**APS360 Winter 2020 Project Final Report**

**Team 62**

**Sean Llera - 1004928536**

**Nathan Jones - 1005003023**

**Priscilla Deng - 1005227826**

**Sadra Tonekaboki - 1004899461**

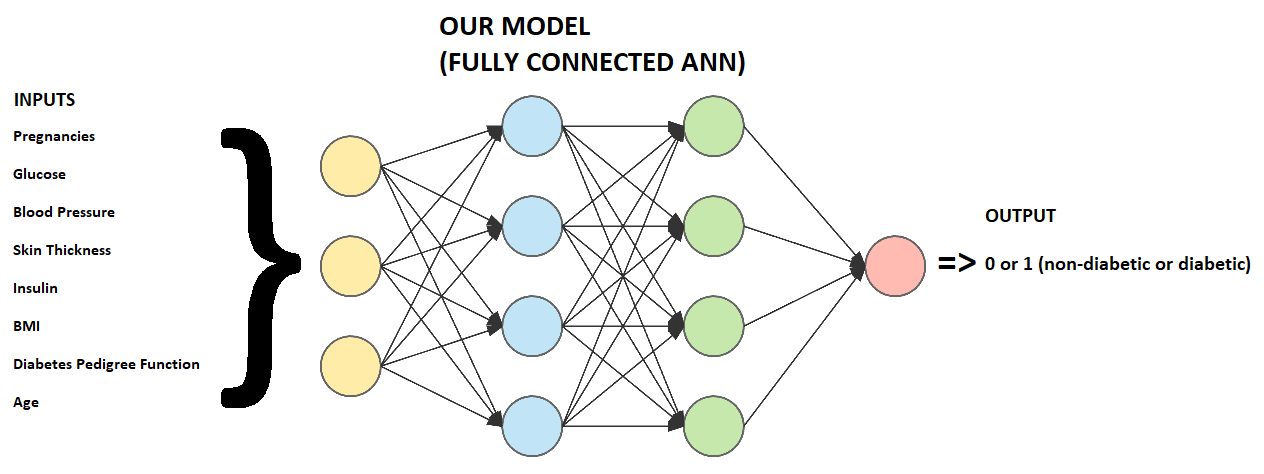
**WORD COUNT: ~1900**

### Introduction

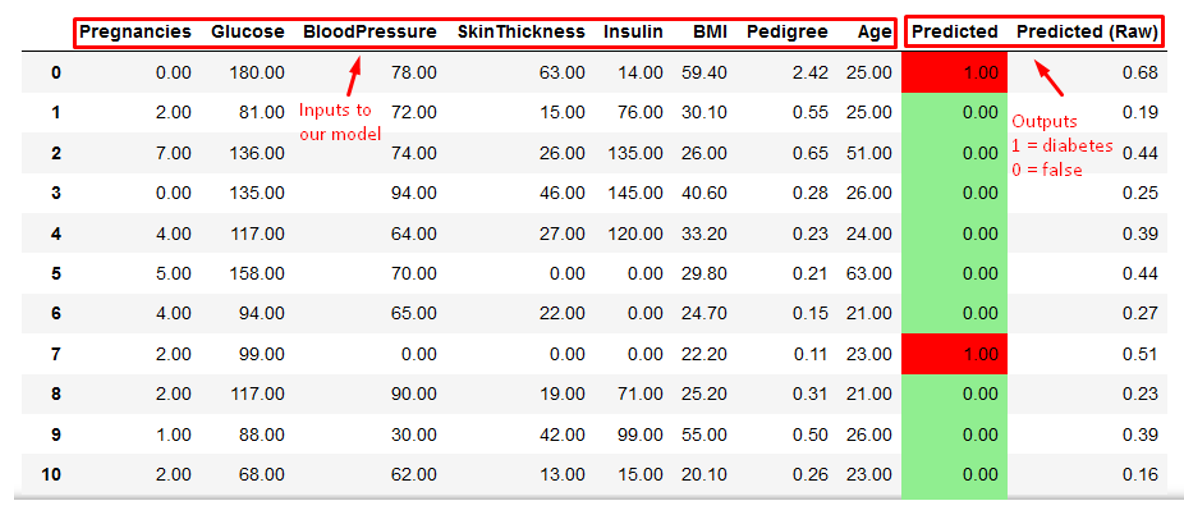
Diabetes is an increasingly prevalent chronic health condition that restricts the body’s ability to metabolize glucose [1]. This disease affects more than 2.3 million Canadians and forces affected individuals to adjust to a difficult and restrictive diet [2], along with necessary invasive tests to determine safe glucose levels. Due to COVID-19, more people are transitioning from the traditional workplace into a more casual setting, putting many at risk to an inactive and sedentary lifestyle. However, physical inactivity and obesity are the top two risk factors in developing type 2 diabetes [1].

Though diabetes is irreversible, early detection can delay or even prevent its development [3]. We decided to create an algorithm that determines whether an individual may have diabetes based on the following factors: age, glucose levels, blood pressure, BMI, insulin, number of pregnancies, skin thickness, and diabetes pedigree. Since diabetes involves binary classification on data analysis, machine learning can be much more time and resource efficient than manual diagnosis. However, the goal of this algorithm is not to replace a diagnosis by a licensed health professional - instead, it aims to serve as a precursor - if a health professional has the information from their patients on hand, they can run it through the algorithm to determine the likelihood the patient has diabetes. If the output results are high, patients can then be contacted for a thorough, in-depth manual diagnosis. When detected early, individuals have a higher chance of delaying or preventing the development of diabetes [3].

### Illustration / Figure

*Figure I: Simple Model Diagram*

1. Obtain data from dataset
2. Import data/parameters into machine learning algorithm
3. Identify trends and extrapolate information
4. Output results



*Figure 2: Machine learning algorithm output - predicts ‘1’ if the model detects diabetes, 0 if false.*

### Background & Related Work

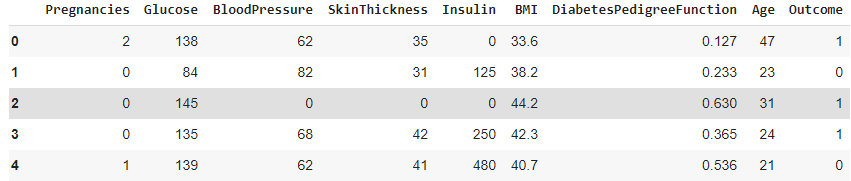
Researchers from the Vellore Institute of Technology developed a machine learning based diabetes mellitus detection algorithm just last year, with a prediction accuracy of 85.6% [4]. However, this project focuses more on diabetes mellitus, a condition that leads to multi-organ failure, rather than diabetes detection as a whole. One of the datasets used was from the National Institute of Diabetes and Digestive and Kidney Diseases, which contains information on diabetic female individuals at least 21 years old of Pima Indian heritage [5]. This dataset can be found on the UCI Machine Learning Repository for public use.

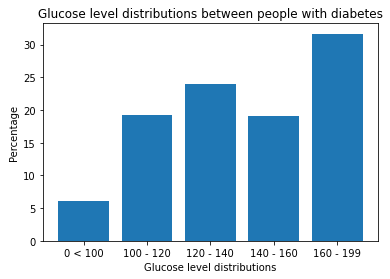
There are other datasets on Kaggle, specifically Early Stage Diabetes Risk Prediction Dataset, which describes the symptoms of individuals at risk of developing type 2 diabetes [6]. Another dataset describes health-related information such as sugar levels and blood pressure of diabetic patients from Germany [7].

### Data Processing

Our data was taken from Kaggle [8]. We found 2 datasets that contained the same types of information, so we decided to combine them to get a larger dataset. The first dataset is a dataset titled *diabetes* and another dataset titled *Pima Indians Diabetes Database* [7][8].

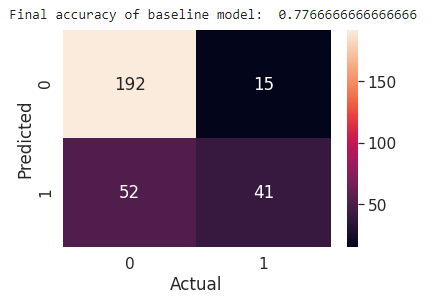
The input of the dataset include number of pregnancies, plasma glucose concentration 2 hours after a glucose administration, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), insulin levels 2 hours after a glucose administration (mu U/ml), BMI, diabetes pedigree function (function that determines likelihood of diabetes based on family history), and Age [8]. The output of the dataset is in a column called Outcome that has a value of 0 meaning non-diabetic or 1 meaning diabetic. The data was split randomly into 75% - 15% - 15% training, validation, and testing sets. The following figures display samples from the dataset and the glucose levels distribution across patients with diabetes.

*Figure 3: Samples from dataset*



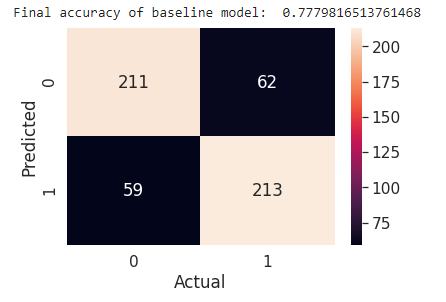
*Figure 4: Graph displaying the glucose level distributions across diabetes patients*

Our initial dataset, which was the dataset titled *diabetes* consisted of 2000 entries but had 1316 non-diabetic entries and 684 diabetic entries. This imbalanced dataset resulted in the baseline model predicting more false positives than true positives, as seen in the following figure.



*Figure 5: Baseline results on the initial imbalanced dataset*

To try balancing and enlarging the dataset we added the entries of the *Pima Indians Diabetes Dataset* to the initial dataset, which resulted in 2768 entries with 1816 non-diabetic entries and 952 diabetic entries. This dataset was larger but still very imbalanced, so we followed what was done in Lab 5 and duplicated 864 diabetic entries so that the number of diabetic entries was equal to the non-diabetic entries. This resulted in the baseline model predicting many more true positives compared to before.



*Figure 6: Baseline results on the larger balanced dataset*

### Architecture

The neural network model we used was a 4 layer sequential ANN model, including the input and output layers. The model used Binary Cross Entropy loss alongside the Adam optimizer. BCE loss was used as the core problem of the model is binary classification. The model was trained with 1600 epochs, a learning rate of 0.001, and a batch size of 50. The size of the intermediate layers were 60, 30, 20 respectively.

### Baseline Model

Since our problem was one of classification and simple data we chose a simple regression model without any hidden layers as our baseline. Since the main ways of detecting diabetes is through the abnormalities in the blood sugar and insulin levels, we perceived that if a simple correlation of a few parameters was all that was needed, then a regression model would suffice. The results showed us that the best that can be achieved with the baseline model is an accuracy of 77% which showed us that all parameters such as the genetic information and pregnancy history played a significant role in predicting diabetes and that the more complex correlations required the use of neural networks.

### Quantitative Results

The key measure of success used by our team for this project is the scale of the area under the ROC graph used by the national institute of health in the United States [9]. The scale can be seen on the table on the right [10]. Following the formulas provided by the health institute we found that our final model has an area of 9.8 which landed us in a comfortable spot.

Now looking at the finer details of the output, we tested the merit of each model once and only once with the testing dataset. Each model was optimized using the validation set and once fully optimized, it was tested with the testing set. If the resulting accuracy was inadequate, the model would be deemed suboptimal and completely disregarded. We only used the testing dataset twice, once to disregard our first model, and once again to conclude that our final model was adequate and therefore the final testing accuracy of 97.8% was obtained free of any bias.

### Qualitative Results

The misdiagnosis of diabetes is very uncommon which means that its symptoms and causes are well studied and easily quantifiable, to a point where the blood sugar levels of a diabetic person and a non-diabetic one can be distinguished even by an untrained individual (as seen on the graph on the right) [11]. Therefore it comes as no surprise that our second model was able to reach a near perfect accuracy. The reason that a perfect accuracy is impossible is that many factors such as an unreported pregnancy or an uncommon trait of a single individual may cause some amount of imperfection in the data.

### Evaluate model on new data

Much effort was taken to ensure that the results were a good representation of the model’s performance on new data. First, the dataset, which consisted of 1816 diabetic entries and 1816 non-diebetic entries, was shuffled and split into training, validation and testing sets as mentioned in the Data Processing section. The model was then trained and optimized using only the training and validation sets. The testing set data, which had a size of 300 entries (evenly split between diabetic and nondiabetic), was left untouched by the model until the final accuracy test.

Due to these efforts, there is very little bias in our model and data, and we are confident that our model’s results are an accurate representation of its performance on new data. If new data samples that included the same input as our dataset were to arise, we believe our model would accurately diagnose the patients.

### Discussion

At a glance, our model has a fantastic testing accuracy of 97.9%, which even by medical standards, is amazing. Such a high accuracy may seem unbelievable at first, but then we realized that diabetes detection is at its core, a binary classification problem. Individuals with diabetes have much higher blood sugar levels (almost twice as much as healthy patients) and other features easily distinguishable from nondiabetic individuals, as mentioned earlier in *Qualitative Results.* These common and quantifiable traits are easily observable by the model, and with enough learning, the algorithm can accurately distinguish healthy and diabetic patients - save for a few outliers.

In the future, we could further improve the model by examining the outliers (low BMI, low blood sugar levels) that were either false-positives or false-negatives, and using these entries to further tune our model to achieve an even higher accuracy. However, this may be difficult, as patterns may be erratic with outliers and cause our project to become unstable.

### Ethical Considerations

For ethical considerations we will make sure that consumer data is held with the highest level of privacy, including making sure that no data collected by this model is used by companies for any nefarious purposes (ie, insurance companies using the data to give higher rates to people who may not have diabetes, but are likely to be diagnosed later in life). We will also make sure to receive consent for all data used in this model (whether that be through using public data, etc). Furthermore, if this model is ever used in medical institutions, patients can sign up using an opt-in/opt-out basis. Not everyone wants their information to be processed by an algorithm, so this choice will allow them to keep their privacy.

### Project Difficulty / Quality

As the project proposal was written at the same time as the first and the second labs, the scope of the project was well within the abilities of the team, and thanks to a great team effort a good model was found fast enough to leave a great amount of time to optimize it while at the same time allowing us to research the topic at hand for both generic and data driven information.

One difficulty that the group ran into was the program of processing multi-variate inputs. As our labs worked with single variable problems, there were many factors that needed to be considered with multiple variables. For instance, how much weight should each factor contribute to the result? How should the loss function be calculated with so many variables?

For the baseline model, we chose a very simple linear regression model that used mean-squared error to account for multiple variables. This loss was minimized with a sigmoid activation function. We considered assigning weights in our final model as well, but in the end, we chose to use batch normalization as this proved to be more stable. Individually initializing the weight for each factor proved too difficult and time-consuming as there were too many permutations with an 8-input model. Batch normalization also allowed us to use cross entropy loss functions, whereas MSE losses proved unstable and would jump to infinity and NaN. Furthermore, since none of us had prior medical knowledge on diabetes, we decided that keeping all weights equal was the most ‘fair’ and unbiased approach to our project.

Project Colab Link:

<https://colab.research.google.com/drive/1sFjGA3iVXgr_Gf-CwkyGVpvKqTpihLfV?usp=sharing>

Reference List

[1] “Physical inactivity and diabetes,” World Health Organization, 12-Nov-2015. [Online]. Available: <https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/news/news/2015/11/physical-inactivity-and-diabetes>. [Accessed: 11-Feb-2021].

[2] P. H. A. of Canada, “Government of Canada,” Highlights: in Canada: Facts and figures from a public health perspective - Canada.ca, 15-Dec-2011. [Online]. Available: <https://www.canada.ca/en/public-health/services/chronic-diseases/reports-publications/diabetes/diabetes-canada-facts-figures-a-public-health-perspective/report-highlights.html>. [Accessed: 11-Feb-2021].

[3] Y. Li, D. D. Wang, S. H. Ley, M. Vasanti, A. G. Howard, Y. He, and F. B. Hu, “Time Trends of Dietary and Lifestyle Factors and Their Potential Impact on Diabetes Burden in China,” Diabetes care, Dec-2017. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/29046327/>. [Accessed: 10-Apr-2021].

[4] J. Chaki, S. T. Ganesh, S. K. Cidham, and S. A. Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review,” Journal of King Saud University - Computer and Information Sciences, 04-Jul-2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1319157820304134>. [Accessed: 11-Feb-2021].

[5] UCI Machine Learning Repository: Diabetes Data Set. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/diabetes>. [Accessed: 11-Feb-2021].

[6] I. Dutta, “Early Stage Diabetes Risk Prediction Dataset,” Kaggle, 21-Sep-2020. [Online]. Available: <https://www.kaggle.com/ishandutta/early-stage-diabetes-risk-prediction-dataset>. [Accessed: 11-Feb-2021].

[7] John, “diabetes,” Kaggle, 25-Apr-2018. [Online]. Available: <https://www.kaggle.com/johndasilva/diabetes>. [Accessed: 11-Feb-2021].

[8] U. C. I. M. Learning, “Pima Indians Diabetes Database,” Kaggle, 06-Oct-2016. [Online]. Available: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>. [Accessed: 22-Mar-2021].

[9] A.-M. Šimundić, “Measures of Diagnostic Accuracy: Basic Definitions,” EJIFCC, 20-Jan-2009. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975285/>. [Accessed: 09-Apr-2021].

[10] Available: <https://effectivehealthcare.ahrq.gov/products/test-performance-metrics/appendixes>. [Accessed: 09-Apr-2021].

[11] D. D. Staff, “Is My Blood Sugar Normal?,” Diabetes Daily, 10-Aug-2020. [Online]. Available: <https://www.diabetesdaily.com/learn-about-diabetes/understanding-blood-sugars/is-my-blood-sugar-normal/>. [Accessed: 09-Apr-2021].