlit app's interface and adding explanations for predictions
- Collaborating with healthcare professionals to validate predictions and explore practical applications.

Next Step

Objective: Enhance hypertension prediction by analyzing retinal images and improving segmentation techniques.

Key Steps:

- Find Paper: https://github.com/nakib103/Hypertensive-Retinopathy-Detection/tree/master
- Use DRIVE or similar datasets, focusing on retinal fundus images.
- Enhancing hypertension detection by refining retinal image segmentation, automating AVR measurement, and improving machine learning models based on prior research.

Goal: Improve diagnostic accuracy of hypertensive retinopathy detection using advanced image preprocessing.

Thank You

I appreciate your time and interest in my project.

I especially thank my mentors from Juniotech who guided me through the machine learning process, and Ms. Lai who provided me with unconditional support. Without them, this project wouldn't be possible.

I am especially grateful for Dr. Leyla baghirzada, clinical assistant professor at the University of Calgary, who guided me through this challenging yet rewarding process.

I am excited to share my research and contribute to the development of health prediction models.

References

[1] CDC. (2024, October 24). *Stroke Facts*. Stroke. https://www.cdc.gov/stroke/data-research/facts-stats/index.html

[2] World. (2019, May 13). *Diabetes*. Who.int; World Health Organization: WHO. https://www.who.int/health-topics/diabetes#:~:text=About%20830%20million%20people%20worldwide,diabetes%20are%20not%20receiving%20treatment

[3] *High blood pressure (hypertension): Controlling this common health problem-High blood pressure (hypertension) - Symptoms & causes - Mayo Clinic*. (2024). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410

[4] *HYPERTENSION IN CANADA HIGH BLOOD PRESSURE (HYPERTENSION) IS THE LEADING RISK FOR DEATH AND DISABILITY WORLDWIDE*. (2016). https://hypertension.ca/wp-content/uploads/2018/12/HTN-Fact-Sheet-2016_FINAL.pdf.

[5] https://www.facebook.com/NIHAging. (2023, February 9). *Stroke: Signs, Causes, and Treatment*. National Institute on Aging. https://www.nia.nih.gov/health/stroke/stroke-signs-causes-and-treatment#:~:text=A%20stroke%20happens%20when%20there%27s,oxygen%20suffer%20and%20eventually%20die

[6] *Types of Stroke*. (2022, December 13). Hopkinsmedicine.org. https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/types-of-stroke.

[7] and, D. (2025, January 23). *What Is Diabetes?* National Institute of Diabetes and Digestive and Kidney Diseases; NIDDK - National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[8] *High blood pressure (hypertension): Controlling this common health problem-High blood pressure (hypertension) - Symptoms & causes - Mayo Clinic*. (2024). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410

[9] *hypertension*. (2025). Vocabulary.com. https://www.vocabulary.com/dictionary/hypertension#:~:text=Hyper%2D%20is%20a%20prefix%20that,strain%20on%20your%20blood%20vessels.

[10] IBM. (2024, August 12). *Predictive AI*. Ibm.com. https://www.ibm.com/think/topics/predictive-ai#:~:text=Predictive%20AI%20uses%20big%20data,biases%20in%20predictive%20AI%20models

[11] Keita, Z. (2022, September 21). *Classification in Machine Learning: An Introduction*. Datacamp.com; DataCamp. https://www.datacamp.com/blog/classification-machine-learning

[12] Dawson, C. (2021, February 11). *A Guide to Logistic Regression for Beginners - Christa Dawson - Medium*. Medium. https://dawsonc96.medium.com/a-guide-to-logistic-regression-for-beginners-c53632fea4e4

[13] *What is Random Forest? [Beginner's Guide + Examples]*. (2020, October 21). CareerFoundry. https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/

[14] *Behavioral Risk Factor Surveillance System*. (2024, November 22). Cdc.gov. https://www.cdc.gov/brfss/index.html

[15] SWASTIK. (2020, October 6). *SMOTE for Imbalanced Classification with Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

[16] *Hyperglycemia: Symptoms, Causes, and Treatments*. (2023, November). Yale Medicine.
https://www.yalemedicine.org/conditions/hyperglycemia-symptoms-causes-treatments#:~:text=Hyperglycemia%20is%20a%20condition
o%20develop%20in%20non%2Ddiabetics