Source: https://unsplash.com/photos/XknuBmnjbKg

Design Experiment about Diabetes

Experimental Design for Data Science

Celio Rodrigues de Oliveira
Student ID 1006570284
MI Human-Centred Data Science
2629 words

April 10, 2022

Table of Contents

Introduction

By delivering the right and available facts and performing analytic doings, Machine Learning structures can contribute to making good verdicts. You must guarantee that the dataset is as clean and interpretable using a set of choices that well-matches with the data arrangements used in your secured and embraced capabilities you possess. The output, whether prediction or any other result, offer you direction to the decision-making process. When you make rulings built on statistics from information arrangements, you're reflecting data from your firms' activities. The information recorded at the job level is organized into multiple setups by the strategic structures.

Objective

Many researchers have conducted various studies on this dataset and have gained many insights into areas diabetes prediction. However, the objective of this experiment is not a thoroughly study on the dataset, but it will serve as a good case study for us to apply and learn various machine learning methods on. Although many researchers have annotated this dataset for their research and most of them are available for public use. Not much processing was done by me in order to gain the insights I seek, which are mentioned in coming sections.

Data Structure and Strategy

The Diabetes dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases and it was made public by Google on Kaggle platform in 1990. This current Kaggle dataset with information about female patients at least 21 years old of Pima Indian heritage from Phoenix, Arizona – USA, and contains over 700 instances organized in 9 features as described below:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U / ml)
- BMI: Body mass index (weight in kg / (height in m) ^ 2)
- DiabetesPedigreeFunction: Diabetes pedigree function (the likelihood of genetic diabetes)
- Age: Age (years)
- Outcome: Class variable (0 for non-diabetic or 1 for diabetic)

The output is a diagnostic of diabetes in a binary classification, where 0 = non-diabetic, 1 = diabetic according to the World Health Organization criteria (eg. If the 2-hour post-load plasma glucose was at least 200mg / dl at any examination or if found during medical routinary care).

As it appears on Kaggle, these are the information about the dataset Source:

a) Original owners: National Institute of Diabetes and Digestive and Kidney Diseases

b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu) Research Center, RMI Group Leader - Applied Physics Laboratory, The Johns Hopkins University Johns Hopkins Road - Laurel, MD 20707 (301) 953-6231

c) Date received: 9 May 1990

```
Pregnancies  is a datatype  int64  and ranges from  0  to  17
Glucose  is a datatype  int64  and ranges from  0  to  199
BloodPressure  is a datatype  int64  and ranges from  0  to  122
SkinThickness  is a datatype  int64  and ranges from  0  to  99
Insulin  is a datatype  int64  and ranges from  0  to  846
BMI  is a datatype  float64  and ranges from  0.0  to  67.1
DiabetesPedigreeFunction  is a datatype  float64  and ranges from  0.078  to  2.42
Age  is a datatype  int64  and ranges from  21  to  81
Outcome  is a datatype  int64  and ranges from  0  to  1
```

Figure 1. Data Structure

The dataset is from class pandas.core.frame.DataFrame, with data shape organized in 768 rows and 9 columns. However, Null values are not present in this study, I found some misinformation about the Insulin variable which there were some 0 values. The 0 value in "Insulin" is missing for unknown reason. It is known that insulin level should never be zero. Therefore, I decided to replace 0 by the insulin mean.

```
diabetes['Insulin']= diabetes['Insulin'].replace(0,diabetes['Insulin'].mean()).astype(int)
diabetes.Insulin.value_counts()
```

```
79     376
105     11
130      9
140      9
120      8
```

Figure 2. Replacing 0 values by the mean

Duplications are not present in this dataset and the column names are, as explained earlier: Pregnancies (int), Glucose (int), BloodPressure (int), SkinThickness (int), Insulin (int), BMI (float), DiabetesPedigreeFunction (float), Age (int), and Outcome (int).

Exploratory Data Analysis

From the correlation plot below, we infer that there is ...

- a strong negative correlation between:
  - SkinThickness and Age;
  - Pregnancies, Insulin, BMI and DiabetesPedigreeFunction

- a positive correlation between:
  - Pregnancies and Age;
  - Glucose and Outcome (positive diagnostic for Diabetes)

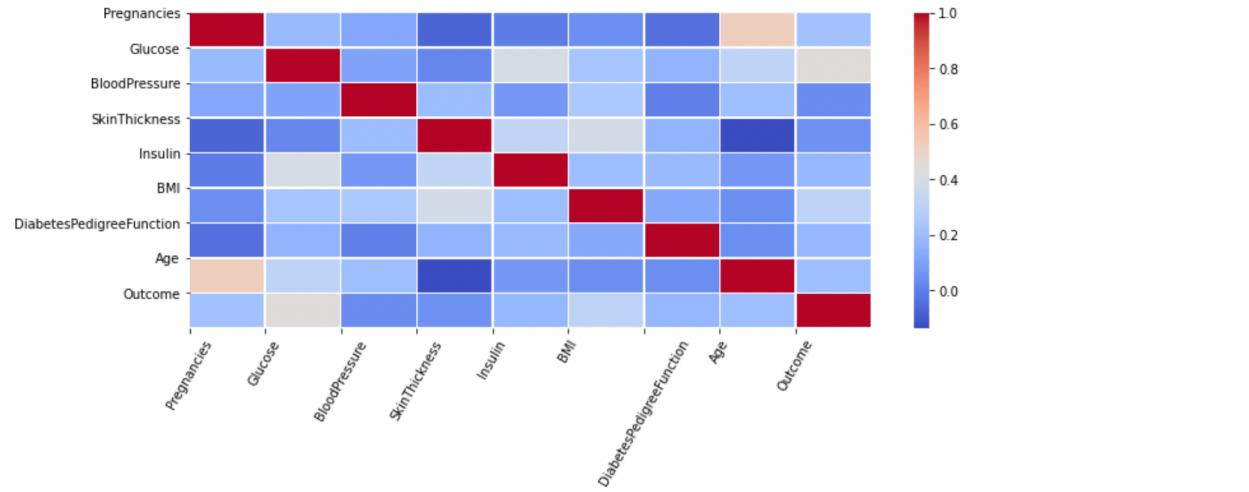

Figure 3. Correlation[1] Analysis

It is important to manage diabetes specially during pregnancies to preserve the health of the mother and also the baby. In opposition, untreated cases lead to depression, high blood pressure, malformation, etc., which increase the risk of pregnancies complications and even loss of a child. A thorough but quick analysis can be performed using a pair-plot as follows on Figure 4.

[1] Correlation is how one or more variables are related to each other. It can also be one of the examples of finding feature importance before start modeling, but I prefer to use a feature importance graph as shown later to express the information gain among variables.

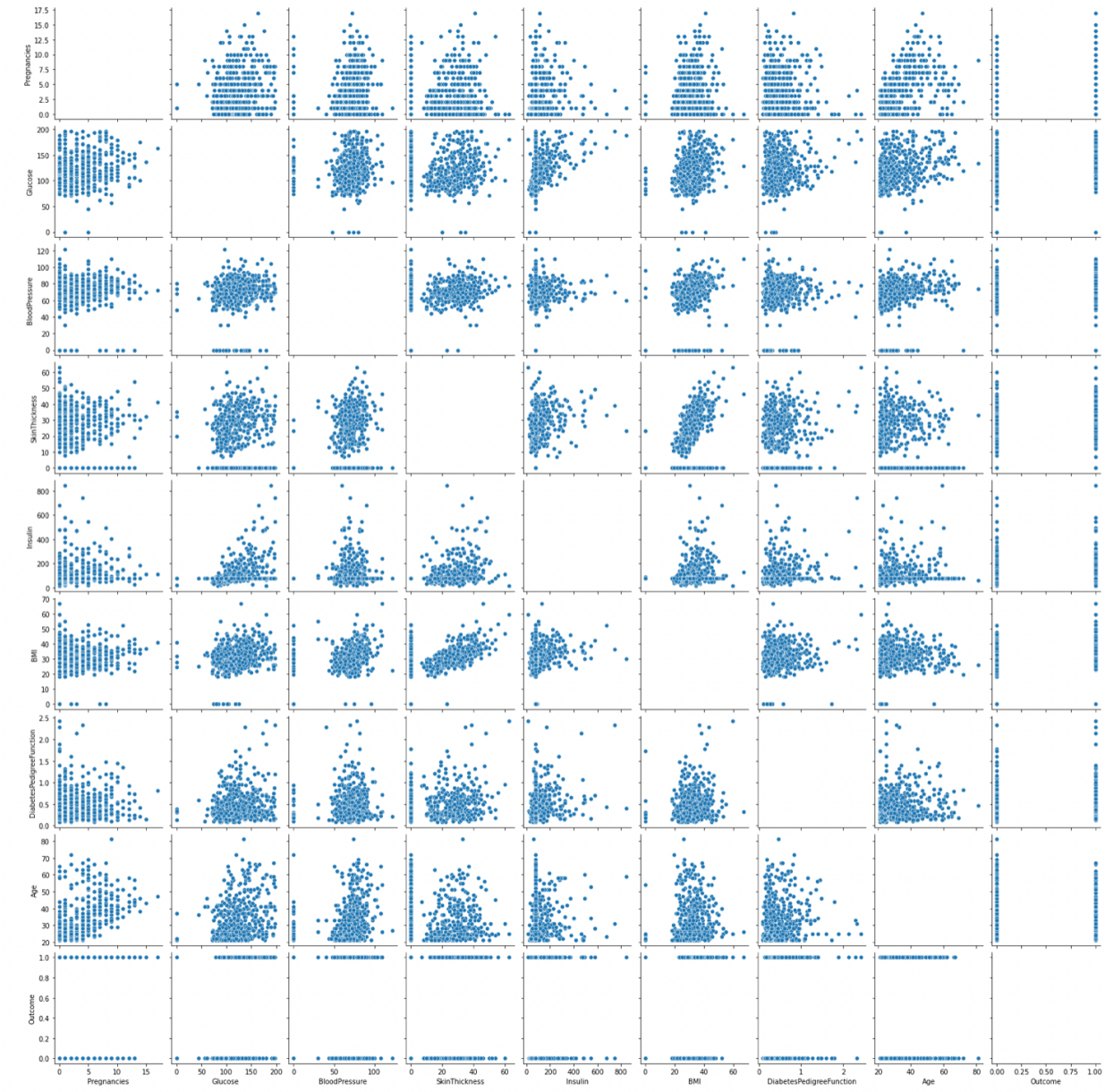

Figure 4. Pair-plot (analysis of linearity / non-linearity)

We do not see clear linear correlation in any of the 9 variables. Therefore, we can foresee that a regression analysis would not enrich our knowledge about this dataset and another type of analysis would be preferred.

Furthermore, I decided to perform a Decision Tree analysis and look for the elbow curve, which enlightened the study that it would need at least 40 splits before reaching a purity level on each node.

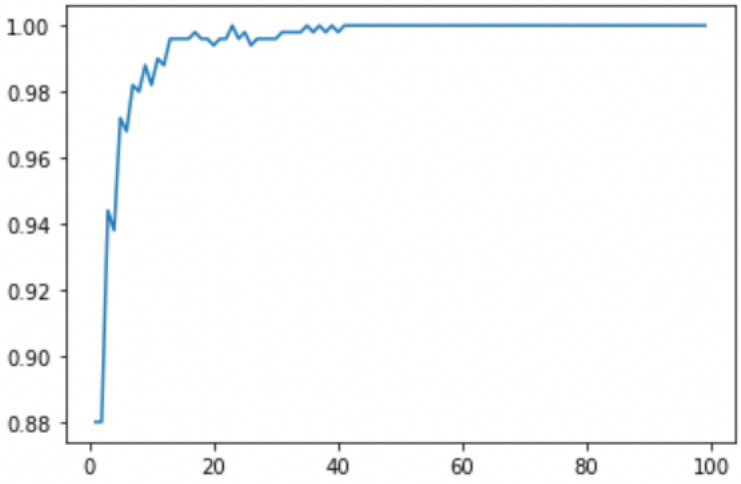

Figure 5. Elbow curve analysis

When deciding about dimensionality reduction for fasten the analysis and optimize the prediction, I tried to perform a density analysis between variables Age and Outcome which was not very representative. I would also say that age variable is unconclusive by its own when predicting diabetes diagnostics as shown below on Figure 6.

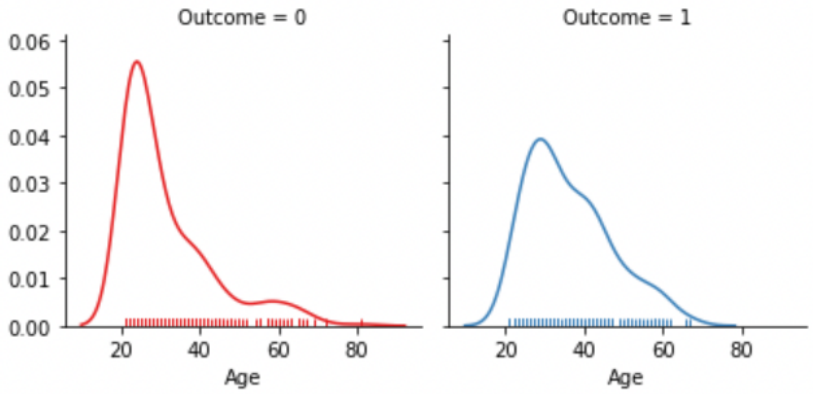

Figure 6. Density analysis (Age and Diabetes propensity)

From the density plot above we can see the non-linear relationship between Age and Outcome (positive diagnostic for diabetes). Therefore, diabetes can be interpreted as either an extrinsic phenomenon or forcefully caused by genetic predisposition. Possible causes might be obesity, inactive lifestyle, pregnancy, environmental factors, etc.

The mean age for this study is 36 years old. Therefore, it is noticeable that the age is not well balanced with a majority of the population between 20 and 30 years old which makes it harder to analyse critical paths for diabetes diagnostics as well as preventive strategies. Moreover, we do not see isolated cases nor outliers what would happen in a dataset better distributed.

According to the magnitude density as shown in the scatter plot below it is clearer the concentration of age by the proximity of the dots. Clearly, the density resides in age group 20-30 years old.
In sequence, I moved to a Feature Importance analysis to calculate the score for all the independent variables in relation to the dependent variable (Outcome -> positive diagnostic for diabetes). The score simply represents the importance of each feature. A higher score means that a specific feature has a higher impact in predicting the outcome over certain variables.

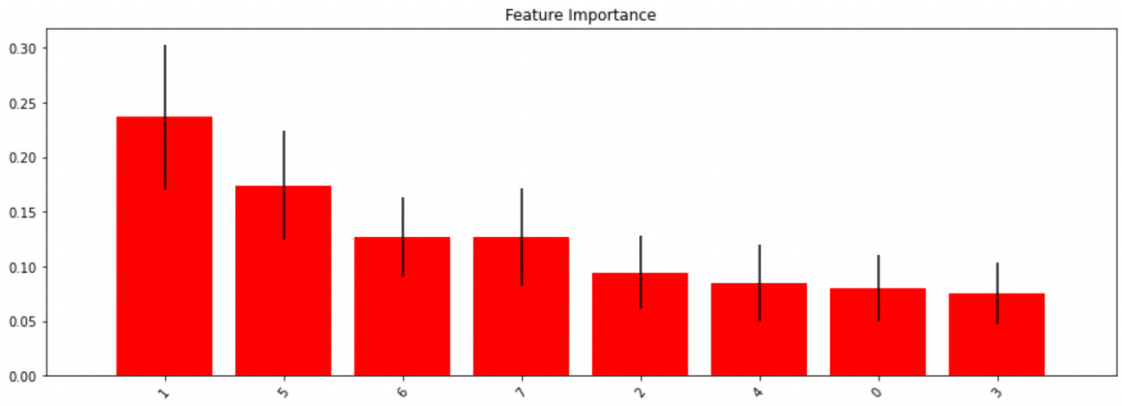

Figure 7. Feature Importance

ANOVA – Analysis of Variance

I then decided to pursue the study of Anova with these two higher impact features (1 = Glucose, 5 = BMI). However, the dataset required a new treatment to cluster these two variables into groups as explained:

- **Blood sugar classification:**
  - less than 140 mg/dL (7.8 mmol/L) is normal
  - between 140 and 199 mg/dL (7.8 mmol/L and 11.0 mmol/L) indicates prediabetes.
  - more than 200 mg/dL (11.1 mmol/L) after two hours indicates diabetes

```
Normal         571
Prediabetes    197
Name: Glucose, dtype: int64
```

Figure 8. Clustered groups Normal, Prediabetes, and Diabetes

BMI -> A the healthy range is 18.5 to 24.9, while a BMI of 25.0 or more is overweight. BMI applies to most adults 18-65 years. According to Diabetes Canada (https://www.diabetes.ca/managing-my-diabetes/tools---resources/body-mass-index-(bmi)-calculator).

- **BMI classification:**
  - 18 or less Underweight
  - 19 - 24 Healthy
  - 25 - 29 Overweight
  - 30+ Obese

```
Obese          472
Overweight     207
Healthy         78
Underweight     11
Name: BMI, dtype: int64
```

Figure 9. Clustered data into Obese, Overweight, Healthy, and Underweight.

Figure 10, below, shows Anova results that can be interpreted by p-values for each of the factors in the output:

- The Glucose p-value is equal to 3.249630e-35
- The BMI p-value is equal to 2.305121e-11
- The Glucose * BMI: p-value is equal to 1.849685e-01

The p-values for BMI turn out to be less than 0.05 which implies that the means of both the factors possess a statistically significant effect on Diabetes. The p-value for the interaction effect (1.849685e-01) is greater than 0.05 which depicts that there is significant interaction effect between Glucose and BMI.

	df	sum_sq	mean_sq	F	PR(>F)
C(Glucose)	1.0	29.970841	29.970841	169.986098	3.249630e-35
C(BMI)	3.0	9.583078	3.194359	18.117498	2.305121e-11
C(Glucose):C(BMI)	3.0	0.853081	0.284360	1.612811	1.849685e-01
Residual	761.0	134.174562	0.176313	NaN	NaN

Figure 10. ANOVA analyses of results

Regression Analysis (Ordinary Least Squares)

OLS is a good substitution of traditional linear regression techniques in analyzing regression models. It is fundamentally comparing differences between individual points of data (Glucose ad BMI) and the prediction (Outcome) finding the best fit line to measure prediction errors also called distance between original and predicted points.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.242
Model:                            OLS   Adj. R-squared:                  0.239
Method:                 Least Squares   F-statistic:                     79.30
Date:                Mon, 28 Mar 2022   Prob (F-statistic):           1.29e-30
Time:                        20:29:48   Log-Likelihood:                -274.44
No. Observations:                 500   AIC:                             554.9
Df Residuals:                     497   BIC:                             567.5
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.7508      0.093     -8.044      0.000      -0.934      -0.567
Glucose        0.0059      0.001      9.804      0.000       0.005       0.007
BMI            0.0126      0.002      5.357      0.000       0.008       0.017
==============================================================================
Omnibus:                       72.183   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.284
Skew:                           0.389   Prob(JB):                     4.37e-07
Kurtosis:                       2.106   Cond. No.                         644.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 11. OLS regression results

The dependent variable is Outcome, the method is OLS, the date and time that this model was created was March 28th, 2022. The number of observations in our dataset is 500 (as I used a subsample for velocity of performance). The degrees of freedom (calculus: number of observations – number of predicting variables – 1), also called DF Residuals is 497. DF Model is the number of independent variables used in this model (Glucose and BMI). The Covariance Type is listed as non-robust (not significant explanation other than a robust covariance is the one that minimizes variables). The R-squared is very important because it measures how much of the independent variables, or features, are explained by changes in our dependent variable, or target. This model explains 24% of the change in the Outcome. Adjusted R-squared analyses the efficacy of multiple dependent variables and it penalizes the R-squared based on the dimension of a dataset.

```
Parameters:  Intercept   -0.750783
Glucose      0.005851
BMI          0.012644
dtype: float64
```

Figure 12. OLS parameters

Another important information is the F-statistics in linear regression models, which compares a linear model against a model that replaces variable effects to zero. Therefore, it is important to use additional information to interpret this number such as alpha and F-table. Prob F-statistics describes the accuracy of the Null Hypothesis, or if the effect is zero. The chance of this happens in this analysis is shown 1.29e-30 or very unlikely. Log-likelihood is used to compare AIC and BIC or efficacy using penalty systems used for feature selection.

Further tunning in model performance would be performed using hyperparameter tunning techniques and also transforming the dataset through feature engineering techniques. The research is essential because it gives a chance for future research that will help in acquiring more knowledge on the problem we would like to solve by analysing this case-study. Exhaustive tests

are needed to be more conclusive in predicting diabetes diagnostics with better ML algorithms, which is not yet object of this study.

ANCOVA

The analysis of covariance (ANCOVA) is used when determining whether or not there is a statistically significant difference between means of independent groups after controlling one or more covariates[2]. Ancova has three main characteristics:

1. It removes the statistical power and reduces error term by removing the variance associated with covariates,
2. It gives adjusted means for each group of a categorical variable and removes the covariates' bias from the model,
3. It uses an approach like multiple regression to study adjusted effects of independent variables on a dependent variable.

I wanted to estimate the differences between groups of the independent categorical variable glucose (primary interest) would have an impact of Diabetic outcome, by statistically adjusting the effect of a covariate BMI of each individual. For the purpose of this study, I used a subsample with 500 trials for each of the three characteristics, as described:

- Factor variable: Glucose (Normal, Prediabetes, Diabetes) -> categorical
- Covariate: BMI (Body mass index of each subject) -> continuous
- Response variable: Outcome (1 = diabetic, 0 = non-diabetic) -> continuous

	Glucose	BMI	Outcome
0	Prediabetes	33.6	1
1	Normal	26.6	0
2	Prediabetes	23.3	1
3	Normal	28.1	0

Figure 13. Ancova subsample dataset

[2] Covariates are characteristic excluding the actual treatment of participants in an experiment and it can be an independent variable (e.g. direct interest) or an unwanted (e.g. non-interest in a study). Covariates are also known as control, concomitant, or confounding variables and they might affect the responsible variable, in some cases.

Assumptions

At each level of a categorical independent variable, the covariate BMI linearly related to the outcome, on that case the relationship was linear. Therefore no adjustment needed to be made to the covariate and bias interference was preserved as we can see o figure 13 below.

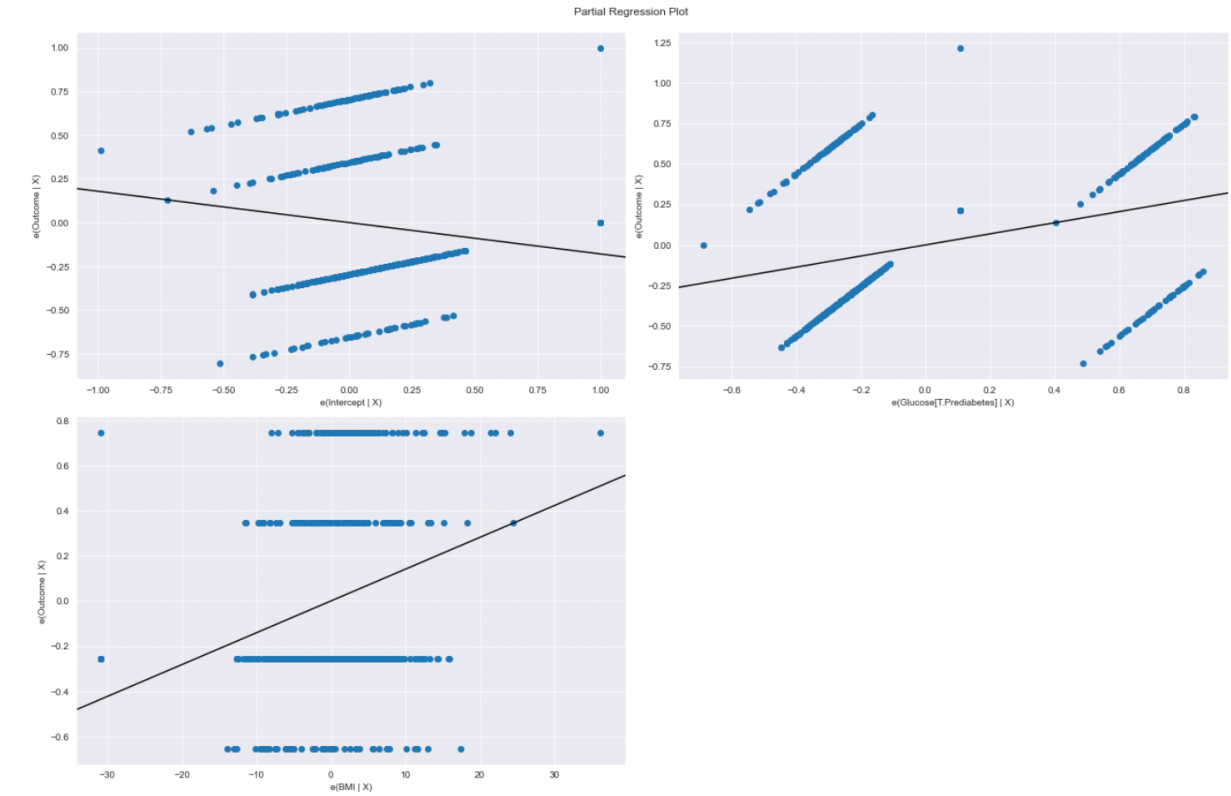

Figure 14. ANCOVA's partition regression grid

The graph shows a dependency between the Outcome (Diabetes) and BMI and also between the Outcome and Glucose, but when these two variables are combined (intercept), the dependency decreases drastically and even invert the direction showing that Diabetes is inversely proportional to the combination between Glucose and BMI, indicating that the dependent variable is not directly dependent on the two independent variables on this example.
Moreover, the interaction between glucose and BMI was non-inexistent and the regression line between BMI and outcome for each group was almost parallel reflecting the same slope. In addition there were no error on the measurement of the covariate BMI as we are using the original measured results from lab tests. (Note that BMI and Outcome were measured on a continuous scale as required for ANCOVA experiments).

Ancova Hypothesis

Null hypothesis: The means of all glucose levels are equal after controlling the effect on BMI i.e. adjusted means are equal.

$$H0: \mu1 = \mu2 = \ldots = \mu p$$

Alternative hypothesis: at least one glucose mean is different from another glucose level after controlling the effect of BMI i.e. adjusted means are not equal.

$$H1: \text{All } \mu \text{ are not equal}$$

Ancova Analysis

I performed a One-way Ancova analysis twice to check the power of different algorithms:

1. Using Ordinary Least Squares (OLS) as a blend between ANOVA and linear regression from General Linear Model (GLM) as proposed by Ronald A. Fisher (1930),
2. Using pingouin.ancova package with one covariate which suits perfectly to this experiment.

```
# ANCOVA from ols
ols_ancova = smf.ols(formula='Outcome ~ Glucose + BMI', data=ancova_diabetes, subset=None, drop_cols=None)
ancova_outcome = ols_ancova.fit()
aov = sm.stats.anova_lm(ancova_outcome, typ=2)
print(aov)
```

```
              sum_sq     df          F        PR(>F)
Glucose    11.059952    1.0  58.687888  9.742684e-14
BMI         6.334855    1.0  33.614909  1.195372e-08
Residual   93.661509  497.0        NaN           NaN
```

Figure 15. #1 Ancova Analysis for hypothesis testing using statsmodel

From both ANCOVA tables, we see that the p-value (p-unc = "uncorrected p-value") for Glucose is 9.742684e-14. Therefore, there are significant differences in outcome means $p < 0.001$ (9.742684e-14) among glucose whilst adjusting the effect of BMI. Since this value is less than 0.05, we can reject the null hypothesis that all glucose levels lead to diabetes and also for BMI that not all higher BMI values lead to diabetic patients. Furthermore, the covariate BMI is significant $p < 0.001$ (1.195372e-08) suggesting it is an important predictor of a diabetic outcome.

```
# ANCOVA from pingouin
ancova(data=ancova_diabetes, dv='Outcome', covar='BMI', between='Glucose')
```

	Source	SS	DF	F	p-unc	np2
0	Glucose	11.059952	1	58.687888	9.742684e-14	0.105613
1	BMI	6.334855	1	33.614909	1.195372e-08	0.063351
2	Residual	93.661509	497	NaN	NaN	NaN

Figure 16. #2 Ancova Analysis for hypothesis testing using pingouin

Conclusion

This study offers thoughts on algorithmic selection and multiple techniques to explore and analyse the diabetes dataset in decision-making. Further study is required to determine the specific benefits received by combining technologies, techniques, and comprehend how AI may be enhanced further medical studies becomes more widely available in terms of capacity, diversity, and velocity. These seemingly diverse activities and complementary elements of a more excellent painting of this picture. In the current competitive world, Machine Learning is essential.

References

Akturk, M. Diabetes Dataset. (2020). Kaggle https://www.kaggle.com/datasets/mathchi/diabetes-data-set?select=diabetes.csv (Accessed on March 28th, 2022)

Ball, P. "Violence in Blue." (2020). *Granta*

Bedre, R. ANCOVA using R and Python (with examples and code). (2022). Data Science Blog. https://www.reneshbedre.com/blog/ancova.html (Accessed on April 10th, 2022)

Czarniak, E., Carter, N., Juneja, K. How to perform an analysis of covariance (ANCOVA) (in Python, using pingouin). (2021). Bentley University. https://nathancarter.github.io/how2data/site/how-to-perform-an-analysis-of-covariance-ancova-in-python-using-pingouin/ (Accessed on April 10th, 2022)

Geeks for Geeks. How to perform a two-way ANOVA in Python. (2021). https://www.geeksforgeeks.org/how-to-perform-a-two-way-anova-in-python/ (Accessed on March 28th, 2022)

Glen, S. Covariate Definition is Statistics. (2015). Statistics How To. https://www.statisticshowto.com/covariate/ (Accessed on April 10th, 2022)

Joshi, G. Diabetes Prediction using Machine Lerning – Python. (2021). https://medium.com/geekculture/diabetes-prediction-using-machine-learning-python-23fc98125d8 (Accessed on March 28th, 2022)

McAleer, T. Interpreting Linear Regression Through statsmodels.summary(). (2020). https://medium.com/swlh/interpreting-linear-regression-through-statsmodels-summary-4796d359035a (Accessed on March 28th, 2022)

Onuoha, Mimi. “When Proof is Not Enough. (2020).” FiveThirtyEight

Seltman, H. J. Experimental design, and analysis. (2018). *Department of Statistics at Carnegie Mellon*