# Science Fair Logbook

By Chenwei Pan
Westmount Charter Mid High School

## Project Timeline

### September 2024

- **September 10-15:** Initial project idea
    - I finalized my project topic: Can the climate of different regions affect your risk of high blood pressure/hypertension?
    - Motivation: I was originally motivated to do this because studies have shown that hypertension may be caused by the weather and temperature. I also have a long family history of hypertension, so I wanted to see if there was any correlation between the climate and your chances of hypertension.
    - https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/effects-of-high-temperatures-on-blood-pressure-heart
- **September 16-20:** Researched the impact of hypertension/what it is
    - Problem: In Canada, the prevalence of hypertension in adults is 22.6% and an additional 20% have prehypertension. More than 70% of adults above the age of 80 have hypertension as well. Over 90% of Canadians are estimated to develop hypertension if they live an average life span. [1]
    - Background info: Another name for high blood pressure is hypertension. Hyper- is a prefix that means "over" or "beyond" — if you're hyper you're wildly energetic. Tension means "stretching" or "straining." Hypertension, therefore, means "straining beyond." With hypertension, your blood pressure is abnormally high, causing a strain on your blood vessels. [2]
    - Hypertension is usually symptomless, but severe cases may lead to headaches, shortness of breath, or nosebleeds. It is commonly linked to

factors like diet, lack of physical activity, genetics, stress, or underlying health issues.

- Types:
  - Primary hypertension: Gradual onset, no clear cause.
  - Secondary hypertension: Results from other conditions or medications
- Having hypertension can cause heart disease, stroke, kidney damage, and more. Lifestyle changes (diet, exercise) and medications are often recommended for control. [3]
- Sources:
  [1]https://hypertension.ca/wp-content/uploads/2018/12/HTN-Fact-Sheet-2016_FINAL.pdf
  [2]https://www.vocabulary.com/dictionary/hypertension#:~:text=Hyper-%20is%20a%20prefix%20that,strain%20on%20your%20blood%20vessels
  [3] https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410

- **September 21-30:** A literature review on climate effects on blood pressure
  - MayoClinic states that “Blood pressure generally is **higher in the winter and lower in the summer**. That's because low temperatures cause blood vessels to temporarily narrow. More pressure is needed to force blood through narrowed veins and arteries. This causes blood pressure to rise.” [4]
  - I will try to see if there is a correlation between the climate and your chance of getting hypertension, and create an app that will tell you your risk of hypertension based on the average temperature of your city and your age. This won't completely stop your chances of developing hypertension, but it is still good to know beforehand. If you are already at risk of hypertension (perhaps due to diet, exercise, lifestyle, etc.), this data

may help you find the best place to live. If you already live in a place with a high risk of hypertension, you will also be able to know quickly and change your lifestyle quickly.
    - Source: [4]https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/blood-pressure/faq-20058250#:~:text=Blood

## October 2024

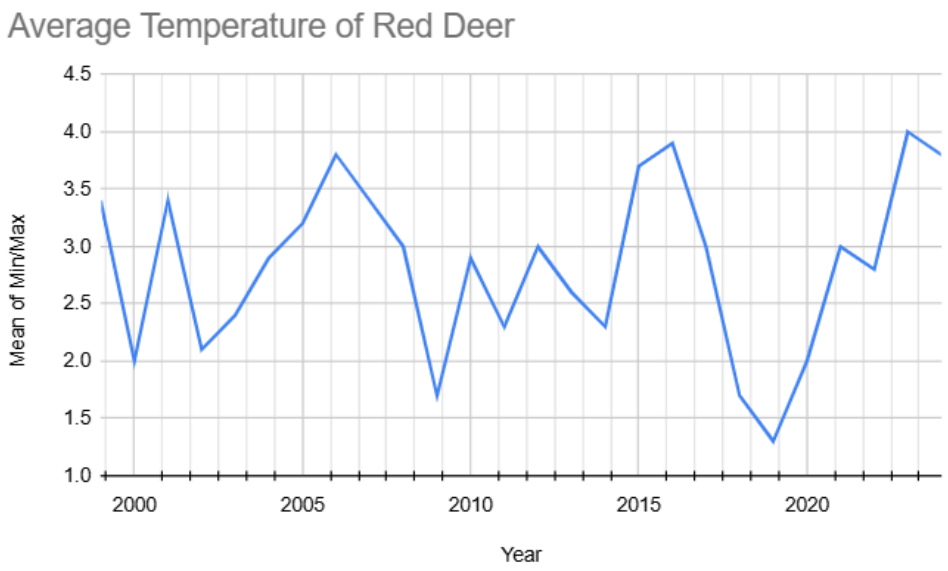
- **October 1-10:** Planning and identification of data sources
    - Hypertension Alberta: http://www.ahw.gov.ab.ca/IHDA_Retrieval/selectSubCategoryParameters.do
    - The datasets for other provinces are less available to the public.
- **October 11-20:** Data collection from climatedata.ca for Alberta cities (Calgary, Edmonton, High Level)
    - I chose these cities because they represent the North (High Level), Middle (Edmonton), and South (Calgary.
- **October 21-31:** Weather trend analysis and graphing for selected cities
    - I compiled the datasets for the weather into a Google spreadsheet.

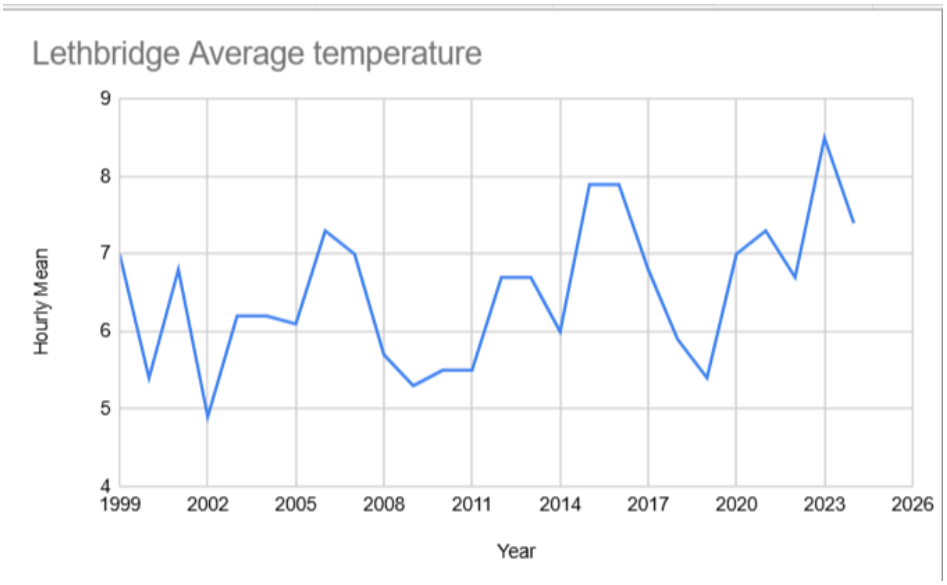
## November 2024

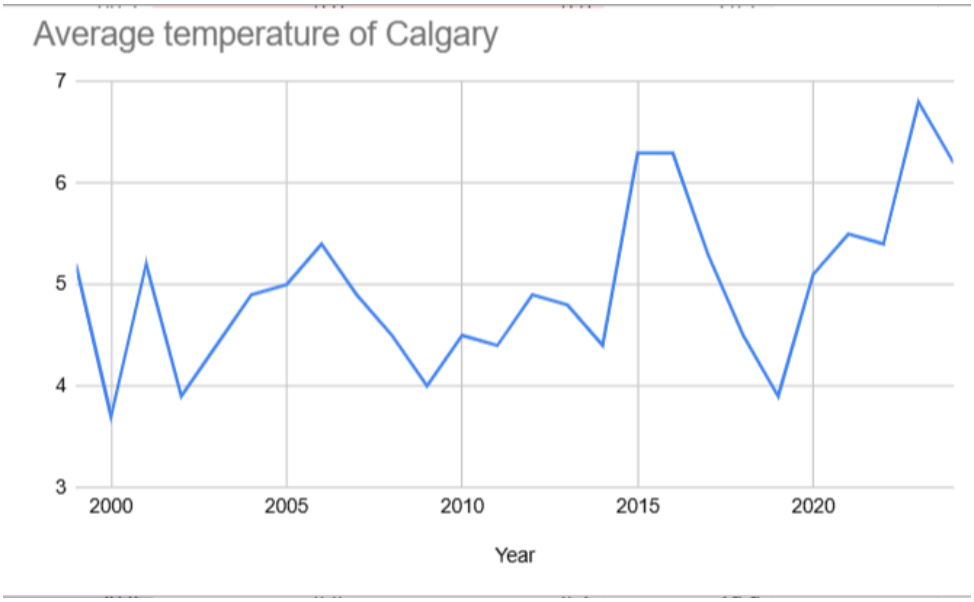
- **November 1-15:** Collection of hypertension statistics for Alberta
    - The surgeon general of the USA is trying to get more people to pay attention to this issue.
        - https://www.hhs.gov/surgeongeneral/reports-and-publications/disease-prevention-wellness/index.html
- **November 16-30:** Provincial hypertension data compilation and analysis
    - I also compiled the hypertension dataset for Alberta into a spreadsheet by gender and age

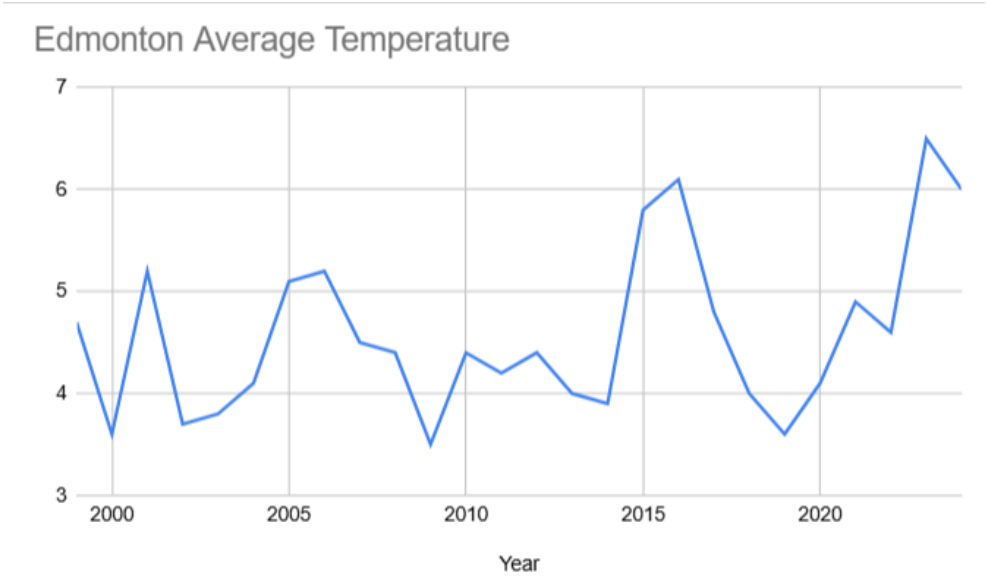
## December 2024

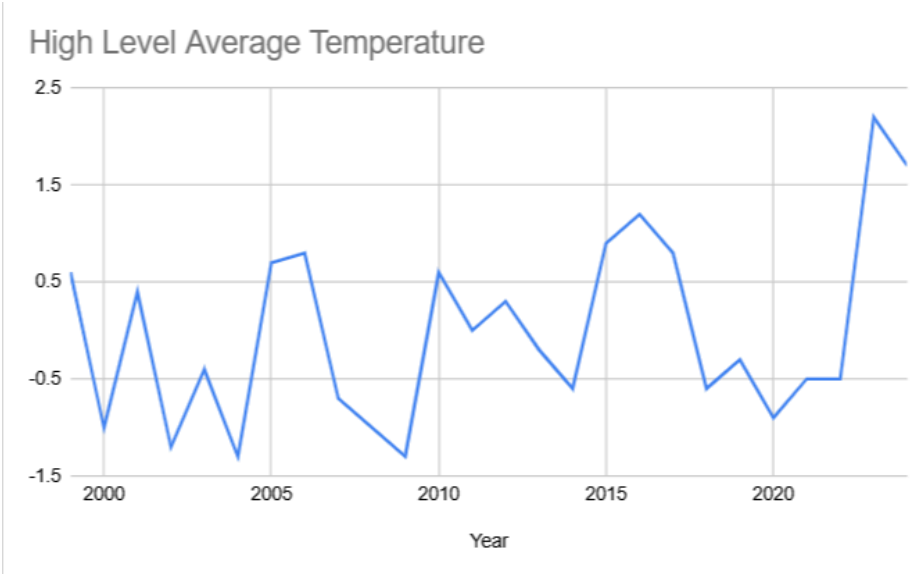
- **December 1-7:** Data correlation analysis between climate trends and hypertension rates
    - Putting these two types of data together to see if there is a trend.
    - Trying to find better data, since it would be nice to know the monthly weather as well as provincial data for ALL Canadian provinces/territories.
- **December 7-14:**
    - I found a better climate dataset since the climate data one was not recent enough:
        - **https://calgary.weatherstats.ca/charts/temperature-yearly.html**
        - https://edmonton.weatherstats.ca/charts/temperature-yearly.html
        - https://highlevel.weatherstats.ca/charts/temperature-yearly.html
- **December 14-21**
    - I have now found a great resource for the different zones/regions of Alberta.
    - These regions are the South Zone, Calgary Zone, Central Zone, Edmonton Zone, and North Zone. To make things easier, I chose one city/town from each zone to find datasets on. For South, I have Lethbridge, Calgary is Calgary, Central is Red Deer, Edmonton is Edmonton, and North is High Level. I also found different statistics on the different zones from the Alberta Health Services Website, including obesity rates which directly impact hypertension.
    - Here are some charts I made of the temperature in each over the years,

- 
Average Temperature of Red Deer
Mean of Min/Max
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
2000
2005
2010
2015
2020
Year

- 
Lethbridge Average temperature
Hourly Mean
9
8
7
6
5
4
1999
2002
2005
2008
2011
2014
2017
2020
2023
2026
Year


Average temperature of Calgary
7
6
5
4
3
2000
2005
2010
2015
2020
Year

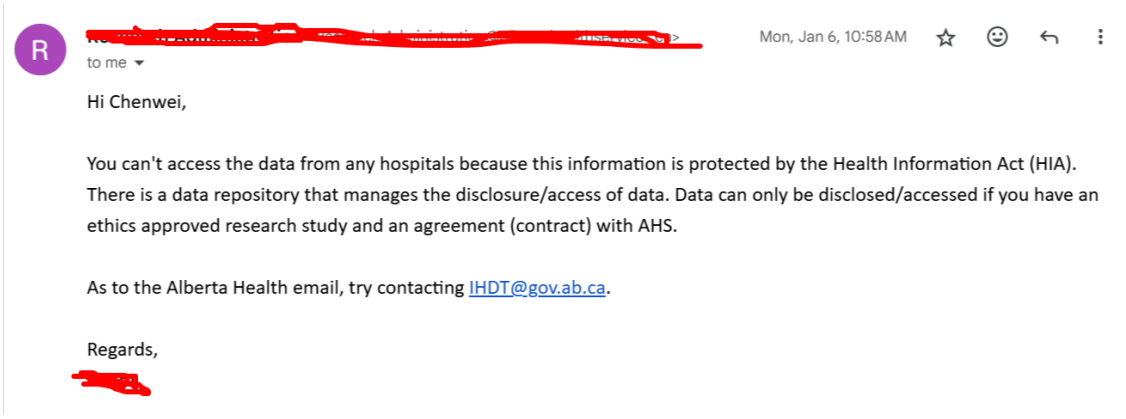
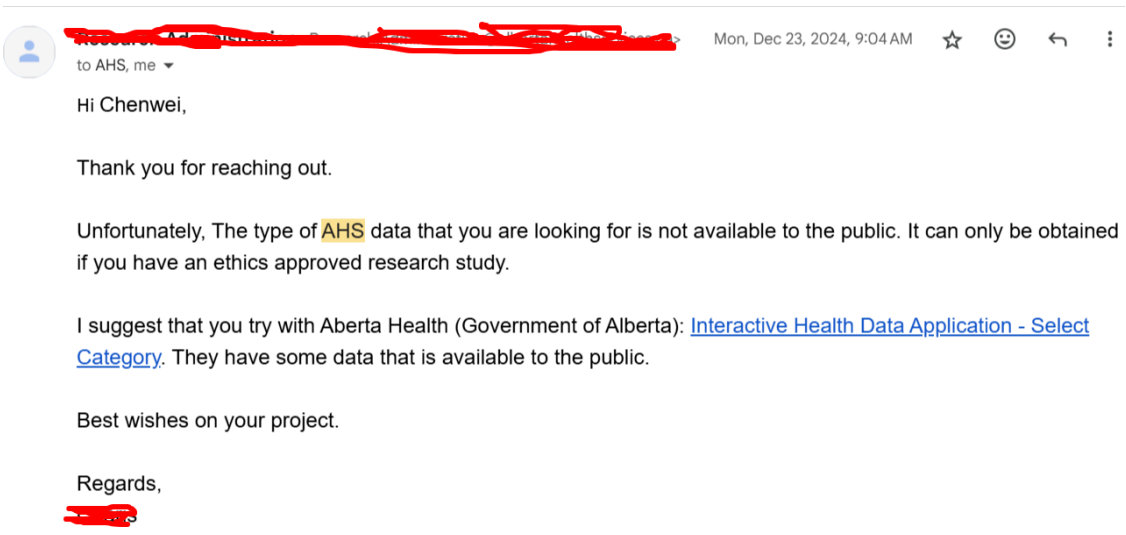

Edmonton Average Temperature
7
6
5
4
3
2000
2005
2010
2015
2020
Year

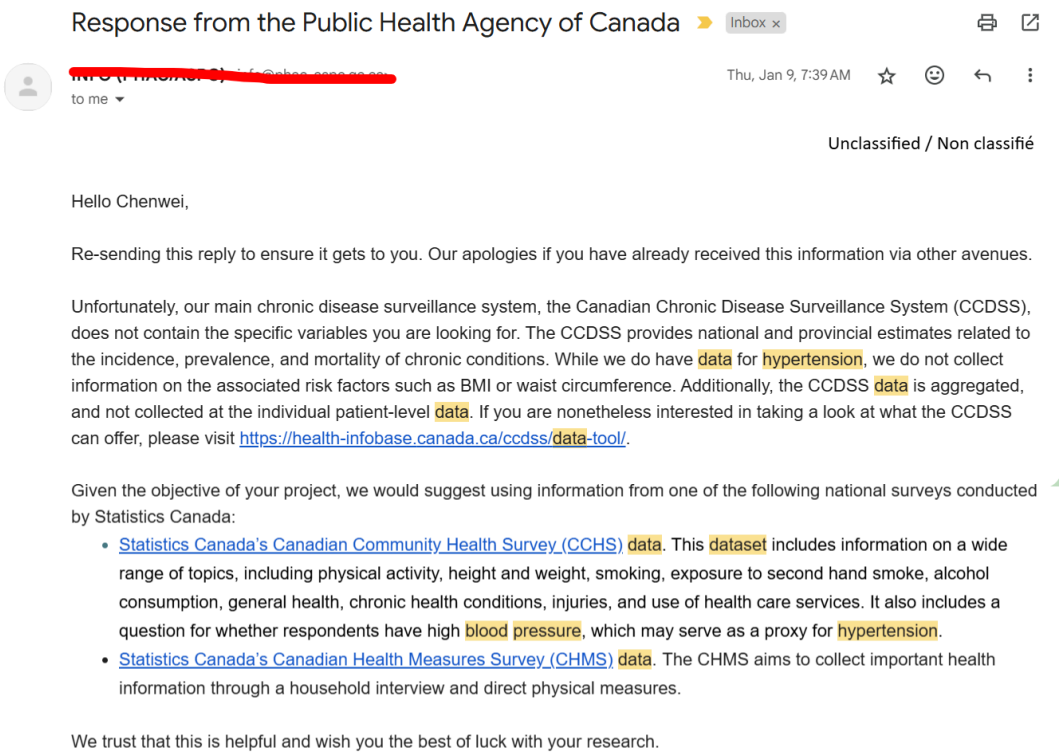



- I edited my hypertension dataset to be able to choose the sex, year, age group, and zone to see the different rates of hypertension for each one.
- I can't currently access these graphs, because my computer broke! (spoiler alert!) - Feb. 5
- Decided to expand my project to Social Economic and Environmental Impacts on Hypertension.
- I can compile datasets on average housing prices or inflation or income for economics, I can compile populations for social, or environmental I can continue with my current work.
- Found all of the weather data for Lethbridge and Red Deer as well, so now I have a full weather dataset for each zone.

- **December 22:** Decision to pivot the project focus
  - Concluded that the lack of sufficient data for the original project made it infeasible to continue.
  - I asked Dr. Leyla Baghirzada about this, and she agreed that this project would be very difficult to do since there are no relevant datasets.
  - Brainstormed alternative project ideas, including focusing on a machine-learning approach to analyze medical data.
  - I found a paper about ways to predict hypertension using easy details in patient information. These details included Body Mass Index (BMI), age,

family history, and waist circumference (WC), and then whether they have hypertension or not.

- https://pmc.ncbi.nlm.nih.gov/articles/PMC8497705/#:~:text=In%20addition%2C%20Body%20Mass%20Index,without%20clinical%20or%20genetic%20data
- I could not find this information online, so I emailed a multitude of health organizations (AHS, CIHI, communityhealth, infostats, etc.). However, 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.
- 
[redacted]
Mon, Jan 6, 10:58 AM
to me
Hi Chenwei,
You can't access the data from any hospitals because this information is protected by the Health Information Act (HIA). There is a data repository that manages the disclosure/access of data. Data can only be disclosed/accessed if you have an ethics approved research study and an agreement (contract) with AHS.
As to the Alberta Health email, try contacting IHDT@gov.ab.ca.
Regards,
[redacted]
- 
[redacted]
Mon, Dec 23, 2024, 9:04 AM
to AHS, me
Hi Chenwei,
Thank you for reaching out.
Unfortunately, The type of AHS data that you are looking for is not available to the public. It can only be obtained if you have an ethics approved research study.
I suggest that you try with Aberta Health (Government of Alberta): Interactive Health Data Application - Select Category. They have some data that is available to the public.
Best wishes on your project.
Regards,
[redacted]



Response from the Public Health Agency of Canada

Inbox ×

[redacted]

Thu, Jan 9, 7:39 AM

to me

Unclassified / Non classifié

Hello Chenwei,

Re-sending this reply to ensure it gets to you. Our apologies if you have already received this information via other avenues.

Unfortunately, our main chronic disease surveillance system, the Canadian Chronic Disease Surveillance System (CCDSS), does not contain the specific variables you are looking for. The CCDSS provides national and provincial estimates related to the incidence, prevalence, and mortality of chronic conditions. While we do have data for hypertension, we do not collect information on the associated risk factors such as BMI or waist circumference. Additionally, the CCDSS data is aggregated, and not collected at the individual patient-level data. If you are nonetheless interested in taking a look at what the CCDSS can offer, please visit https://health-infobase.canada.ca/ccdss/data-tool/.

Given the objective of your project, we would suggest using information from one of the following national surveys conducted by Statistics Canada:

- Statistics Canada's Canadian Community Health Survey (CCHS) data. This dataset includes information on a wide range of topics, including physical activity, height and weight, smoking, exposure to second hand smoke, alcohol consumption, general health, chronic health conditions, injuries, and use of health care services. It also includes a question for whether respondents have high blood pressure, which may serve as a proxy for hypertension.
- Statistics Canada's Canadian Health Measures Survey (CHMS) data. The CHMS aims to collect important health information through a household interview and direct physical measures.

We trust that this is helpful and wish you the best of luck with your research.

- 
    - However, private datasets are not necessarily the best option to use. Public datasets are most commonly used for a reason - they are ethics-approved. Public datasets are safe to use, they are free to access, widely available, and used by professionals daily!

## January 2025

- **January 1-5:** Finalized the new project topic
    - Decided to shift focus to using machine learning to analyze factors related to stroke and diabetes.
    - Goal: Use machine learning to identify patterns and correlations that could help predict or understand stroke and diabetes risk.
- **January 6-10:** Researched machine learning methods and data requirements
    - Explored common machine learning models and tools suitable for medical data analysis.
    - Machine Learning:
      Machine learning is the use and development of computer systems that

can learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. (Oxford Dictionary)

- Predictive Models:

  Predictive AI 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. (IBM)
  There are different types of predictive models and different algorithms. The two main types of algorithms that may benefit my project are regression and classification. Both regression and classification are types of supervised machine learning algorithms, where a model is trained according to the existing model along with correctly labeled data. The most significant difference between regression and classification is that while regression helps predict a continuous quantity, classification predicts discrete class labels. One simpler classification model is logistic regression, (ironic, I know), which is what I used for my first prototype. This model predicts accuracy by comparing it to a linear relation. (datacamp)

  Logistic regression is a statistical method used to predict a binary outcome, like yes/no or 0/1. Unlike linear regression, which predicts continuous values, logistic regression calculates probabilities and classifies outcomes based on them (e.g., over 50% = yes). Logistic regression is ideal for classification problems, like determining if an email is spam, a customer will buy a product, or a student will pass an exam. It works best with large datasets and independent, meaningful variables. (medium)

  Another popular algorithm is called random forest. This is what I'm using on my current and most accurate model. Random Forest is a machine learning algorithm that makes predictions by combining many decision

trees. Imagine each tree is like an expert analyzing data to answer. By combining the answers from all the trees, the Random Forest reduces errors and improves accuracy. It 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, which happens when a model gets too specific and struggles with new data. (career foundry)

- Investigated publicly available datasets that included stroke-related medical data.
    - The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based health survey by CDC designed to collect data on health-related behaviours, 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 behaviours. Its primary purpose is to monitor public health trends, guide policy development, and support research to improve health outcomes.
- **January 11-15:** Began preliminary work on a new project
    - Identified and downloaded datasets related to stroke risk factors.
        - Downloaded from the above dataset
    - Started cleaning and preprocessing the data for use in machine learning models.
        - In code
- **January 16-24:** Experimented with using random forest and logistic regression
    - Random forest and logistic regression are two of the models I mentioned before. I used them both to predict stroke and diabetes, and I want to see which one works better.
    - For stroke prediction, I got this result:

        - Logistic Regression - 68%, Random Forest - 99%
        - This may be a sign of overfitting, but I did cross-validate and the result remains the same
    - For diabetes prediction, I got this result:
        - Logistic Regression - 74%, Random Forest - 75%
        - This shows that random forest works better than logistic regression in my case.
    - Work on a presentation for the school science fair
        - The top 15 projects will be chosen for the city fair.
    - Did background information for the medical section of my project:

      Factors affecting your health:

      Hypertension:

      Another name for high blood pressure is hypertension. Hyper- is a prefix that means "over" or "beyond" — if you're hyper you're wildly energetic. Tension means "stretching" or "straining." Hypertension, therefore, means "straining beyond." With hypertension, your blood pressure is abnormally high, causing a strain on your blood vessels. (Vocabulary)

      Hypertension is usually symptomless, but severe cases may cause headaches, shortness of breath, or nosebleeds. It is commonly linked to factors such as diet, lack of physical activity, genetics, stress, or underlying health issues.

    - Types:
        - Primary hypertension: Gradual onset, no clear cause.
        - Secondary hypertension: Results from other conditions or medications
    - Having hypertension can cause heart disease, stroke, kidney damage, and more. Lifestyle changes (diet, exercise) and medications are often recommended for control. (MayoClinic)

Hyperglycemia:

Hyperglycemia is a condition in which you have higher amounts of glucose in your blood than normal. It is sometimes called “high blood sugar”.

Transient hyperglycemia typically doesn't result in long-term issues. However, if high blood sugar levels continue, it can result in serious complications such as vision problems, kidney damage, nerve damage, cardiovascular disease, and stroke.

Non-diabetic people who have hyperglycemia are at high risk of developing type-2 diabetes. [13]

Even mildly raised blood sugar levels can put you more at risk of a stroke. [14]

Stroke:
A stroke occurs when blood flow changes the brain. Blood brings oxygen and nutrients to brain cells. If blood can't flow to a part of the brain, cells that do not receive enough oxygen suffer and eventually die. Once brain cells die, they can’t be repaired. As a result, someone who has had a stroke may have trouble speaking, thinking, or walking. (nia)

Stroke is the third highest leading cause of death globally, right after COVID-19. (WHO)

Types of Stroke:
Ischemic: Makeup up about 87% of cases, occurs 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: Accounting for 13% of cases, results 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. [15]

Diabetes:
Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Glucose is your body's main source of energy. Your body can make glucose, but glucose also comes from the food you eat.

Insulin is a hormone made by the pancreas that helps glucose get into your cells to be used for energy. If you have diabetes, your body doesn't make enough—or any—insulin or doesn't use insulin properly. Glucose then stays in your blood and doesn't reach your cells.

Diabetes raises the risk of damage to the eyes, kidneys, nerves, and heart. Diabetes is also linked to some types of cancer. Taking steps to prevent or manage diabetes may lower your risk of developing diabetes health problems. (niddk)

Stroke risk increases with age, but strokes can—and do—occur at any age. Early action is important for stroke. Strokes are a leading cause of serious long-term disability. (CDC)

About 830 million people worldwide have diabetes, the majority living in low-and middle-income countries. More than half of people living with diabetes are not receiving treatment. (WHO)

- Worked on the Streamlit app, and finished it!

        - Streamlit is an open-source Python library that allows you to quickly build interactive web applications.
- **January 25-31:**

    - Finished my presentation and started practicing presenting for the school fair
        - I presented in front of experts in coding and healthcare and got some valuable feedback.
            - Don’t be defensive!!
            - Put in more details on the medical-related aspect.
            - Make sure your problem is clear and concise, and each section of the presentation is relevant.
    - Finishing up on my CYSF platform
    - Final details for my app (added a diet and exercise tip section)
    - Probability over 30% = high risk
    - Added session state to streamlit app, so that my app wouldn’t refresh every time you click a button.
    - I changed the colour scheme of my app to make it more aesthetically pleasing.
    - My computer BROKE so I had to start over on my presentation. I was using PowerPoint without an account, so my progress wasn’t saved.
        - Use Google Slides instead!

## February 2025

- **February 1-7:** Week leading to the school science fair
    - Quickly caught up on the slideshow because my computer broke
    - Decorating trifold by printing out my Google slides
    - App is done!! All of my sections are there, and my new colour scheme looks a lot better
    - Logbook is up-to-date
    - CYSF platform components are (mostly) finished

        - I wrote my analysis and conclusion.
- **February 8-15:** Week after science fair (school)
    - Present presentation at school fair
    - Continue tweaking MINOR details in my presentation
    - Waiting for feedback and results
    - Starting a scientific report about my project
- **February 16-23:**
    - Got my results and feedback back!
    - I made it into the next round and will be going to cities!
    - Here is your feedback:

| Scores | Feedbacks |
|---|---|
| 86 | What might be some limitations of your project? Ensure you have tested your hypothesis a number of times, connect the findings with the problem |
| 91 | What improvements could be made for your project? |
| 92 | |
| 98 | |
| 94 | Good |

- Implementing feedback into my project
- Improvements:

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

- More Diverse Data: Expanding the dataset to include a wider range of demographics and medical histories could enhance accuracy.
- Feature Refinement: Incorporating additional health metrics, such as sleep patterns, may improve predictions.
- Model Optimization: Testing other machine-learning algorithms or fine-tuning hyperparameters could further boost performance.
- User Experience: Enhancing my Streamlit app's interface and adding explanations for predictions could make it more user-friendly.
- Real-World Validation: Collaborating with healthcare professionals to validate predictions and explore practical applications would strengthen credibility.

- **February 24-28:**
  - Continue working on slideshow
  - Starting phase 2 of my project, detecting hypertension with retina imaging
  - https://github.com/nakib103/Hypertensive-Retinopathy-Detection/tree/master
  - Extension of the above paper
  - I will try to remove the background of DRIVE datasets, or datasets containing images of retinas, to predict hypertension more accurately

## March 2025

- **March 1-8:**
  - Working on reading the paper and understanding the logistics of the code
    - This includes AVR calculation, segmentation, etc.
  - Starting on my code to mimic their segmentation and AVR calc code
  - Editing all of what I am submitting on the CYSF platform so everything flows and makes sense.
    - Method, Problem, Conclusion, etc.

- **March 9-16:**
    - Working on the CYSF platform because soon there will be no editing time
        - Making sure what I wrote makes sense
        - Wrapping up phase 1, and finishing up the scientific report about phase 1 (will do formatting later)
            - I am not planning on submitting the report yet, but if I finish it before the actual fair, then I will let the judges take a look at it.
        - Recording a presentation to upload online
        - Ensuring ALL required elements are there
        - Working on my scientific report
        - Below you will find my edited version of what I will be submitting to the platform.
            - I changed all of my in-text citations and citations so that they were in numerical order. That is why they might not be matching up with what I wrote previously.

Problem:

My project aims to investigate different health factors that may contribute to stroke and diabetes.

Stroke risk increases with age, but strokes can occur at any age. Early intervention is crucial, as strokes are a leading cause of serious long-term disability [1]. On the other hand, approximately 830 million people worldwide have diabetes, with the majority residing in low- and middle-income countries. More than half of those affected do not receive treatment [2].

Based on research, hypertension may be one of the primary factors contributing to stroke, diabetes and many other chronic diseases [3]. In Canada, the prevalence of hypertension in adults is 22.6%, with an additional 20% classified as prehypertensive. More than 70% of adults over 80 have hypertension. If the average lifespan is reached, over 90% of Canadians are estimated to develop hypertension [4].

On a personal level, I am deeply invested in this issue due to a long family history of hypertension. To better understand what might happen to my family, or myself, I aim to determine whether this condition significantly increases the risk of stroke and diabetes. To explore this, I developed a machine-learning model that analyzes patient data to predict the risks of stroke and diabetes. By identifying these risks, I hope to contribute to a broader understanding of hypertension's impact on overall health.

My project not only delves into the relationship between hypertension and stroke as well as diabetes but also identifies other key factors that may influence their development. The findings could provide insights into how a healthy lifestyle affects health outcomes, potentially encouraging individuals to adopt healthier habits.

Background info:

First of all, let's look at how stroke and diabetes occur.

A stroke occurs when blood flow to the brain is disrupted. Blood supplies oxygen and nutrients to brain cells, and if blood flow is blocked, these cells suffer damage and eventually die. Once brain cells die, they cannot be repaired, leading to difficulties with speech, cognition, and mobility [5].

There are two types of strokes, ischemic strokes and hemorrhagic strokes. Ischemic strokes account for approximately 87% of cases. They occur when blood flow to the brain is obstructed, either by a clot forming in the brain (thrombotic stroke) or a clot travelling from elsewhere in the body (embolic stroke). Hemorrhagic strokes, comprising 13% of cases, result from bleeding in or around the brain. The two subtypes are intracerebral hemorrhage and subarachnoid hemorrhage, caused by factors such as high blood pressure or aneurysms. Hemorrhagic strokes require urgent medical intervention to control bleeding and mitigate risks [6].

Diabetes occurs when blood glucose (sugar) levels are too high. It develops when the body does not produce enough insulin or cannot use it effectively, causing glucose to remain in the bloodstream rather than being used for energy. By definition, glucose, the body's primary energy source, comes from both internal production and dietary intake. Insulin, a hormone produced by the pancreas, facilitates glucose absorption into cells. Diabetes increases the risk of complications such as eye, kidney, and nerve damage, cardiovascular disease,

and certain cancers. Preventive measures and disease management can significantly reduce these risks [7].

It is reported that hypertension, as a significant factor along with many other factors, can lead to stroke, diabetes, heart disease, kidney damage, and other serious complications [8]. Hypertension, or high blood pressure, occurs when the force of blood against artery walls is consistently too high. The term "hypertension" derives from "hyper-" (over/beyond) and "tension" (stretching/straining), meaning "straining beyond." This condition places excessive strain on blood vessels [9].

There are two types of hypertension, primary and secondary. Primary hypertension has a gradual onset, it develops slowly over time. Other underlying medical conditions cause secondary hypertension which is often more sudden and severe [3].

My goal for this study is to use machine-learning approaches to help model the relationships between hypertension plus other health factors and the diseases of stroke and diabetes and develop an app to predict the likelihood of strokes and diabetes based on different lifestyle factors using the machine-learning models.

Method:

Machine learning involves the development of computer systems that learn and adapt without explicit programming, using algorithms and statistical models to analyze patterns and draw inferences from data (Oxford Dictionary).

It employs data analytics and machine learning algorithms to examine historical data, patterns, and trends. The more data available to machine learning algorithms, the better the predictions [10].

Various predictive models and algorithms exist. The two primary types relevant to my project are regression and classification models. Both fall under supervised learning, where models are trained on existing data with labelled outcomes. The key difference is that regression predicts continuous values, while classification assigns discrete labels.

A simple classification algorithm is logistic regression. Despite its name, logistic regression is designed for classification tasks. It predicts outcomes by calculating probabilities, and categorizing data points based on threshold values (e.g., >50% = class 1, <50% = class 0). Logistic regression is effective for binary

classification problems, such as determining whether an email is spam or whether a patient is likely to develop a condition [11, 12].

A more advanced model, random forest, is an ensemble learning method that constructs multiple decision trees and aggregates their predictions for improved accuracy. Each tree functions as an independent "expert," and the collective decision-making process reduces errors and overfitting. This algorithm has been widely adopted for medical predictions, such as assessing disease risks, and performed well with large datasets [13].

Both logistic regression and random forest were used in my project. Currently, my most accurate model employs random forest to predict the likelihood of stroke and diabetes based on health factors. By refining this model, I aim to provide a clearer understanding of how hypertension influences these conditions.

My Dataset Source [14]:

- The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based health survey by CDC or Centers for Disease Control and Prevention designed to collect data on health-related risk behaviours, 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 behaviours. Its primary purpose is to monitor public health trends, guide policy development, and support research to improve health outcomes.

In this study, I gathered two datasets from CDC. Each of them contains health information, stroke or diabetes respectively, along with hypertension, and other relevant factors. Using this data, I  trained and validated machine-learning models to predict the likelihood of stroke or diabetes with high accuracy. The model's performance was assessed based on its ability to analyze and interpret the relationships between the two diseases and the health factors including hypertension.

The detailed information in each database is given below:

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 (hypertension)
- Diabetes (yes or no)

The following steps were implemented in my Python code:

1. Import all necessary libraries (Numpy, pandas, etc.)
2. Upload and Prepare Data
3. Process data
4. Find missing values and replace them with a 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 a grid search for the best pseudo-parameters
9. Cross-validate and print results
10. Generate predictions and evaluate
11. Generate a confusion matrix (true positive, true negative, etc)
12. Plot feature importance
13. Calculate and print metrics (accuracy, precision, etc)

At first, I used the logistic regression model to predict the risks. The accuracy was 68% and 74% for stroke and diabetes respectively.

Then, I switched the algorithm to random forest and used the SMOTE technique, to improve accuracy because the dataset was imbalanced. SMOTE stands for Synthetic Minority Oversampling Technique. It's a technique used in machine learning to address imbalanced datasets by recognizing the data contains a minority class, similar to rare disease cases in a medical dataset [15].

After many tries, and fixing the errors, I finally got both models to work correctly using random forest. My stroke prediction model has improved dramatically with an accuracy of 99%. My cross-validation has confirmed no overfitting. My diabetes prediction model now has an accuracy of 75%, not a big improvement, but it still shows how switching the algorithm might impact the results.

Since my stroke prediction model appears more satisfactory, I decided to only make a stroke prediction app using Streamlit, leaving the development of a diabetes app to the future when I further improve the diabetes prediction model accuracy.

Streamlit is a Python framework that allows you to build interactive apps. My learning curve on Streamlit was built with a combination of YouTube videos and hours of practice. The stroke prediction app has the following features:

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

Below in Figure 1 and Figure 2, you can see some pictures of my app (although it is cut off).

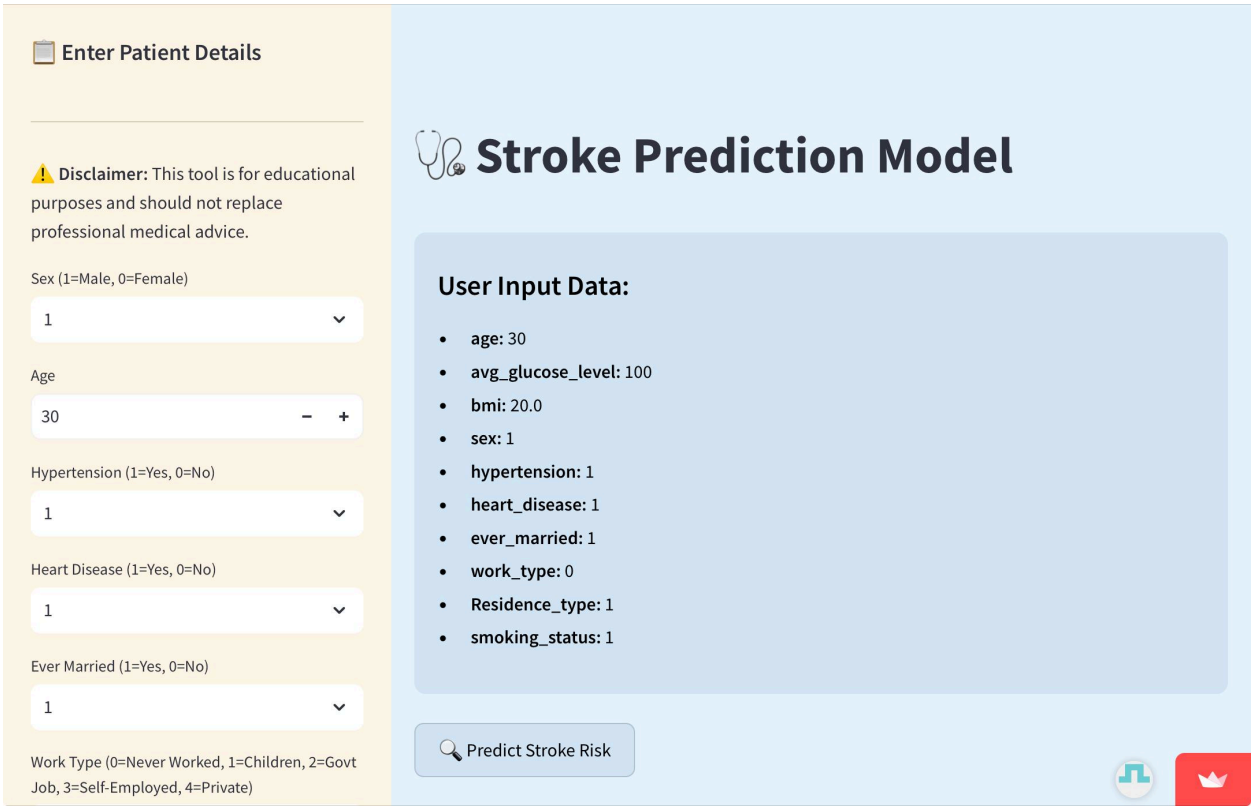



Figure 1. Input screen of Streamlit app

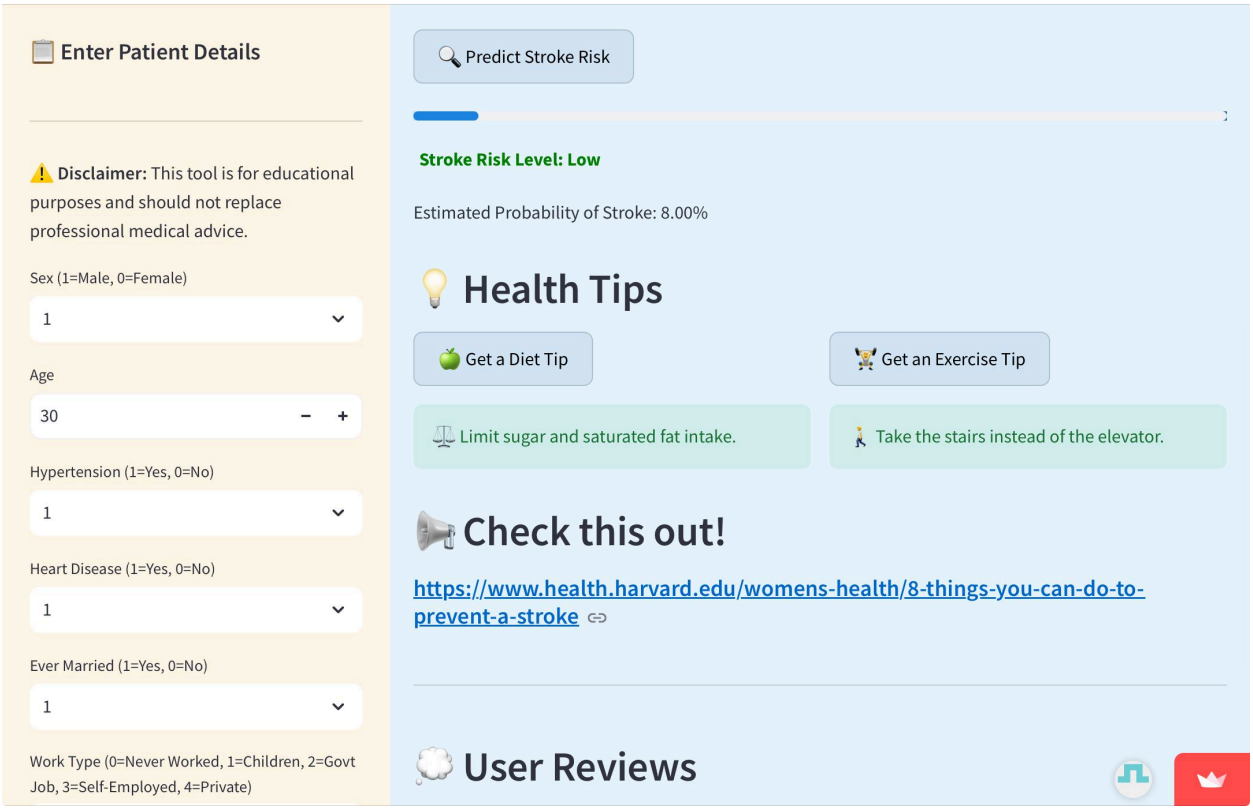



Figure 2. The resulting risk prediction screen with health tips

Analysis:

The accuracy of model prediction from this work is summarized in Table 1 below. The comparison shows that random forest is typically a better model to use when predicting health outcomes compared to logistic regression. This finding may help choose models for future medical-related projects.

Table 1. The accuracy of model prediction by logistic regression and random forest.

| Dataset | Accuracy with Logistic Regression | Accuracy With Random Forest |
|---|---|---|
| Stroke | 68% | 99% |

| | | |
|---|---|---|
| Diabetes | 74% | 75% |

Moreover, the feature importance from the random forest model has provided some insights into the most important factors for stroke and diabetes. The results are given below in Figure 3 and Figure 4.

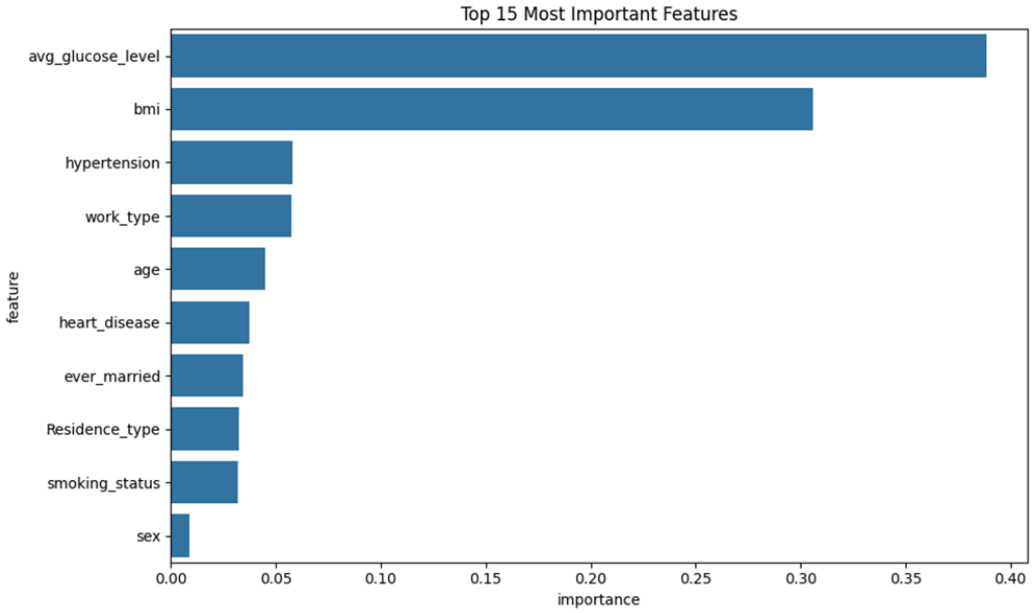



Figure 3. Top 15 most important factors for Stroke Prediction.

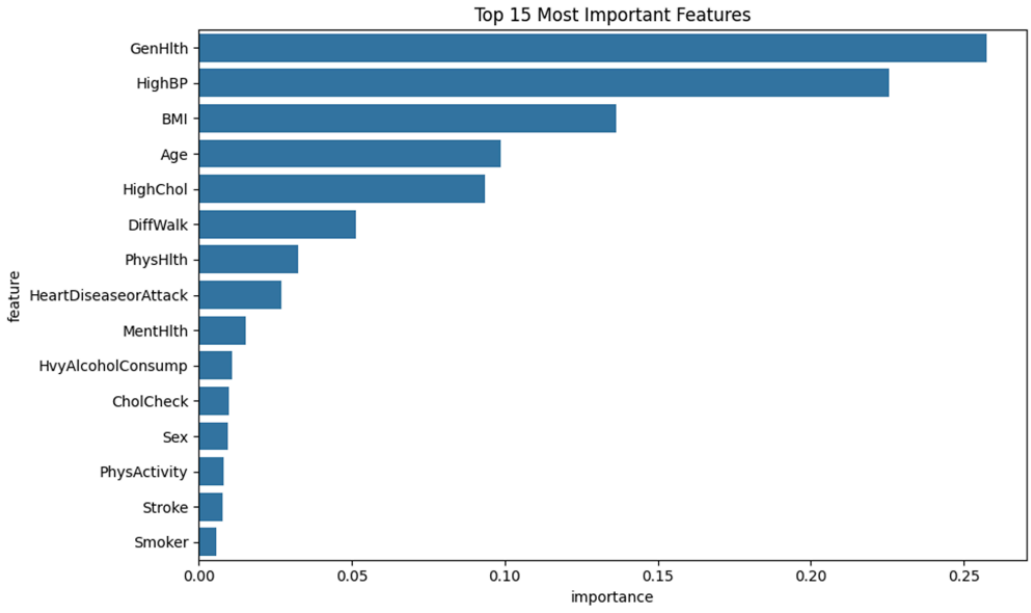


Figure 4. Top 15 Most Important Factors for Diabetes Prediction

The results suggest that glucose level and BMI are the two most important factors for stroke prediction. It is reported that high blood sugar, also called hyperglycemia, occurs when glucose levels exceed normal ranges. While transient hyperglycemia may not cause long-term issues, prolonged elevation can result in complications such as stroke, cardiovascular disease, vision impairment, kidney damage and nerve damage [16].

My model results confirmed the relationship between the blood sugar level and the risks of strokes. This result could indicate the potential stroke risks for diabetes patients.

In addition, besides high blood pressure, general health and BMI are among the top 3 important factors for diabetes prediction. Both stroke prediction and diabetes prediction suggest that general health and BMI impact your chances of developing chronic illnesses and other damaging health conditions.  Awareness of health conditions and lifestyle changes remain crucial to decrease the risks of strokes and diabetes, and thereby to live a long, healthy life.

Based on this study, hypertension is the 3rd most important factor for stroke and the second most important factor for diabetes. This confirms that hypertension can lead to serious medical implications. However, I believe that hypertension could have scored even higher. This may be because people who have hypertension will take medications to lower their blood pressure. Thus, people who have hypertension will be less likely to experience the syndromes it may bring about.

The results demonstrated that we can use machine learning to guide preventive healthcare.

Impact:

This project could provide insights into the predictive power of machine learning in healthcare, highlighting how analyzing health factors can assist in early detection and intervention. By demonstrating the potential of such models, this project underscores the value of integrating technology into healthcare to improve outcomes and encourage proactive health management.

My app could be used for future patient assessment surveys, helping to guide individuals in assessing their potential risk for stroke. When a person suspects they may be at risk, they can use my app as an initial screening tool before

seeking medical attention. This could save time and money, as individuals with an extremely low risk may not need to visit a doctor unnecessarily.

However, I acknowledge that no predictive model is 100% accurate, and my model will always have limitations. If this app is used in the future, it should not be solely relied upon for medical decisions. Instead, it should serve as a supplementary tool alongside professional medical advice.

It is important to recognize that similar technologies have already been developed. Many existing models use machine learning and artificial intelligence to predict stroke risks based on various health factors. However, my project aims to ensure that such tools are more user-friendly and practical for early health assessments.

Conclusion:

- Having a healthy lifestyle is essential for preventing serious health conditions like stroke and diabetes. Hypertension can lead to serious medical implications. My machine-learning model identifies key risk factors, reinforcing the importance of maintaining good health habits.
- To make these insights more accessible, I created a Streamlit app using my stroke prediction model. I invited my family members, close friends, and classmates to test the app, and their feedback was mostly positive. Through this process, I gained valuable insights into how machine learning can help raise awareness and encourage proactive health decisions.
- Additionally, my analysis showed that the random forest algorithm generally performs better than logistic regression, suggesting that more complex models may be more effective in predicting health risks when sufficient data are available. This reinforces the potential of AI-driven tools in medical prediction and prevention.

Limitations:

- Initially, I was hoping to use local health data in Alberta and Canada. I emailed a multitude of health organizations (AHS, CIHI, community health, 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. These private datasets

also bring up ethical issues, which is another reason I ultimately chose an ethics-approved public dataset.
- However, 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:

- More Diverse Data: Expanding the dataset to include a wider range of demographics and medical histories could enhance accuracy.
- Feature Refinement: Incorporating additional health metrics, such as sleep patterns, may improve predictions.
- Model Optimization: Testing other machine-learning algorithms or fine-tuning hyperparameters could further boost performance.
- User Experience: Enhancing my Streamlit app's interface and adding explanations for predictions could make it more user-friendly.
- Real-World Validation: Collaborating with healthcare professionals to validate predictions and explore practical applications would strengthen credibility.

My next step:

Introducing Phase 2: Using vessel segmentation to determine hypertensive retinopathy.
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.

Citations:

[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%20in,also%20develop%20in%20non%2Ddiabetics

Acknowledgement:

I acknowledge Ms. Lai who supported and guided me through this exciting process.

I acknowledge my mentors from Juniotech, Tim and Irada for providing their opinions and helping improve my project with their expertise in the computer science area.

Finally, I acknowledge Dr. Leyla Baghirzada, clinical assistant professor at the University of Calgary for her great help and guidance in the health aspect of my project.