Course: INF2178

Term: 2022 Winter

Name: Kuo-Lun Chang

Student number: 1007618641

Instructor: Shion Guha

1. Introduction

1.1. Problem Statement and Importance

Diabetes has a long history of being one of the most prevalent diseases worldwide and causes many deaths every year. Rather than being a standalone disease, diabetes could also trigger many complex complications, which led to even more deaths [1]. Researchers conducted numerous tests and medication methods to help patients recover or delay deterioration from diabetes that improve a lot of people's lives' quality. In this paper, I will use a past dataset with 768 anonymous patients' data to build a prediction model to see if patients would have diabetes or not.

1.2. Background

According to the World Health Organization, diabetes is the ninth leading cause of death with 1.5 million deaths directly related to diabetes worldwide in 2019 [2]. In Canada, 12.1% of the population, around five million Canadian, will be under the threat of getting diabetes at the end of 2025 [3].

In general, there are three main types of diabetes. Type I diabetes is an autoimmune condition. Both genes and cells in the pancreas fail to make insulin could be the potential cause. Patients with this type of diabetes are usually diagnosed in their youth. Type II diabetes is non-insulin-related diabetes, mainly raised from obesity, fat, and unhealthy living styles, accounting for more than 90% of diabetes patients nowadays [4]. The third type of diabetes is Gestational diabetes. Some women will have some form of insulin resistance during pregnancies, which later causes diabetes. Some people also categorize prediabetes as a type of diabetes. However, since prediabetes represents patients with higher-than-usual blood sugar levels but not high enough to diagnose as Type II diabetes, prediabetes would not be in the scope of this research.

Thus, it is particularly critical to building a model that could predict potential diabetes patients with several clinical outcomes. There are some advantages of building this prediction model. First, this model could potentially reduce physicians' time manually to diagnose diabetes. Although diabetes is not a rare disease and requires advanced medical knowledge and training to be diagnosed, this model could be implemented along with the electronic health records and sent alerts to physicians proactively to identify potential patients. Second, with careful explanation and slight modification of the model, the public could use this model as a diabetes risk assessment. Thus, people with higher risk could discuss prevention methods with their family doctors before convicting to diabetes.

1.3 Research Question and Objectives

The purpose of this project was inspired by the National Institute of Diabetes and Digestive and Kidney Diseases' diabetes dataset [5]. The object of this dataset is to predict whether a patient has diabetes by using several measurable outcomes. This research will conduct multiple measurable outcomes (factors) comparisons and ultimately find two factors with the highest correlations to predict diabetes.

The following hypotheses were derived from our research question: (1) some factors in the original dataset may have high correlations with the existence of diabetes while they may be diabetes' syndrome instead of direct cause. (2) since the ANOVA test could only use on categorical data, I logically grouped and assigned affected data into new categories. 2. Methodology

2.1 Data Collection

The dataset contained 768 anonymous patients' clinical information from the National Institute of Diabetes and Digestive and Kidney Diseases institution. This institution is a part of the United States National Institute of Health and aims to support research, training, and communication regarding diabetes and other diseases. The dataset included pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and final diabetes verdict.

2.2 Data Analysis Method

In this research, I first cleaned the dataset with the tidy data concept and summarized the dataset using descriptive statistics. Before applying two-way ANOVA and two-way ANCOVA to test the dataset, I made a copy of the original dataset but categorized 'Glucose,' 'Blood Pressure,' 'Insulin,' 'BMI' into different levels according to variable related categorization criteria. After using ANOVA and ANCOVA with a post-hoc test to figure out the best two diabetes predicators, I used the linear regression function to build the final model, find the accuracy, and test the power.

2.3 Data Cleaning Method

2.3.1 Missing data

There is no missing data in this dataset. However, after looking into the data for great detail, multiple 0s existed across variables, which did not make any sense. For example, it is impossible for any living human being to have 0 skin thickness, blood pressure, or BMI. Thus, I did some data processing for these 0s. In general, there are two ways to deal with missing data-delete or imputation dataset. Since the dataset only contains 768 datapoints with at most 375 0s in one variable, it did not make sense to delete all these data points and only analyze the rest of them. Thus, this research uses the imputation method to deal with missing data. More specifically, the study used the variable median to replace 0s. Below is the comparison of the overall data profile before and after data imputation.

Picture1. Before missing data imputation

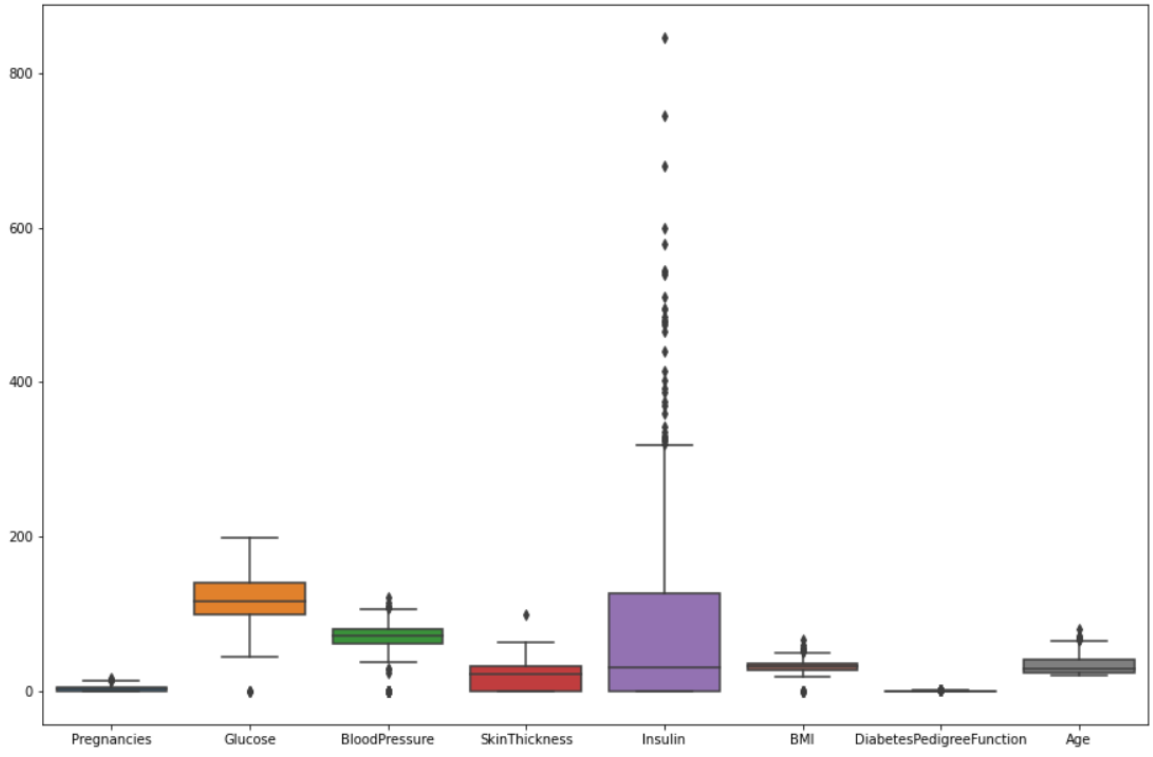
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000

Picture 2. After missing data imputation

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.681605	72.386719	27.334635	94.652344	32.450911	0.471876	33.240885	0.348958
std	3.369578	30.436016	12.096642	9.229014	105.547598	6.875366	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000

2.3.2 Outliers

In the data preprocessing step, I recognized some outliers embedded in the dataset. However, since the size of the dataset is relatively small, I could not justify if outliers are indeed outliers, or these cases were selected for the sake of representation. Thus, I would not perform any treatment for outliers. The below charts show the distribution of the outliers for variables.



2.3.3 Data categorization

Since this research mainly focuses on using ANOVA and ANCOVA methods to test research questions, one condition of using ANOVA was variables must be categorical. Thus, I copied the original dataset and assigned glucose, blood pressure, insulin, and BMI into categories according to their unique severe levels. In general, glucose could be separate into three levels- normal (<=

140mg/dL), intermediate (140-199mg/dL), and high (>=200mg/dL) [6]. Blood pressure could be categorized into four categories based on diastolic- normal (<80), stage1 (80-89), stage2 (90-119), hypertension (>=120) [7]. Insulin could be categorized into three levels- deficiency (<200 poml/L), intermediate (200~600 poml/L), substantial (>=600 poml/L) [8]. BMI could be categorized into four categories- underweight (<18.5), normal (18.5~24.9), overweight (25~29.9), obese (>30) [9]. Lastly, age could be categorized into three categories- youth (15-24), adulthood (25-64), seniority (>=65) [10]. Below is the comparison of variables snapshots before and after categorizing them into relevant categories.

Picture 1.

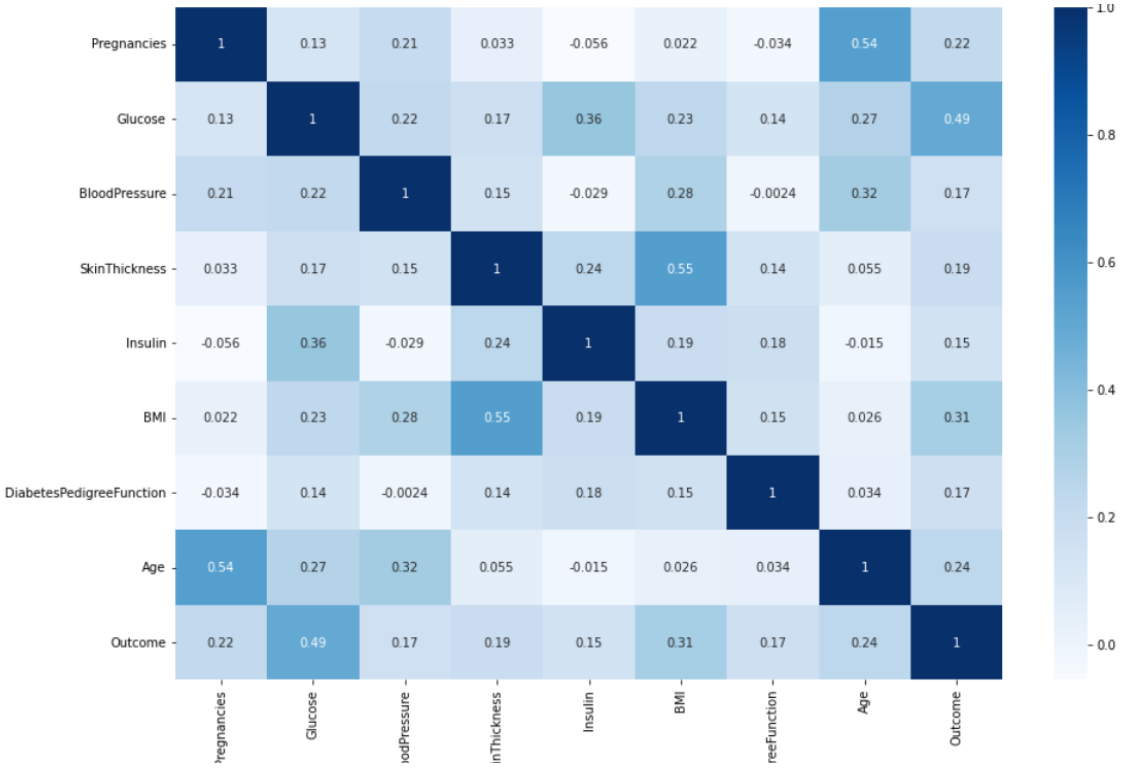
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72	35	30.5	33.6	0.627	50	1
1	1	85.0	66	29	30.5	26.6	0.351	31	0
2	8	183.0	64	23	30.5	23.3	0.672	32	1
3	1	89.0	66	23	94.0	28.1	0.167	21	0
4	0	137.0	40	35	168.0	43.1	2.288	33	1

Picture 2.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	Intermediate	Normal	35	Deficiency	Obese	0.627	Adulthood	1
1	1	Normal	Normal	29	Deficiency	Overweight	0.351	Adulthood	0
2	8	Intermediate	Normal	23	Deficiency	Normal	0.672	Adulthood	1
3	1	Normal	Normal	23	Deficiency	Overweight	0.167	Youth	0
4	0	Normal	Normal	35	Deficiency	Obese	2.288	Adulthood	1

2.3.4 Variables correlations- Heatmap

Before going to the testing directly, I generated a heatmap to illustrate overall data factors' correlations with each other. We could see from the heatmap that glucose and BMI had relatively high correlations with outcomes. Thus, I assumed that in the following test, glucose or BMI plus the other one variable would have the higher chance to successfully predict the diabetes outcome.



3. Parameter Estimates: T-test

Before applying ANOVA and ANCOVA tests, I used the t-test method as a pre-determination test to see if the variables affected the diabetes prediction. If variables affect the diabetes prediction, ANOVA and ANCOVA tests will be implemented to determine the effects level. After conducting the t-test, I concluded that all attributes affect predicting diabetes outcomes. The below chart summarizes the t-test result of attributes.

Attribute	t-test	p-value
Pregnancies	28.47	1.40e-123
Glucose	110.46	0
Blood pressure	164.91	0
Skin thickness	80.92	0
Insulin	24.76	6.32e-100
BMI	129.08	0
Diabetes pedigree function	5.87	5.6e-09
Age	77.45	0

4. One-way ANOVA processes and Post-hoc Tukey Test

One-way ANOVA provided the information used to identify each attribute's strongness level towards the outcome. In this section, I used ANOVA as the first step to test the effectiveness of each attribute. According to the one-way ANOVA definition, attributes have a significant impact on outcomes if the p-value <0.05. After testing five attributes and performing

post-hoc Tukey test, I concluded that all five attributes significantly impact predicting diabetes outcomes.

Attribute 1. Glucose

- One-way ANOVA

Attribute	Sum square	F-value	p-value
Glucose	58.68	154.85	1.66e-32

- Post-hoc test

Group	Coefficient	Standard error	t statistic	Confidence interval	p-value	Effect size
Intermediate-Normal	0.45	0.04	12.44	0.38~0.52	4.972716e-32	1.036987

Attribute 2. Insulin

- One-way ANOVA

Attribute	Sum square	F-value	p-value
Insulin	3.35	7.48	0.00061

- Post-hoc test

Group	Coefficient	Standard error	t statistic	Confidence interval	p-value	Effect size
Deficiency – Intermediate	0.21	0.05	3.87	0.1~0.32	0.000357	0.02

Attribute 3. BMI

- One-way ANOVA

Attribute	Sum square	F-value	p-value
BMI	16.41	26.64	2.19e-16

- Post-hoc test

Group	Coefficient	Standard error	t statistic	Confidence interval	p-value	Effect size
Overweight–Normal	0.15	0.06	2.75	0.04~0.27	2.39e-02	0.34
Obese- Normal	0.38	0.05	7.76	0.29~0.48	1.6e-13	0.85
Obese-Overweight	0.23	0.04	5.77	0.15~0.31	5.92e-08	0.51

Attribute 4. Blood Pressure

- One-way ANOVA

Attribute	Sum square	F-value	p-value

Blood pressure	2.67	3.96	0.01

- Post-hoc test

Group	Coefficient	Standard error	t statistic	Confidence interval	p-value	Effect size
Stage2–Normal	0.18	0.06	2.7	0.05~0.3	0.04	0.37

Attribute 5. Age

- One-way ANOVA

Attribute	Sum square	F-value	p-value
Age	13.72	32.65	2.47e-14

- Post-hoc test

Group	Coefficient	Standard error	t statistic	Confidence interval	p-value	Effect size
Adulthood-Youth	0.3	0.04	8.03	0.22~0.38	1.07e-14	0.64

5. Two-way ANOVA, Two-way ANCOVA processes

In order to find the two best predictors that can predict the potential diabetes patients, I used two-way ANOVA and two-way ANCOVA methods to perform the processes. Due to ANOVA test original limitation (needed to be categorical information), I categorized information based on their unique individual characteristics as mentioned above. After transforming the information, I made a paired-wise two-way ANOVA test among various factors, including glucose, blood pressure, insulin, BMI, and age. Thus, fifteen paired tests have been conducted by two-way ANOVA to find out the best two factors to predict diabetes' existences.

After conducting the total of twenty-five paired with two-way ANOVA and ANCOVA tests, I concluded that seven paired factors had the highest indicator that they may be the great diabetes predictors. They were glucose vs. blood pressure, glucose vs. BMI, and BMI vs. age, pregnancies vs. glucose, skin thickness vs glucose, diabetes pedigree function vs. glucose, diabetes pedigree function vs. insulin. The successful candidates must have a p-value <0.05 to ensure that they had a statistically significant contribution to predicting the outcome. Below are the results of the two-way ANOVA testing of these three groups. Seven of them had a very small p-value to indicate that they may be the great diabetes predictors.

Chart 1. Glucose vs. Blood pressure ANOVA test

Group	sum square error	mean square error	F value	p value
Glucose	29.42	14.71	78.21	1.29e-31
Blood Pressure	0.9	0.3	1.6	1.89e-01
Glucose vs Blood Pressure	1.85	0.31	1.64	1.34e-01

Chart 2. Glucose vs. BMI

Group	sum square error	mean square error	F value	p value
Glucose	29.69	14.85	84.99	5.22e-34
BMI	9.64	3.21	18.39	1.6e-11
Glucose vs BMI	1.66	0.28	1.58	1.49e-01

Chart 3. BMI vs. Age

Group	sum square error	mean square error	F value	p value
BMI	16.41	5.47	28.24	2.63e-17
Age	9.0	4.5	23.24	1.61e-10
BMI vs Age	1.16	0.19	1.0	4.24e-01

Chart 4. Pregnancies vs. Glucose

Group	sum square error	F value	p value
Glucose	52.08	142.46	3.1e-30
Pregnancies	5.29	28.95	9.87e-08

Chart 5. Skin thickness vs Glucose

Group	sum square error	F value	p value
Glucose	52.02	138.89	9.22e-30
Skin Thickness	2.91	15.63	8.42e-05

Chart 6. Diabetes pedigree function vs. Glucose

Group	sum square error	F value	p value
Glucose	54.12	145.62	8.1e-31
Diabetes pedigree function	2.99	16.1	6.6e-05

Chart 7. Diabetes pedigree function vs. Insulin

Group	sum square error	F value	p value
Insulin	2.58	5.93	0.003
Diabetes pedigree function	4.51	20.69	0.000006

Interestingly, one group of variables did not meet the p-value <0.05 criteria. Thus, I concluded that there is no significant statistical evidence that distributions of 'Blood Pressure' and 'Insulin' are different. No further testing had been performed for this group of variables.

Group	sum square error	mean square error	F value	p value
Blood Pressure	2.67	0.89	4.02	0.007477
Insulin	3.0	1.5	6.78	0.001209
Blood Pressure vs Insulin	1.68	0.28	1.27	0.270982

6. Two-way ANOVA, Two-way ANCOVA Post-hoc Tukey Test

However, only looking at the two-way ANOVA and two-way ANCOVA result was not enough because they did not point out which group in each pair made this pair a better predictor than the other groups. Thus, it was essential to apply the Tukey test, a type of ANOVA and ANCOVA post-hoc test, to confirm the importance of the variables. I used the Tukey test as the post-hoc test method because the Tukey test would estimate the variance from the whole dataset rather than be justified from a part of the dataset. The successful candidates must have the p-value <0.05 after the Tukey test to ensure that they had a statistically significant contribution to predicting the outcome. Below information presented the statistical significance difference (p-value < 0.05) of each group based on different effect factors.

Group 1. Glucose vs. Blood pressure

Chart 1.1 Glucose as main effect factor

Group 1	Group 2	Difference	Lower limit	Upper limit	q- value	p-value
Intermediate	Normal	0.46	0.39	0.53	17.85	0.001

Chart 1.2 Blood pressure as main effect factor

Group 1	Group 2	Difference	Lower limit	Upper limit	q- value	p-value
Normal	Stage 2	0.18	0.02	0.34	4.17	0.02
Normal	Stage 1	0.11	0.01	0.21	3.85	0.03

Chart 1.3 both Glucose and Blood pressure were main effect factors

Group 1	Group 2	Difference	Lower	Upper	q-value	p-value
Intermediate, Normal	Normal, Normal	0.49	0.35	0.62	609771	0.001
Intermediate, Normal	Normal, Stage 2	0.42	0.16	0.68	7.04	0.001
Intermediate, Normal	Normal, Stage 1	0.36	0.18	0.54	8.73	0.001
Intermediate, Stage 2	Normal, Normal	0.61	0.33	0.9	9.2	0.001
Intermediate, Stage 2	Normal, Stage 2	0.55	0.18	0.91	6.5	0.001
Intermediate, Stage 2	Normal, Stage 1	0.48	0.18	0.79	6.74	0.001

Intermediate, Stage 2	Normal, Normal	0.4	0.2	0.61	8.53	0.001
Intermediate, Stage 2	Normal, Stage 2	0.34	0.04	0.63	4.83	0.02
Intermediate, Stage 2	Normal, Stage 1	0.28	0.04	0.51	5.06	0.01

Group 2. Glucose vs. BMI

Chart 2.1 Glucose as main effect factor

Group 1	Group 2	Difference	Lower limit	Upper limit	q- value	p-value
Intermediate	Normal	0.46	0.39	0.53	17.76	0.001

Chart 2.2 BMI as main effect factor

Group 1	Group 2	Difference	Lower limit	Upper limit	q- value	p-value
Obese	Overweight	0.23	0.13	0.33	8.53	0.001
Obese	Normal	0.38	0.26	0.51	11.49	0.001
Overweight	Normal	0.15	0.02	0.29	4.07	0.02

Chart 2.3 both Glucose and BMI were main effect factors

Group 1	Group 2	Difference	Lower limit	Upper limit	q- value	p-value
Intermediate, Obese	Intermediate, Overweight	0.36	0.1	0.61	5.95	0.001
Intermediate, Obese	Normal, Obese	0.45	0.32	0.58	14.73	0.001
Intermediate, Obese	Normal, Overweight	0.58	0.42	0.73	16.18	0.001
Intermediate, Obese	Normal, Normal	0.74	0.57	0.92	18.17	0.001
Intermediate, Obese	Normal, Underweight	0.76	0.09	1.43	4.91	0.01
Intermediate, Overweight	Normal, Normal	0.38	0.11	0.65	6.12	0.001
Intermediate, Normal	Normal, Normal	0.48	0.04	0.92	4.69	0.02
Normal, Obese	Normal, Normal	0.29	0.14	0.45	8.09	0.001

Group 3. BMI vs. Age

Chart 3.1 BMI as main effect factor

Group 1	Group 2	Difference	Lower limit	Upper limit	q- value	p-value

Obese	Overweight	0.23	0.13	0.32	8.76	0.001
Obese	Normal	0.38	0.27	0.5	11.80	0.001
Overweight	Normal	0.15	0.02	0.29	4.18	0.02

Chart 3.2 Age as main effect factor

Group 1	Group 2	Difference	Lower limit	Upper limit	q- value	p-value
Adulthood	Youth	0.29	0.21	0.37	11.97	0.001

Chart 3.3 both BMI and Age were main effect factor

Group 1	Group 2	Difference	Lower limit	Upper limit	q- value	p-value
Obese, Adulthood	Obese, Youth	0.28	0.13	0.43	8.58	0.001
Obese, Adulthood	Overweight, Adulthood	0.19	0.04	0.34	5.81	0.002
Obese, Adulthood	Overweight, Youth	0.47	0.28	0.66	11.45	0.001
Obese, Adulthood	Normal, Adulthood	0.41	0.22	0.61	9.82	0.001
Obese, Adulthood	Normal, Youth	0.49	0.27	0.72	10.03	0.001
Overweight, Adulthood	Overweight, Youth	0.28	0.06	0.5	5.97	0.001
Overweight, Adulthood	Normal, Adulthood	0.23	0.003	0.45	4.7	0.04
Overweight, Adulthood	Normal, Youth	0.31	0.05	0.56	5.62	0.004

Group 4. Pregnancies vs. Glucose

Group 1	Group 2	Difference	Lower limit	Upper limit	p- value
Intermediate	Normal	0.93	1.48	0.38	0.0009

Group 5. Skin thickness vs Glucose (Glucose was the main effect factor)

Group 1	Group 2	Difference	Lower limit	Upper limit	p- value
Intermediate	Normal	3.19	1.7	4.69	0.0

Group 6. Diabetes pedigree function vs. Glucose (Glucose was the main effect factor)

Group 1	Group 2	Difference	Lower limit	Upper limit	p- value
Intermediate	Normal	0.08	0.03	0.14	0.0031

Group 7. Diabetes pedigree function vs. Insulin (Insulin was the main effect factor)

Group 1	Group 2	Difference	Lower limit	Upper limit	p- value
Deficiency	Intermediate	0.12	0.03	0.2	0.0061
Deficiency	Substantial	0.59	0.15	1.04	0.0051
Intermediate	Substantial	0.48	0.03	0.93	0.03

7. Power Analysis

Power analysis is a useful tool to determine whether sufficient sample sizes covered the existence of Type I and Type II errors. Type I error means a false positive rate, while Type II error is a false negative rate. In order to validate the research, each factor's individual required sample size should be less than 768 samples, which is the total input data for this research. After conducting the power analysis for the entire dataset, I concluded that the dataset used in this research had sufficient samples to cover Type I and Type II errors. The below chart shows the result of the power analysis.

Attributes	Required sample sizes
Glucose	15.62
Insulin	81.05
Blood pressure	115.76
BMI	16.75
Age	38.74

8. Discussion

I conducted several tests and analyses to find out sets of variables that may be the best indicators to predict diabetes above. Here, I will discuss why these testing and analysis methods were selected and the meaning of the testing and analysis results.

At the beginning of the research, I used the heat map as the initial and simple filter to see the correlation between all variables and prediction outcomes. The result shows that correlations exist. At this stage, I also used a parameter estimation function, mainly a t-test, to confirm the result from the heat map. By performing the t-test, people could tell if the variables significantly impact the prediction outcome. According to the t-test outcome, I concluded that variables significantly impact the prediction outcomes. Although I could tell from the above testing that variables contributed toward the final prediction outcomes, I was not sure of the scale of impact of each variable. Thus, I decided to run the one-way ANOVA test. This test revealed critical information: coefficient, standard error, confidence interval, p-value, and effect size. Therefore, I had the idea of which variables have more impact on the prediction outcomes. Also, as the research topic focuses on finding two variables that contribute the most to predicting outcomes, one-way ANOVA results provided a brief outline telling which variable sets may be the best predictors.

After conducting the above tests, it was time to jump into the research topic- finding the variable sets that could predict diabetes the most. In order to find the result, I used both two-way ANOVA and two-way ANCOVA tests as the primary testing methods. The difference between these two methods is that ANOVA testing focuses on categorical variable sets while ANCOVA testing could work on categorical and numerical variable sets. Some variables were

categorized into different categories with a structured methodology, such as glucose, insulin, BMI, age, and blood pressure. ANOVA tested these variable sets. On the other hand, pregnancies and diabetes pedigree function were variables that could not be categorized easily. Thus, variable sets that contained these two variables were tested by ANCOVA. After conducting ANOVA and ANCOVA tests, I implemented a post-hoc test to figure out further which group in the variable sets performed the best to predict the result. This method could also help us understand why some variables were better than others to predict diabetes outcomes.

Lastly, no matter which testing was used in the research, Type I (false positive) and Type II (false negative) errors were the inevitable side effects. Therefore, I conducted the power analysis to confirm that the number of samples in the dataset provided enough amount to cover Type I and Type II errors. The result showed that the Type I and Type II errors were well covered among all variables. Until this point, I concluded that the research was valid and variable sets were ready to fit into the prediction model.

9. Model fitting and Prediction accuracy

In this research, I decided to use the linear regression model to build the diabetes prediction model because linear regression had (1) solid mathematical history that numerous statisticians have tested without prime controversies and (2) easy to interpret features that could be understood by people who may not have a solid technical or clinical background. Based on the ANOVA, ANCOVA, and post-hoc test result, I concluded that (1) Glucose vs. Blood pressure, (2) Glucose vs. BMI, (3) BMI vs. Age, (4) Pregnancies & Glucose, (5) Skin thickness vs Glucose, (6) Diabetes pedigree function vs. Glucose, (7) Diabetes pedigree function vs. Insulin were the best seven groups of diabetes prediction and applied them into the linear regression model. The below chart illustrates the final prediction score and coefficient of determination. Based on this dataset, glucose, and BMI together are the best predictors to predict the existence of diabetes.

Group	Score	Coefficient of determination	Mean squared error
Glucose & Blood pressure	24.18	0.25	0.16
Glucose & BMI	30.31	0.21	0.18
BMI & Age	15.52	0.12	0.19
Pregnancies & Glucose	25.78	0.3	0.15
Skin thickness vs Glucose	25.01	0.26	0.16
Diabetes pedigree function vs. Glucose	25.13	0.26	0.16
Diabetes pedigree function vs. Insulin	3.85	0.06	0.2

10. Result and Conclusion

The primary purpose of this research was to find out the two best predictors that could best predict the existence of diabetes. In order to avoid squeezing all variables into the linear regression model, which could decrease system efficiency and waste of time, I decided to do the two-way ANOVA and two-way ANCOVA test first, followed by the Tukey HSD post-hoc testing to filter out incapable variables at the initial stage. Before performing the two-way ANOVA test and

two-way ANCOVA, fifteen groups of variables needed to fit simple parameter estimation test. After implementing two-way ANOVA and Tukey HSD post-hoc test, only seven groups of variables required model fitting process to find out the best set. Ultimately, glucose and BMI were the best predictors based on linear regression model accuracy score and coefficient of determination. This process was especially critical and helpful when there were a lot of sets of variables that needed to be tested. Instead of throwing them into the model at once, people could first filter out some incapable variables in a short period and let the linear regression model work on remaining suitable variables.

11. Limitation and future analysis

11.1 Only measurable clinical outcomes were used in the research

In this research, I primarily used measurable clinical outcomes to estimate high diabetes risks patients. However, some critical yet difficult-to-collect factors are not included in the research when building the model and algorithms under this condition. For example, patients' self-reports on eating habits, living styles, and past family genetic diabetes histories were not considered here. Thus, the model introduced in this research may have limited capability in predicting outcomes. Further study with the above details may suggest another more accurate model.

11.2 P-value < 0.05 as statistically significance different criteria

In this research, I primarily used the p-value < 0.05 as a standard threshold to judge the effectiveness of factors because 0.05 was one of the most acceptable criteria used in the experiment world [11]. However, researchers may modify these criteria to meet their experiment's needs. Different p-value levels may lead to slightly different implications at the end. Also, in order to concise this report, I chose the three groups that could potentially build the most accurate model. Testing results for other groups, who did not meet $p<0.05$ criteria, were included in the Python file and this research paper.

11.3 Small size dataset and incorrect data

Speaking of the size of the dataset, this was a relatively small dataset with only 768 anonymous patients' data to build the model compared to actual diabetes patients. In addition, missing data/ incorrect data were identified in the dataset, which accounted for almost one-third of the dataset for some variables. Thus, the predicted model's result may be skewed due to outliers or underrepresented syndrome from patients. People who would like to use these predicting models must adjust them according to experimental needs and unique situations. Researchers may find the demands of acquiring more patient data to make a more comprehensive predicting model further.

Reference:

[1] Diabetes: Types, Risk Factors, Symptoms, Tests, Treatments & Prevention. (2021, March 28). Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview

[2] Diabetes. (2021, November 10). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] Advocacy Reports. (2017). DiabetesCanadaWebsite. http://www.diabetes.ca/en-CA/advocacy---policies/advocacy-reports

[4] Types of Diabetes Mellitus. (2008, August 13). WebMD. https://www.webmd.com/diabetes/guide/types-of-diabetes-mellitus#091e9c5e80238a99-1-2

[5] Kang, H. (2013). The prevention and handling of the missing data. Korean Journal of Anesthesiology, 64(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

[6] Diabetes Testing. (2022, March 2). Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/getting-tested.html

[7] Understanding Blood Pressure Readings. (2022, January 5). Www.Heart.Org. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

[8] J. (2022, February 14). Exeter Clinical Laboratory International. Eexeterlaboratory. https://www.exeterlaboratory.com/test/c-peptide-plasma/

[9] All About Adult BMI. (2022, March 17). Centers for Disease Control and Prevention. https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

[10] Ameen, Z. (2021, February 19). The Ultimate Guide to Facebook Ads (Demographics & Ages Ranges). Canz Marketing. https://www.canzmarketing.com/facebook-demographics-and-age-ranges/

[11] [McLeod, S. A. (1970, January 1). [what a P-value tells you about statistical significance]. P. Retrieved from https://www.simplypsychology.org/p-value.html