Chronic Disease Prevention Program

Alona Kryshchenko, Cynthia Flores, Terrance Barroso, Antonio Hernandez,
Nathalie Huerta, Angel Mora-Larscheid

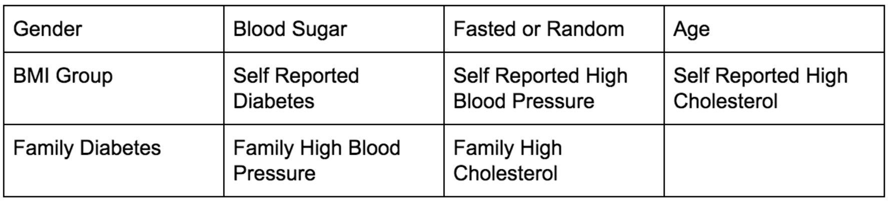
1 Abstract

As our population continues to grow, health professionals in the U.S. have a growing concern for the current and future population related to diabetes mellitus. Diabetes is an underlying disease that occurs when one's blood sugar level is too high for a prolonged period of time.(1) When untreated, short-term and long-term effects are detrimental. Acute complications include: "diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death." (3) Moreover, the long-term effects include: "cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes."

Diabetes is a growing epidemic causing health professionals to research prevention methods as well as a way to diagnosis patients based on certain characteristics. As a result, the Chronic Disease Prevention Program (CDPP) provides blood sugar testing in a non-traditional setting (e.g. grocery stores, libraries, etc.). By using the CDPP data set and applying the tools of machine learning we will predict whether someone is diabetic or requires additional testing. Machine learning is a way to develop algorithms, allowing the computers to learn. The attributes that will be analyzed in the data set are: BMI group, age, gender, blood sugar, self diabetes, and whether the testing was done during fasting or randomly. These attributes were analyzed using Linear Regression to learn more about the relationship between the response variable (i.e. blood sugar) and the explanatory variable. Besides applying Linear Regression, we used Multiple Linear Regression as well a K-Nearest Neighbors, and Decision Tree.

2 Data Description

The data was obtained from Ventura County Health System, specifically from the Chronic Disease Prevention Program (CDPP). The goal is to "link people to care for diabetes that might not otherwise have a usual source of care". The data was collected by testing blood sugar levels in "non-traditional" venues such as grocery stores, libraries, or community centers. The CDPP data set consists of 11 attributes shown below.



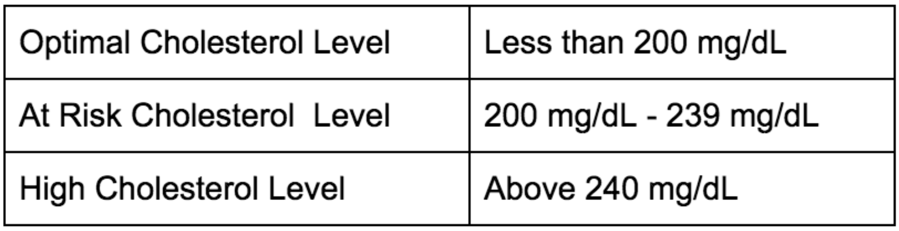
Gender	Blood Sugar	Fasted or Random	Age
BMI Group	Self Reported Diabetes	Self Reported High Blood Pressure	Self Reported High Cholesterol
Family Diabetes	Family High Blood Pressure	Family High Cholesterol	

The dataset contains several binary variables, namely: Fasted or Random, Self Reported Diabetes, Family High Cholesterol, Self Reported High Cholesterol, Family Diabetes, Family High Blood Pressure, and Family High Cholesterol.

The Fasting or Random variable comes from testing the blood sugar of each person. It is assumed that "Fasting" was defined in the data as not having eaten in at least six hours. If this held true for someone being tested then they would be assigned the "Fasting" value, otherwise they were assigned "Random". This is important to know because depending on what level the blood sugar is at for each respective test, there is a threshold for which one's blood sugar level should be at before concerns for diabetes should be considered. We also know if the person has reported to being diabetic, so using this we can re-evaluate their blood sugar levels and determine whether or not they should be tested again, or if testing for diabetes should be a concern for them at all.

For each of the "Self Reported" attributes, it is assumed that if they reported "TRUE" that they have been diagnosed by a medical professional. Similarly, for the familial conditions, it is assumed that they have a relative who was diagnosed by a medical professional for each respective condition.

There are two types of cholesterol, LDL and HDL. Too much of LDL type cholesterol is considered bad for the heart, whereas more HDL is considered healthy (heart.org). For the two cholesterol variables in the dataset, it is assumed that they refer to the total cholesterol score rather than just LDL level. In this case, cholesterol levels are broken into three categories:

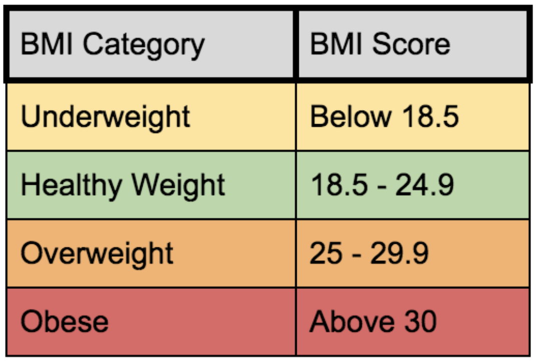


Optimal Cholesterol Level	Less than 200 mg/dL
At Risk Cholesterol Level	200 mg/dL - 239 mg/dL
High Cholesterol Level	Above 240 mg/dL

(healthywomen.org)

Since the variables are binary and not each person's actual cholesterol score, we assume that for each TRUE value the person has a cholesterol of at least 240 mg/dL.

The Body Mass Index (BMI) attribute included in the set is defined as the ratio of one's weight to one's height squared multiplied by a conversion factor of 703 (cdc.gov). For people 20 years old and up, there are four main BMI categories:



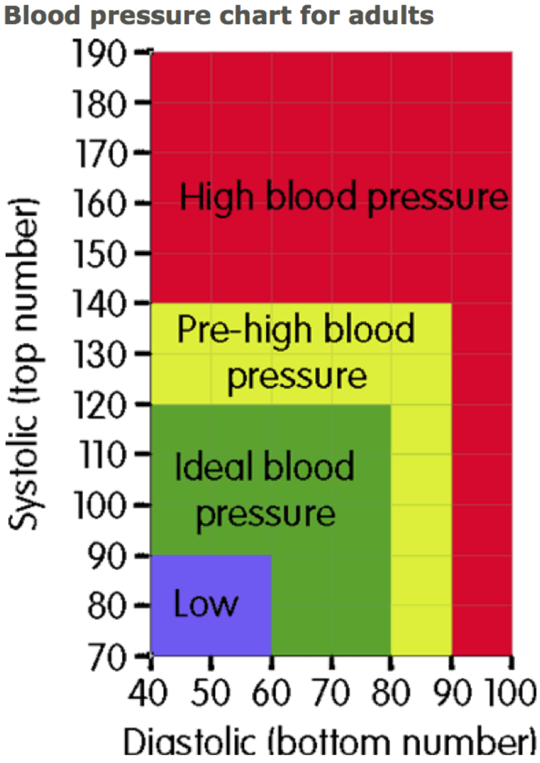
BMI Category	BMI Score
Underweight	Below 18.5
Healthy Weight	18.5 - 24.9
Overweight	25 - 29.9
Obese	Above 30

(cdc.gov)

Since the BMI variable in our data is already split up into these categories, these scores have already been calculated and each observation classified. However, this variable also includes Morbidly Obese as a potential BMI category. It is assumed that for this variable, the morbidly obese category has been defined by the CDPP and they have properly classified those observations.

An example of a BMI classification would be someone who is 5 feet and 9 inches tall and weighs 170 lbs. Using the definition of BMI calculation mentioned above, they would have a BMI of 25.1. This person would be just on the cusp of the overweight category. In order for them to remain in the range of a healthy BMI category, they would need to be anywhere from 125 lbs to 169 lbs (cdc.gov).

A binary variable is observed when a respective individual has high blood pressure versus not having it at all. The graph below explains when one is considered to have high blood pressure in general.



Blood pressure is expressed by two measurements, the systolic and diastolic pressures, which are the maximum and minimum pressures, respectively. For most adults, normal blood pressure at rest is within the range of 100–130 millimeters mercury (mmHg) systolic and 60–80 mmHg diastolic. On the other hand, high blood pressure is present if the resting blood pressure is persistently at or above 130/90 or 140/90 mmHg. Different numbers apply to children. Ambulatory blood pressure monitoring over a 24-hour period appears more accurate than office-based blood pressure measurement.

In our data set we have two blood presure vairables: Self reported, and Family. These can show us how genetics play a roll in having high blood sugar and thus a connection to diabetes.

3 Descriptive Analysis

In the table of summary statistics below we observe our given data including gender, age, blood sugar level, BMI grouping, whether they were randomly tested or fasted, whether they self reported diabetic, having high blood pressure, and high cholesterol, also if they have a family history of the those same variables. Overall our data set contained 454 participants, but for our work we excluded individuals with miss information such as age. There is a higher number of female than males with the exact amount being 121 males and 313 females. Also we see that most of the individuals are in the overweight and obese BMI groups , and that most of the individuals were randomly tested for their blood sugar level instead of being fasted.

As stated above some participants were excluded, specifically the participants missing information such as age. The reason which these participants were dropped was because they were outliers. Since we used common statistical procedures, like linear regression, which are based on parametric statistics, like means or standard deviations, an outlier that is due to incorrectly entered or measured date would be sensitive to these statistics and would compromise the analysis of a linear regression or another statistical procedure.

For the use of understanding what was done and used in our analysis it would be use to know what the parametric statistics.

- **Mean** would be the average or sum of values divided by the number of values.
  - $\hat{x} = \frac{x_1+x_2+...+x_n}{n}$

- **StandardDeviation** is a measure that is used to quantify the amount of variation of a set of data values.
  - std $= \sqrt{\frac{\sum_{i=1}^{N}(x_i - x)^2}{N-1}}$

As seen in the table below the low weight BMI group is of minimal value and can be recognized as an outlier. The low weight BMI group is seen to have a mean value of .002 and is of little significance to our results. Even though generally dropping data is not a good idea for analysis we recognized that this low Weight outlier would create a significant association.

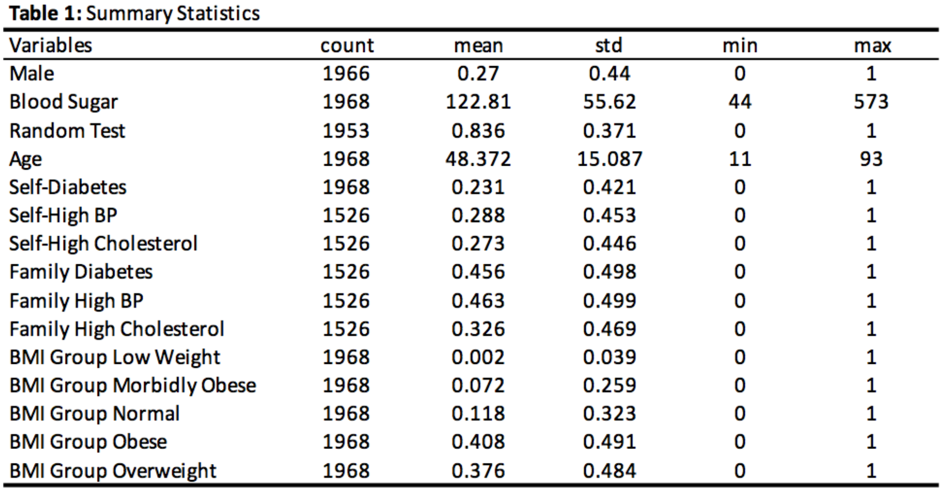


Table 1: Summary Statistics

Variables	count	mean	std	min	max
Male	1966	0.27	0.44	0	1
Blood Sugar	1968	122.81	55.62	44	573
Random Test	1953	0.836	0.371	0	1
Age	1968	48.372	15.087	11	93
Self-Diabetes	1968	0.231	0.421	0	1
Self-High BP	1526	0.288	0.453	0	1
Self-High Cholesterol	1526	0.273	0.446	0	1
Family Diabetes	1526	0.456	0.498	0	1
Family High BP	1526	0.463	0.499	0	1
Family High Cholesterol	1526	0.326	0.469	0	1
BMI Group Low Weight	1968	0.002	0.039	0	1
BMI Group Morbidly Obese	1968	0.072	0.259	0	1
BMI Group Normal	1968	0.118	0.323	0	1
BMI Group Obese	1968	0.408	0.491	0	1
BMI Group Overweight	1968	0.376	0.484	0	1

Summary statistics are broken down in table above for all aforementioned variables. Although total observations equal out to approximately 3000, with a lot of missing data leads to a variable's respective value as shown in the table. Additionally, BMI groups are broken down into a respective binary for each class. The mean for each of these can be interpreted as the total proportion of a respective class occurring.

Additionally, for the continuous variables, blood sugar and age, values above 29 and 0 respective were kept in this dataset. However, with a max of blood sugar of gave possible concern for outliers in general. Moreover, highest blood sugar level ever recorded was approximately 2600 mg, thus indicating slim probability that the max value was correctly recorded.

Referencing the multivariate linear regression results table and the imputation linear regression results table in the results section, we see that age, male gender, randomly tested for blood sugar, self diagnosed diabetic, morbidly obese, and Family history of diabetes populations have higher blood sugar level.

4 Inferential Analysis

When testing for blood sugar, a patient who is randomly tested will result in higher total blood sugar levels compared to someone who fasted prior to the test. The dataset contains the variable, Self-Diabetic- the patient diagnosis himself/herself as being being diabetic. If the participant identifies as Self- Diabetic (i.e. response is true), then we should expect a positively statistically significant relationship with blood sugar levels. This may result from the respective individual already having been diagnosed with diabetes, prior to this testing, resulting in higher blood sugar levels in general.Therefore, we are looking to obtain inference with the remaining aforementioned variables, with the following hypothesis as our guide:

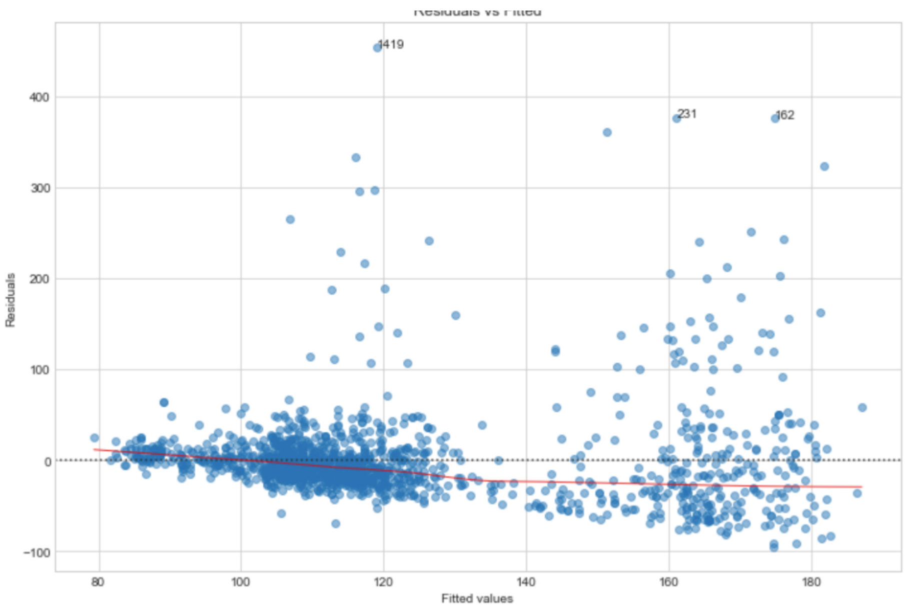
- There is no relationship between age, gender, a respective BMI group, family history of diabetes, high blood pressure, and high cholesterol respectively; and self-diagnosis of high blood pressure and high cholesterol respectively with blood sugar.

- There is a relationship between our mentioned predictors with blood sugar.

The conventional estimation of ordinary least squares (OLS) is used for this empirical analysis with the following equation:

$$\begin{aligned} BSL = \hat{\beta}_0 + \hat{\beta}_1 AGE + \hat{\beta}_2 BMI + \hat{\beta}_3 GEN + \hat{\beta}_4 FA + \hat{\beta}_5 DIA \\ + \hat{\beta}_6 HBP + \hat{\beta}_7 HC + \hat{\beta}_8 FAMD + \hat{\beta}_9 FHC + \hat{\beta}_{10} FHBP + \epsilon \end{aligned} \tag{1}$$

As seen in the equation above we applied the natural log transformation to our response variable. Insight of this transformation was gained from the following graph below.



As seen in the figure above, non-constant variance results on the right-hand side of the figure. In comparison to the left-hand side, we can see a clustered group of observations close to 0 indicating overall good model fit. However, when blood sugar gets larger our model accuracy struggles to capture the predictors relationship with the response. Hence, this is known as heteroscedasticity, and common practice to fix this issue is with natural log or square root transformation of the response variable. This fixed issue was confirmed with adjusted R-squared increasing in value of approximately 6% with the transformation compared to without. Additionally, this provides evidence of non-linearity of our relationship in general, and that nonparametric modeling may result in more sufficient results. However, linear regression is a great tool for interpretation of the relationship between our predictors and response, thus we move forward.

5 Methods

5.1 Feature Engineering

Feature engineering is a vital task in preparing data for machine learning. It "involves the application of transformation functions such as arithmetic and aggregate operators on given features to generate new ones" (Nargesian, Samulowitz, Khurana, Khalil, & Turaga).

It is mathematically manipulating the given attributes in order to create new attributes. The attributes that are developed by feature engineering will lead to an improved prediction model (Nargesian, et al.).

5.2 Principle Component Analysis

Principal Component Analysis (PCA) is procedure that reduces the dimension of the data set, revealing the most important attributes for classification. The first step in PCA is creating the covariance matrix. The covariance matrix is a vital step in PCA because it used to find the redundancy within the attributes given in your data set.

5.2.1 Constructing the Covariance Matrix

Let us consider two observations x_1 and x_2. So, $x_2 = [\, a_1 \quad a_2 \quad \cdots \quad a_m \,]$ and $x_2 = [\, b_1 \quad b_2 \quad \cdots \quad b_m \,]$. Now, the covariance can be re-expressed as a dot product computation where the mean is assumed to be zero for geometric understanding or visualization. Hence,

$$\sigma^2_{x_1x_2} \equiv \frac{1}{n} x_1 x_2^T.$$

Now, let us consider the entire data set. With two observations, the covariance was found by taking the dot product of the two row vectors. By using the data matrix, defined as $\mathbf{X}$, where $\mathbf{X}$ is an $m\times$ n :

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{bmatrix}$$

Hence, the definition of the covariance matrix $\mathbf{C_X}$is:

$$\mathbf{C_X} \equiv \frac{\mathbf{1}}{\mathbf{n}}\mathbf{XX^T}.$$

Consider the matrix $\mathbf{C_X} \equiv \frac{\mathbf{1}}{\mathbf{n}}\mathbf{XX^T}$. The ij^{th} element $\mathbf{C_X}$ will be the dot product between the vector of the i^{th} measurement type with the vectors of the j^{th} measurement type. The properties of the covariance matrix $\mathbf{C_X}$ are:

- $\mathbf{C_X}$ is a square symmetric $m \times m$ matrix.
- The main diagonal terms are the variance.
- The off diagonal terms are the covariance.

5.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a machine learning algorithm, widely used in classification problems, but can also be used in regression. Figure 4 is a visual representation of KNN. In Figure 4, one uses KNN to classify the green dot with two different k values, k=3 and k=5. When k=3, the algorithm looks at the three closest data points to the green dot and classifies it based on what is the most common shapes in that area. The green dot will be classified as a red triangle because there are 2 red triangles versus 1 blue square. Looking at the outer circle with 5 neighbors, i.e. k=5, the green dot is classified as a blue square since there are 3 blue squares versus 2 red triangles. Although this visual makes KNN seem simple and it is regarded as the simplest classification algorithm, there are things one must consider. When implementing KNN one must consider the amount of neighbors as well as the metric; in other words: how distance is defined in the algorithm. The three metrics we used are as followed:

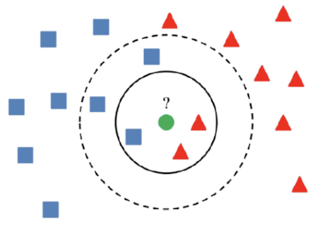
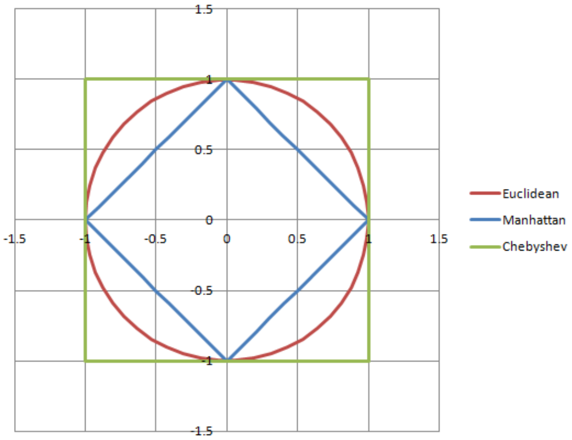


Figure 1: Classifying Using K-Nearest Neighbors (Nathalie Huerta, Angel Mora)

- Euclidean Metric
- Manhattan Metric
- Chebyshev



Euclidean metric is the default metric that KNN follows. It is defined when given points, for example (x_1, y_1) and (x_2, y_2), as

$$D_e = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

We also tried using the two other metrics listed above; the Taxicab metric is used when we define $k = 1$. Taxicab metric is defined to be the sum of the distance of points. For example, given two points (x_1, y_1) and (x_2, y_2), the distance between them is

$$D_t = |x_1 - x_2| + |y_1 - y_2|.$$

The Max metric, or the Chebyshev metric, is the defined when given two points p_i and q_i the distance between them is

$$D_c := \max(|p_i - q_i|).$$

As previously mentioned, the metric used and how we define k are important to consider when implementing KNN because if we do not know how distance, or "near"-ness is defined in our feature space then we might run into comprehension errors. We might not understand why new data is being classified a certain way when a different classification might seem more intuitively appropriate, and knowing how distance is defined for our algorithm could potentially help remedy any mistakes or misunderstandings. Typically this happens in higher dimensions, or when there are many input variables. This is because it is possible for points to have similar attributes and also be far away from each other geometrically. This has been called the "Curse of Dimensionality".

5.4 Multiple Linear Regression

Multiple linear regression models the relationship between 2 or more input, or predictor, variables and a response variable. To do this, we find a line that best fits the data. The equation comes in the form :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p + \epsilon. \tag{2}$$

Where the x_is are the predictor variables and each β_i represents the regression coefficients, i.e. how much association its corresponding x_i has with the response variable y and ϵ is the error term. Then using our variables we have

$$\begin{aligned} BSL = \hat{\beta}_0 + \hat{\beta}_1 AGE + \hat{\beta}_2 BMI + \hat{\beta}_3 GEN + \hat{\beta}_4 FA + \hat{\beta}_5 DIA \\ + \hat{\beta}_6 HBP + \hat{\beta}_7 HC + \hat{\beta}_8 FAMD + \hat{\beta}_9 FHC + \hat{\beta}_{10} FHBP + \epsilon. \end{aligned} \tag{3}$$

In linear regression, each beta value is chosen to reduce the Sum of Residual Squared (RSS) which is the distance between each observation and the line that is fitting the data, i.e. the error in our linear estimation. In Simple Linear regression there are only two coefficients and the lines are of the slope-intercept form. The coefficients represent the average effect that x_i has on our response variable while not taking into consideration the other input variables. Multiple Linear Regression coefficients represent the average effect that its respective x_i has on the response variable while holding the other variables constant (An Introduction to Statistical Learning).

Finding these coefficients for the simple linear regression method is motivated by minimizing the RSS and we have only two coefficients to solve for, β_0 and β_1 which are solved using the following:

$$\beta_1 = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{p}(x_i - \bar{x})^2},$$

$$\beta_0 = \bar{y} - \beta_1\bar{x},$$

where $\bar{x}$ and $\bar{y}$ are the means (An Introduction to Statistical Learning).

When calculating the coefficients for multiple linear regression we use vectors to store all the values. The matrix of the predictor values is represented by

$$X = \begin{pmatrix} 1 & X_{12} & X_{13} & ... & X_{1p} \\ 1 & X_{22} & X_{23} & ... & X_{2p} \\ & & ... & & \\ 1 & X_{n2} & X_{n3} & ... & X_{np} \end{pmatrix},$$

the response vector is

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ ... \\ Y_n \end{pmatrix},$$

the coefficients vector, or slope vector is

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ ... \\ \beta_n \end{pmatrix},$$

and the error vector is

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ ... \\ \epsilon_n \end{pmatrix}.$$

Using this notation we can rewrite equation (2) as something analagous to the formula of a line,

$$Y = X\beta + \epsilon \tag{4}$$

where the variables are put into their respective vector representations. Still, to find the coefficients we follow the same approach as in simple linear regression by reducing the RSS which is found by minimizing the norm of the error term, or in this case the error vector ϵ. From equation (4) we see

$$||\epsilon|| = ||Y - X\beta||.$$

Then minimizing the RSS would be to minimize

$$\sum_{i=1}^{N} ||\epsilon||^2 = \sum_{i=1}^{N} ||Y - X\beta||^2.$$

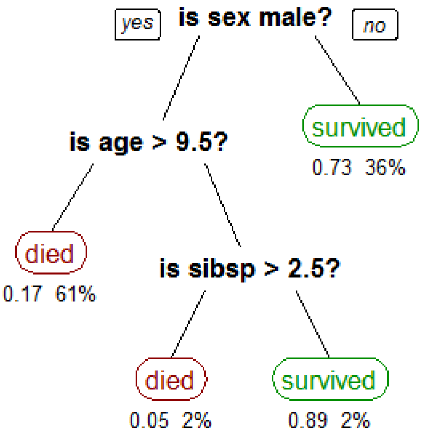
We find that the β vector that minimizes this equaiton is found by

$$\beta = (X'X)^{-1}(X'Y).$$

Once we have the coefficients for each variable X_i we have a full linear approximation of the data and thus can estimate how each predictor variable in our data set effects the response variable and the relationship between each predictor variable (Linear Regression Analysis).

5.5 Decision Tree

Decision Tree is a machine learning algorithm that partitions the data into subsets. It is used in classification problems as well as in regression. They work in a flow-chart structure where each node denotes a test on an attribute, each branch represents the outcome of the test, and each node holds a class label, where the topmost node in a tree is the root node. A visualization of a decision can be seen below.



Along with Decision Tree comes the impurity that measures how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset, and is computed as follows,

$$I_g(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J} (p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} P_i^2 = 1 - \sum_{i=1}^{J} p_i^2.$$

6 Results

6.1 Multiple Linear Regression

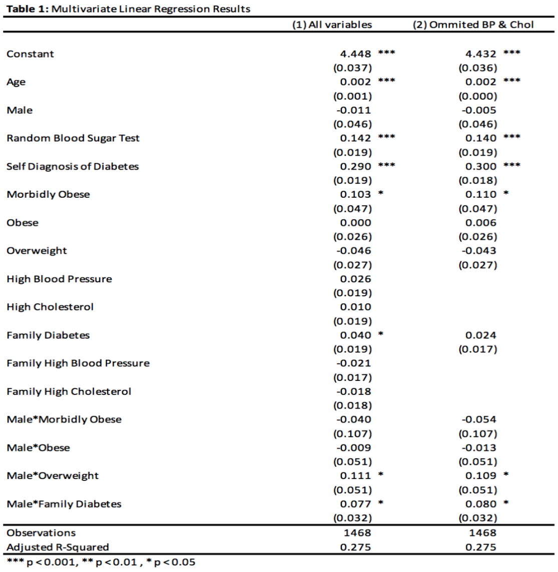


Table 1: Multivariate Linear Regression Results

	(1) All variables	(2) Ommited BP & Chol
Constant	4.448 ***	4.432 ***
	(0.037)	(0.036)
Age	0.002 ***	0.002 ***
	(0.001)	(0.000)
Male	-0.011	-0.005
	(0.046)	(0.046)
Random Blood Sugar Test	0.142 ***	0.140 ***
	(0.019)	(0.019)
Self Diagnosis of Diabetes	0.290 ***	0.300 ***
	(0.019)	(0.018)
Morbidly Obese	0.103 *	0.110 *
	(0.047)	(0.047)
Obese	0.000	0.006
	(0.026)	(0.026)
Overweight	-0.046	-0.043
	(0.027)	(0.027)
High Blood Pressure	0.026	
	(0.019)	
High Cholesterol	0.010	
	(0.019)	
Family Diabetes	0.040 *	0.024
	(0.019)	(0.017)
Family High Blood Pressure	-0.021	
	(0.017)	
Family High Cholesterol	-0.018	
	(0.018)	
Male*Morbidly Obese	-0.040	-0.054
	(0.107)	(0.107)
Male*Obese	-0.009	-0.013
	(0.051)	(0.051)
Male*Overweight	0.111 *	0.109 *
	(0.051)	(0.051)
Male*Family Diabetes	0.077 *	0.080 *
	(0.032)	(0.032)
Observations	1468	1468
Adjusted R-Squared	0.275	0.275

*** p < 0.001, ** p < 0.01 , * p < 0.05

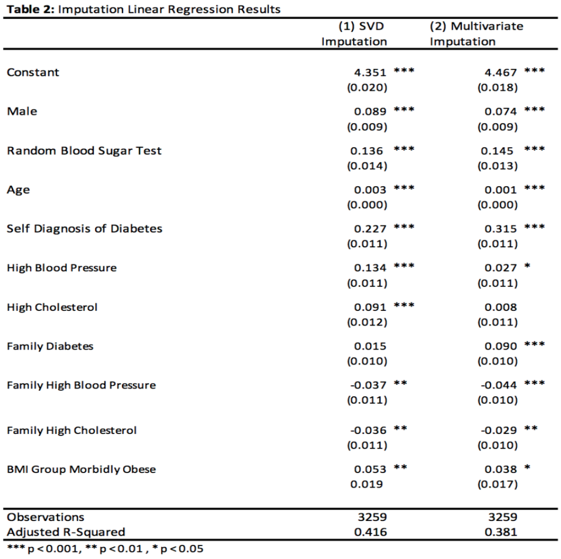


Table 2: Imputation Linear Regression Results

	(1) SVD Imputation	(2) Multivariate Imputation
Constant	4.351 ***	4.467 ***
	(0.020)	(0.018)
Male	0.089 ***	0.074 ***
	(0.009)	(0.009)
Random Blood Sugar Test	0.136 ***	0.145 ***
	(0.014)	(0.013)
Age	0.003 ***	0.001 ***
	(0.000)	(0.000)
Self Diagnosis of Diabetes	0.227 ***	0.315 ***
	(0.011)	(0.011)
High Blood Pressure	0.134 ***	0.027 *
	(0.011)	(0.011)
High Cholesterol	0.091 ***	0.008
	(0.012)	(0.011)
Family Diabetes	0.015	0.090 ***
	(0.010)	(0.010)
Family High Blood Pressure	-0.037 **	-0.044 ***
	(0.011)	(0.010)
Family High Cholesterol	-0.036 **	-0.029 **
	(0.011)	(0.010)
BMI Group Morbidly Obese	0.053 **	0.038 *
	0.019	(0.017)
Observations	3259	3259
Adjusted R-Squared	0.416	0.381

*** p < 0.001, ** p < 0.01 , * p < 0.05

Table 1 displays that a statistically significant inference was obtained with several variables including: age, gender, morbidly-obese BMI-Group, and family history of diabetes. As expected, positive statistical significance was achieved with the self-diagnosis of diabetes being true and taking a random blood sugar test versus the 8-12 hours fasting test. The P-value obtained approached 0 for these variables, indicating the strongest possible relationship between them and blood sugar. The results also concluded that BMI groups, excluding the morbidly obese group, had no statistical significance, indicating the possible glaring issue involved with using BMI Groups, in particular, the normal, overweight, and obese categories.

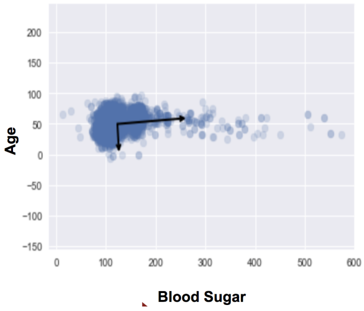
Results were obtained with non-statistically significant relationships, in particular the predictors of high blood pressure, cholesterol, and whether an individual had one of the two mentioned issues in their family, respectively. This implies the lack of usefulness regarding these measurements in relation to blood sugar levels, and this issue is accounted for with the following modified regression displayed in table 2. Overall, for comprehension of the extent to how much improvement was made with omitting the previously mentioned variables, viewing adjusted R-squared presented us with evidence that this model was a better fit without these variables. Particularly, there was no change in adjusted R-squared when omitting those four variables. Moreover, this may imply that these binary variables aren't an appropriate measurement regarding an individual's blood sugar levels. Although, to the contrary, it's known in practice that blood pressure and cholesterol levels do in fact impact an individual having higher/lower blood sugar levels. Perhaps a better measurement would've been a continuous variable with an actual value associated with these variables.

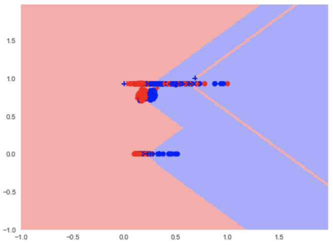
From Table 2 one is able to see that with age, 95% of the time, one expects an individual to see an increase of blood sugar levels of approximately .12% to .31% with each year they age. Although in this dataset, age spans from 11 to 88 years old, one obtains inference through this subset of people. Since the data contains ages in the early teens, this explains why positive relationship exists with blood sugar.

The table also displays the disparity between genders, indicating that 95% of the time, males have approximately 3.84% to 10.04% higher levels of blood sugar than females. These results are uncommon since there is no statistical significance regarding gender and blood sugar. This disparity between males and females, may be a result of the numbers of male participants versus the number of female participants. Also, with a family history of diabetes, 95% of the time, this leads to approximately 1.72% to 7.43% higher blood sugar levels versus no family diabetes history. This result is not unexpected due to genetics playing a factor regarding how an individual processes food and thus whether an individual has a predisposition for diabetes or not. If someone's parent has diabetes, then their child will likely deal with similar if not the same issues.

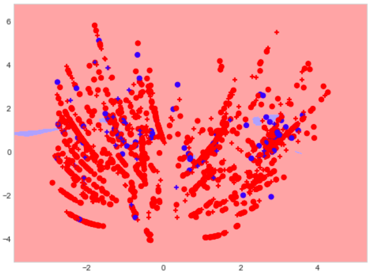
When an individual is categorized in the morbidly obese BMI group, 95% of the time, we expect to see an increase between 2.15% to 18.65% in blood sugar compared to the three other groups explored upon. This interval span is a lot larger compared to other statistically significant due to a lower p-value. This result is expected as well, because it's common knowledge that people categorized as obese have their body struggle with implementing insulin to control their blood sugar levels (http://www.obesity.org/content/weight-diabetes).

6.2 Principal Component & K-Nearest Neighbors Results





```
training score 0.902618657938
testing score 0.896932515337
```

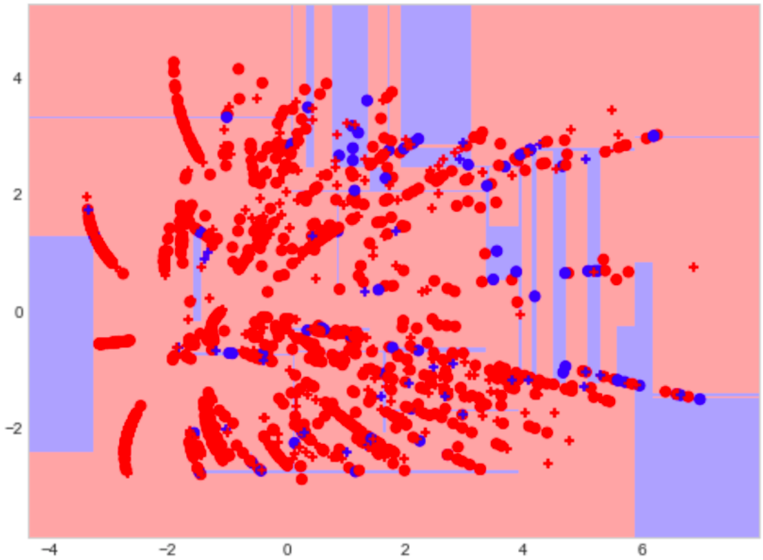


```
training score 0.930785123967
testing score 0.93
```

The picture on the left is a plot of the principle components. This graph indicates the the principal components and the direction with the most variance and least redundancy.

The plot on the center displays K-Nearest Neighbors with without feature engineering. Without feature engineering and using the raw data the algorithm performs at a 89% versus a 93% with feature engineering. These two plot displays the importance of feature engineering. Although our score was good without feature engineering, it improved with a couple features created. This indicates that if we create more features that are relevant our score will show even more improvement.

6.3 Decision Tree Results



```
training score 0.993801652893
testing score 0.89
```

As seen in our decision tree results above we have a training score of .993 and a testing score of .89. Based on these scores we can conclude that we have some overfitting in our results, which means that the noise or random fluctuations in the training data is picked up and learned as a concepts by the model. Thus the problem is that these concepts do not apply to new data and negatively impact the models ability to generalize. In this case the problem of overfitting can be solved by pruning a tree after it has learned in order to remove some of the detail it has picked up.

7 Conclusion

By using Linear Regression we were able to see the attributes that have a relationship with blood sugar on a statistically significant basis are age, family history of diabetes, males, and being morbidly obese. Moreover, the best non-parametric model results included the use of the KNN and the decision tree classifiers. Future work that can be done is to differentiate between type I and type II diabetes and to distinguish not only the levels of risk but also which type of diabetes should be a concern for an individual.