PREDICTING DIABETES ONSET

Ashfi Hasnain (1163397) | Abhinav Mishra (1173645)

Lakehead University

COMP4310: Web Health Informatics

Dr M. Rathore

26 November 2024

Predicting Diabetes Onset

A very common disease is diabetes, which "is a condition that happens when your blood sugar (glucose) is too high. (Cleveland Clinic, 2023). There are various types of diabetes including Type 1, Type 2, Gestational, and others. According to the Canadian government, "Around 3.8 million people in Canada over a year old live with diagnosed diabetes", which accounts for 9.8% of the entire country's population. (Public Health Agency of Canada, 2024) According to the Mayo Clinic, there are several contributing factors, which include excessive glucose, not enough insulin, environmental factors, geographical factors, genetics, ethnicity, and whether or not the individual is overweight or suffering obesity. (Mayo Clinic, 2024) In the extreme form, diabetes can cause complications "like kidney disease, foot and leg problems, eye disease (retinopathy) that can lead to blindness, heart attack & stroke, anxiety, nerve damage, amputation and erectile dysfunction" (Canadian Diabetes Association, n.d.) Unfortunately, "Diabetes-related complications can be very serious and even life-threatening. Properly managing blood sugar levels reduces the risk of developing these complications," however. (Canadian Diabetes Association, n.d.)

Unfortunately, "There's currently no cure for diabetes, but some treatments may help to manage it and improve quality of life" (Public Health of Canada, 2024). Thus, we want to ensure that the worst complications can be prevented. But in order to do so, we want to know if someone is at risk or not from having diabetes. Once a person knows they are at risk, they can consult a doctor at their earliest convenience to plan out managing diabetes and maintaining a good quality of life. Therefore, it would be convenient if there was some software that could help them in this step – which can be fullfilled by the *Predicting Diabetes Onset* Tool.

Initial Steps on Development

We come to the process of developing such a program. It would need some dataset to train the machine learning model. We also would need a compiler and a language to compile the program.

Our first question is what language do we use? Our answer is Python, which has several advantages. (Terra 2024) Firstly, Python is a flexible language, which is "Ideal for developers who want to script applications and websites". It is also very simple and readable for both user and computer. It achieves this by using less code to do tasks than any other language. Python can run on many platforms and has support for many open-sourced libraries, which also makes it useful for "Data manipulation, Data Visualization, Statistics, Machine Learning", all of which we will need in such program. Finally, Python is an industrial standard and lots of support resources exist for this language. Ergo, Python is the perfect language to use. A very good compiler for Python is Google Colab, which allows you to attach a .csv dataset and several collaborators to work on development.

We have a language and compiler to develop the machine learning model. But we need a dataset to train the machine learning model. We used the Pima Indians Diabetes Dataset collected by the National Institute of Diabetes and Digestive and Kidney Disease, from 1988 (Smith et al.,1988). It shows whether a sampled individual has or does not have diabetes depending on eight attributes:

1. **Pregnancies** – How many time the individual was pregnant
2. **Glucose** – Plasma Glucose Concentration [Average = 85]
3. **Blood Pressure** – Diastolic Blood Pressure (mm Hg)

4. **Skin Thickness** – Thickness of Triceps Skin Fold (mm) [Average = 18.7]
5. **Insulin** – Insulin concentration (µU/ml) [Average = 10]
6. **Body Mass Index**
7. **Diabetes Pedigree Function**
8. **Age**

Now, we have a dataset. That means we have all the prerequisites to start developing the program. First, we import several libraries we will need to develop the program. Then we import the dataset, which we also attached to the project to ensure the program can use it, irrespective of who the user is. We want to hold the data to high quality standards. This means removing the incomplete data and duplicate data. Thus, the first part of the program is one that does several null checks and duplicate data checks. All incomplete data is replaced with averages, so that they are complete without altering the trends. Fortunately, there were no duplicate data, so we did not have to remove anything.

Data Findings

We had 768 tuples representing 768 individuals in our datasets. 500 individuals tested negative for diabetes and 268 tested positive, as seen in figure 1.

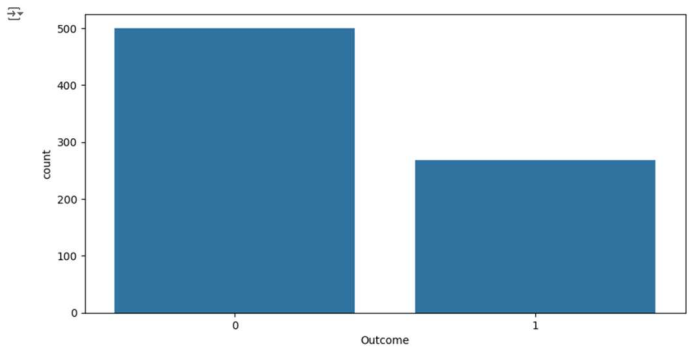

Figure 1: The Outcomes of Diabetic and Non-Diabetic

We also analysed the data to see if any of the eight factors had any outliers or not (Figure 2). Pregnancies only had three outliers while glucose levels had no outliers. All the remaining factors had numerous outliers, mostly on the extremely high end, except for blood pressure, which also had some outliers on the extremely low end.

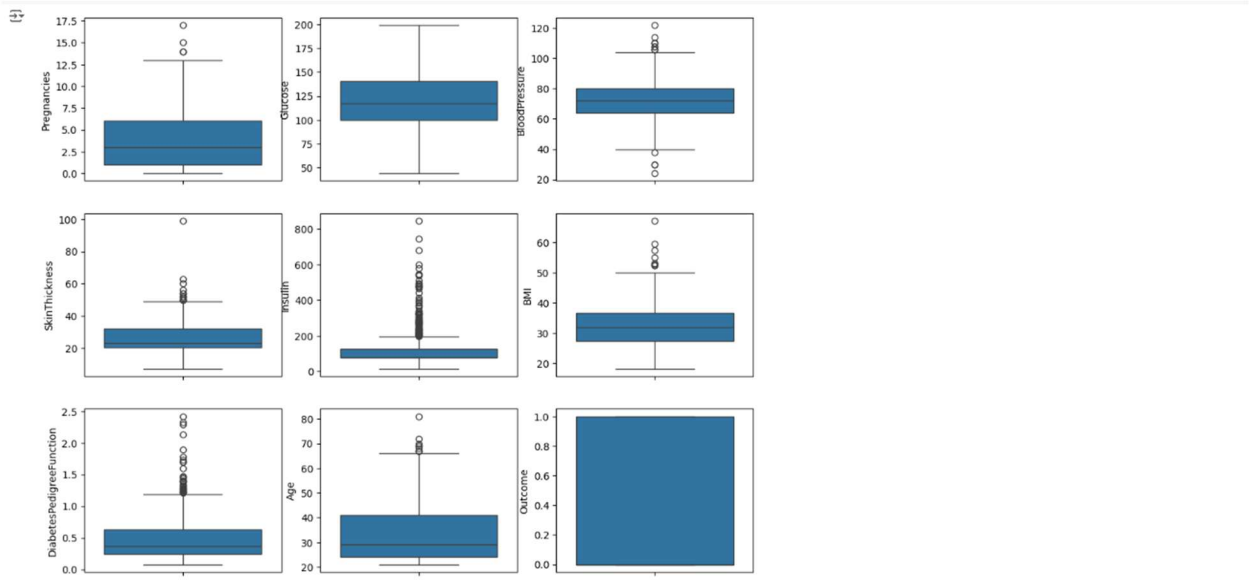

Figure 2: Outliers in each dataset

While we had some outliers, the data seems to be usable. For each of the factors, we also analysed the distribution of each factor (Figure 3), which proves that none of the data was unreliable. Some of the data, dependening on their nature, were obviously skewed towards one end, such as pregnancies, but that is natural as there are more people with 0-3 pregnancies than over 10. Another example is the age, but this only implies that younger individuals went to check their diabetes.

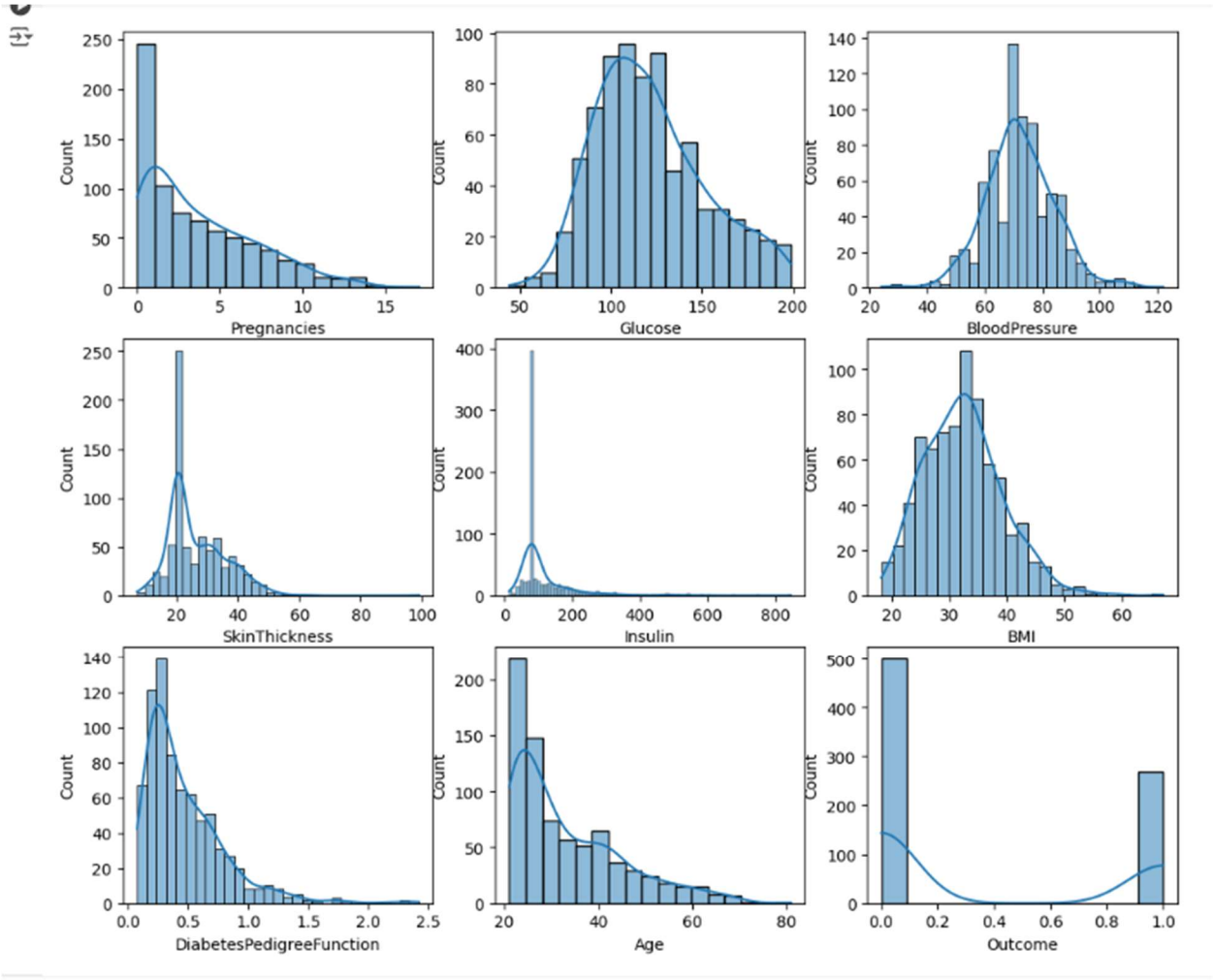

Figure 3: Distribution Histograms of Each Factor

Now that we have defined analysed each factor, we can now see how each factor impacts another factor (Figure 4). This will be useful, especially when studying which factors are most related to diabetes, which will also be compared in figure 4. To see correlation between two factors, such as *glucose levels* and *diabetes outcome*, we can use Pearson's correlation coefficients, such as in a heatmap like on figure 4.

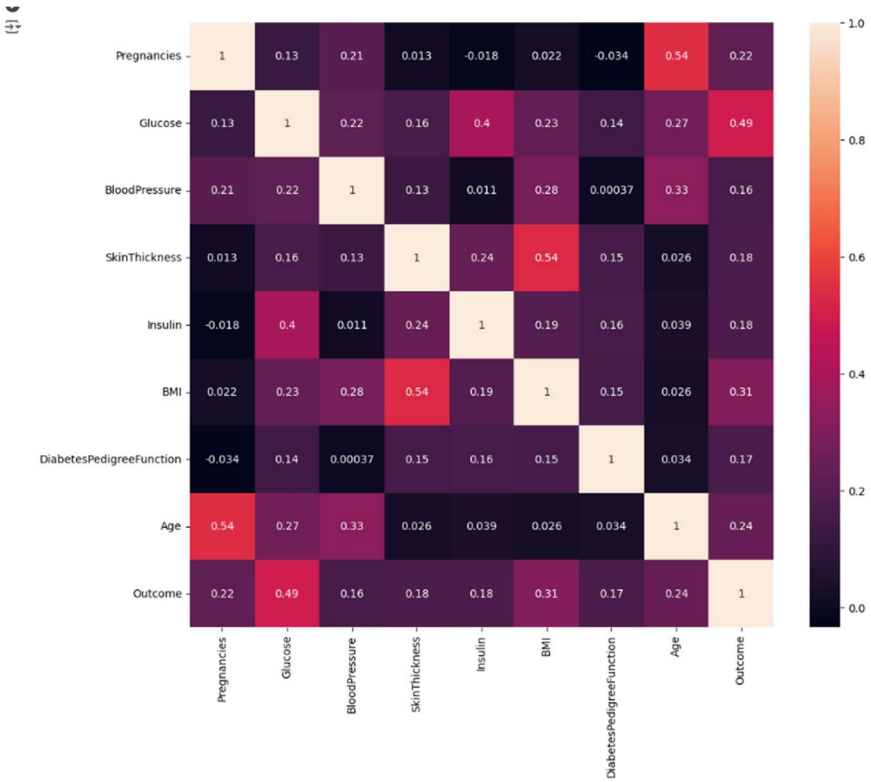

Figure 4: Heatmap of Correlation Between Two Factors

From Figure 4, we can see not much correlation between factors, except for a few pairs:

- Pregnancies and Age
- Skin Thickness and Body Mass Index
- Glucose and Diabetes Outcome

However, instead of using data and identifying trends manually, we will use a machine learning model to analysing correlations and drawing conclusions on which we can build a program to test for diabetes risk.

Machine Learning Model

Model Training and Preparation

The model training process began by defining the feature matrix (X) and the response vector (Y) from the dataset. The outcome variable, which indicates the presence or absence of diabetes, was designated as the target (Y). The remaining features served as predictors (X). Subsequently, the dataset was split into training and testing subsets, with 80% of the data allocated for training and 20% for testing. This split was conducted to ensure the model's generalizability when applied to unseen data. For consistency, the random_state parameter was set to 42.

To streamline the modeling process, a pipeline was constructed using Scikit-learn. This pipeline integrated preprocessing and model training steps, ensuring a standardized workflow. StandardScaler was employed to normalize the features, a critical step for optimizing algorithm performance. Several machine learning algorithms, including Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forests, and Gradient Boosting Classifiers, were incorporated into separate pipelines to evaluate their predictive capabilities.

Figure 5: Data Split, Training and Pipelining

Standardization of Data

Standardization is a vital preprocessing step in machine learning pipelines. It scales the features to have zero mean and unit variance, thus enhancing the performance of algorithms sensitive to feature scaling. Each algorithm was wrapped in a pipeline alongside StandardScaler to ensure consistent feature scaling. For example, the Logistic Regression pipeline contained both the scaler and the classifier. Similar configurations were applied for other classifiers such as K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forests, and Gradient Boosting.

Figure 6: Data Standardization

Algorithm Selection and Evaluation

To identify the optimal model, training datasets were passed through the respective pipelines. The models were then evaluated on the testing data to measure their predictive accuracy. Each pipeline's accuracy was calculated and displayed, revealing the performance differences among the algorithms. The Random Forest classifier emerged as one of the best-performing algorithms, achieving a notable accuracy on the test dataset. This step informed the subsequent selection of the Random Forest model for further analysis.

Implementation of Random Forest

After identifying Random Forest as the preferred algorithm, a dedicated implementation was carried out. This involved training the model using the training data and subsequently saving it for reuse. The joblib library was utilized for model serialization, allowing the trained model to be stored and loaded efficiently for future predictions.

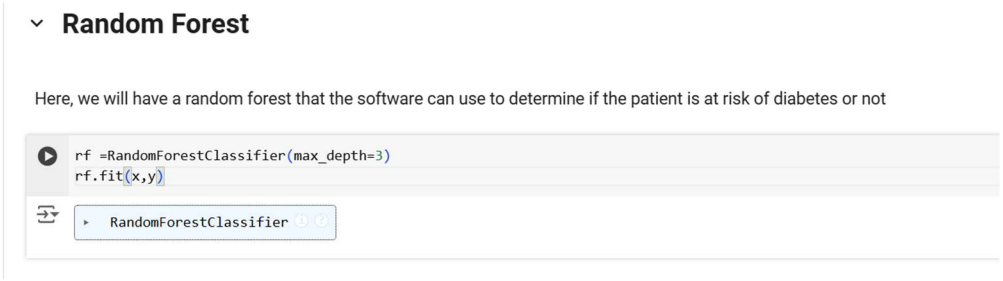

Figure 7: Algorithm Implementation

Model Testing and Validation

To validate the trained model's performance, a dummy dataset was created to simulate real-world input scenarios. This dummy data was fed into the trained Random Forest model to evaluate its prediction capabilities. The results confirmed the model's accuracy and robustness, ensuring its readiness for deployment.

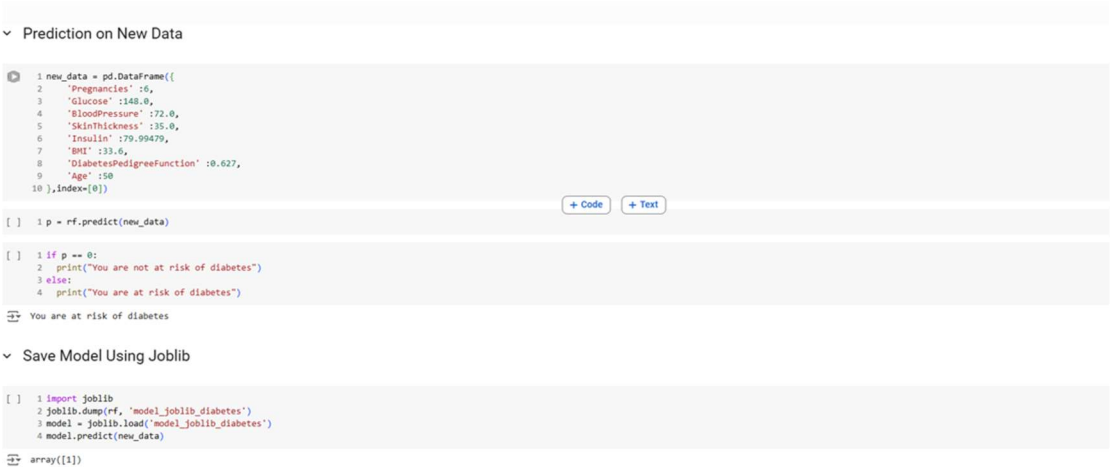

Figure 8: Model Prediction Testing

The Graphical User Interface

Now that we have successfully developed a machine learning model to determine if a user is or is not at risk of diabetes, we can use this knowledge to develop a graphical user interface for the software. Figure 9 shows a screenshot of the graphical user interface.

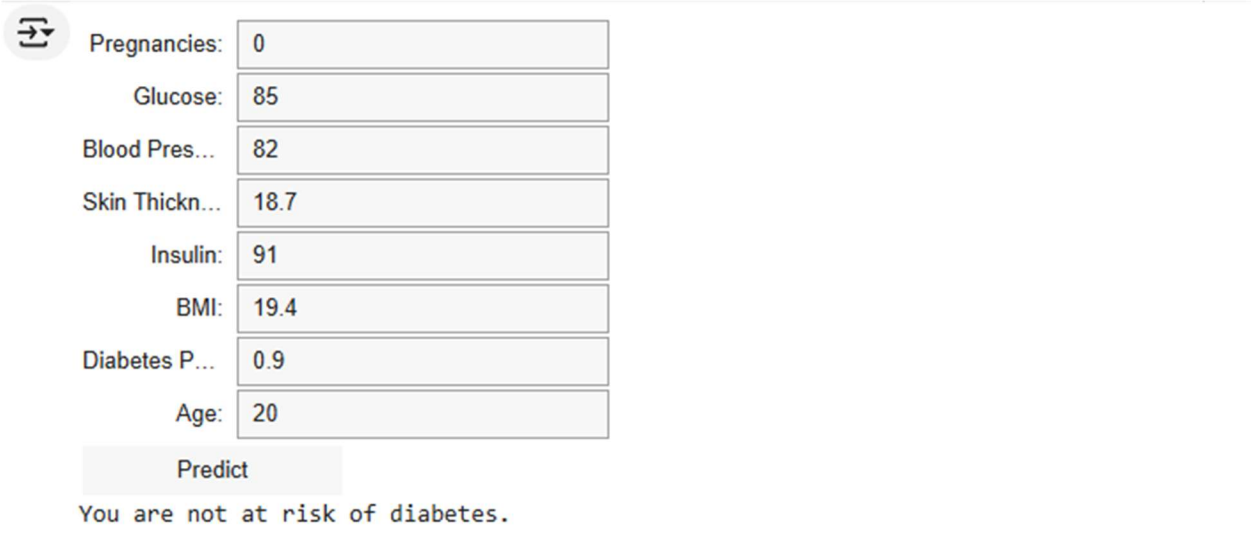

Figure 9: A Graphical User Interface for our Software

In figure 9, you can see the options to select a value for each of the eight variables that can affect the risk of diabetes. Once the user is satisfied with their selection, the press predict to get their diagnosis. In figure 9, the user enters the values of [0, 85, 82, 18.7, 91, 19.4, 0.9, 20] and gets a result of being safe from diabetes.

Challenges and Potential Improvements

We faced several challenges as we worked on developing this program. They included these problems:

- Initially, the CSV data was uploaded through a connection from the Google Colab notebook to Ashfi's Google Drive. However, this meant Abhinav could not access
    - We instead uploaded the CSV data file into the program and resolved the problem
- We wanted the GUI to be its own window (independent from the Colab file), but Google Colab does not support such
    - It would require using a different IDE

After addressing the difficulties we encountered, it is also a opportunity to brainstorm potential improvements for such a software:

1. Find a training data that supports all ethnicities – not just one as this dataset does.
    - This dataset was based on the Pima Indians, the indigenous community based in Arizona, US. Ideally, we want a datasets that has studied with other ethnicities.
2. Make this GUI a separate desktop and mobile application, supporting as many operating systems as possible
    - We could use android studios to develop an android version and use other compilers (such as Visual Studio or IntelliJ) to develop such a GUI.
3. Have a medical doctor, preferably one specialised in diabetes, to assess the accuracy of this software

Conclusion

At last, using data analysis, machine learning, and a graphical user interface (GUI), we have created a software that can help a user identify whether or not a user is at risk of diabetes or not. With a few further steps, this software could be widely used. This would be a great time to reflect on this project:

Ashfi: "*This project taught me a lot about data analysis, as well as using computer/programming technology to do such tasks. While I am still not confident in my Python abilities, I feel this project turned me in the right direction. In my teen years, I wanted to be a doctor but that did not work out. This project made me feel reconnected with a past life.*"

Abhinav: "*The project has been a very good recap of my prior knowledge of analysis. Learning how to make a machine learning model and be able to implement it, is a success in my journey of becoming a data scientist. The project also reflected the importance of learning python language and how efficient the language is. We created a tool which can be used by anybody to check if they have diabetes or not.*"

Works Cited

Canadian Diabetes Association. (n.d.). *What is Diabetes*. Retrieved from Diabetes Canada: https://www.diabetes.ca/about-diabetes-(3)/what-is-diabetes

Cleveland Clinic Medical Professionals. (2023, February 17). *Diabetes*. Retrieved from Cleveland Clinic: https://my.clevelandclinic.org/health/diseases/7104-diabetes

Li, W., Yin, H., Chen, Y., Liu, Q., Wang, Y., Qiu, D., . . . Geng, Q. (2022, May 10). *Associations Between Adult Triceps Skinfold Thickness and All-Cause, Cardiovascular and Cerebrovascular Mortality in NHANES 1999–2010: A Retrospective National Study.* Retrieved from NIH National Library of Medicine: https://pmc.ncbi.nlm.nih.gov/articles/PMC9127233/#:~:text=In%20the%20study%20population%2C%20the,vs%2014.3%20%C2%B1%206.8%20mm).

Mayo Clinic Staff. (2024, March 27). *Diabetes*. Retrieved from Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

Public Health Agency of Canada. (2024, October 23). *Diabetes: Overview*. Retrieved from Government of Canada: https://www.canada.ca/en/public-health/services/chronic-diseases/diabetes.html

Riley, L. (1999). *Mean fasting blood glucose.* Retrieved from World Health Organization: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2380#:~:text=The%20expected%20values%20for%20normal,and%20monitoring%20glycemia%20are%20recommended.

Smith, J., Everhart, J., Dickson, W., Knowler, W., & Johannes, R. (1988). *Pima Indians Diabetes Database.* Retrieved from Kaggle: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data

Straus, J. F., & Barbieri, R. L. (2014). *Yen & Jaffe's Reproductive Endocrinology.* Boston: Elsevier Ltd.

Terra, J. (2024, October 12). *Why Python Is Essential for Data Analysis and Data Science*. Retrieved from SimpliLearn: https://www.simplilearn.com/why-python-is-essential-for-data-analysis-article