MATH-6357 FINAL PROJECT REPORT

Group 7: Diabetes Prediction

Names	PSID
Mahamadou Brehima Dagnoko	8011289
Swajan Reddy Gaddampally	2299019
Karthik Reddy Mettu	2264640
Jacob Smith	209923

1. Introduction

Diabetes is a chronic medical condition characterized by the inability of the body to properly regulate elevated blood sugar (glucose) levels. These issues arise due to the pancreas inability to produce enough insulin (Type I diabetes) or the body's ineffective use of insulin (Type II diabetes). Insulin which is a vital hormone that is derived from the food we eat, allows glucose to enter cells and be used as energy. Without proper regulation of blood sugar levels, diabetes arises and increases the risk of developing hyperglycemia which can damage various organs and systems over time.

In the United States, diabetes ranks among the top 10 most prevalent chronic conditions, affecting over 37 million people which represent over 11% of the population, according to the Centers for Disease Control and Prevention (CDC). It is the leading cause of severe health complications like heart disease, kidney failure, and blindness. Interestingly, the most common type of diabetes is Type II diabetes which can be avoided because it is mostly due to our lifestyle. Alarmingly, many individuals remain unaware of their condition until complications arise. Not only does this impact negatively the population's health but also the economy of the country. According to the CDC, it is estimated that The U.S. health care system spend over $327 billion on diabetes alone. This underscores the need for increased awareness and early intervention.

Addressing diabetes is crucial not only to improve the quality of life for millions of affected individuals but also to reduce the societal and economic strain it imposes. These are the reasons that pushed our team to use machine learning in diabetes prediction. In this study we focus on developing regression models to predict whether someone is at risk of developing diabetes (R1), on determining the key factors of Diabetes (R2), and on using those results to educate on ways to prevent diabetes (R3).

2. Results

2.1 regression models to predict diabetes (R1)

After conducting an exploratory data analysis to understand the relation between diabetes and our predictors, we wanted to quantify their impact (Details in section 3). We started by splitting the data into training and testing. We constructed our models on both the original training sample and an oversampled version of the training sample. We did this oversampled the minority class (Has-diabetes) using the ROSE library in R. First, we built a simple multilinear logistic regression model because a simple logistic regression provides great explainability of the relationship between response and predictors. However, if our model is not strong enough then we could be making wrong assumptions about the relation between diabetes and the attributes. To avoid this, we decided to build more flexible models and compared them to the simple one. Using the chi-squared test and a comparison of the Bayesian Information Criterions (BIC), we discovered that the simple model was enough to explain variabilities in the response. Knowing that, we had to figure out whether the oversampled training set was useful. This was done by testing the models' performance on the testing set, and by comparing the F1 scores and confusion matrixes. The results concluded that the simple linear logistic regression on the oversampled training set was the best model to both predict and explain diabetes using the available features.

2.2 key factors of Diabetes (R2)

Conducting an exploratory data analysis allowed us to uncover the main factors of diabetes. Using plots, we were able to observe that the percentage of diabetes patient increases within the male gender group. Similarly, people who do not have hypertension or a heart disease are usually unaffected by diabetes. Also, people who never smoked have the lowest percentage of diabetes followed by those who stopped smoking. Finally, people who have diabetes tend to be older and tend to have higher body max indexes (BMI). However, these observations could be due to the randomness in the sample. To solidify our interpretations, we developed a logistic regression model to quantify the impact of each factor on the likelihood of someone developing diabetes and used a z-test to confirm the importance of all the features (R1). The most important feature from the model is hypertension. Having hypertension increases the odds of having diabetes by 122%. Following that is having a heart disease which increases those odds by 95%. Next, being a male increases the odds by 42%. Unsurprisingly, age is also a significant factor because for each additional year, your odds of developing diabetes increase by 4.6%. Lastly, smoking is another factor in diabetes prediction. Not smoking for example reduces your chance of having diabetes by 19% and being a former smoker i.e quitting smoking decreases those odds by 5.5%. These results helped us understand the disease better allowing us to come up with solutions to prevent it.

2.3 Preventing diabetes (R3)

Preventing diabetes mostly requires maintaining a healthy lifestyle. The risk of having hypertension, heart diseases, and higher BMI can be avoided by exercising regularly. This should be done in addition to eating a balanced diet, staying hydrated, increasing fiber intake, and reducing sugar and alcohol consumption. Moreover, do not start smoking or quit it to reduce your chances of developing diabetes. Also, whereas we cannot stop aging or change gender, it is important to be more alert and follow these pieces of advice because they increase the chances of you having the decease. Finally, a regular health check like checking blood sugar levels, weight, and heart rate will help you detect irregularities in advance. So, do not wait to have diabetes to start having daily health routines because prevention is always better than cure.

3. Methods

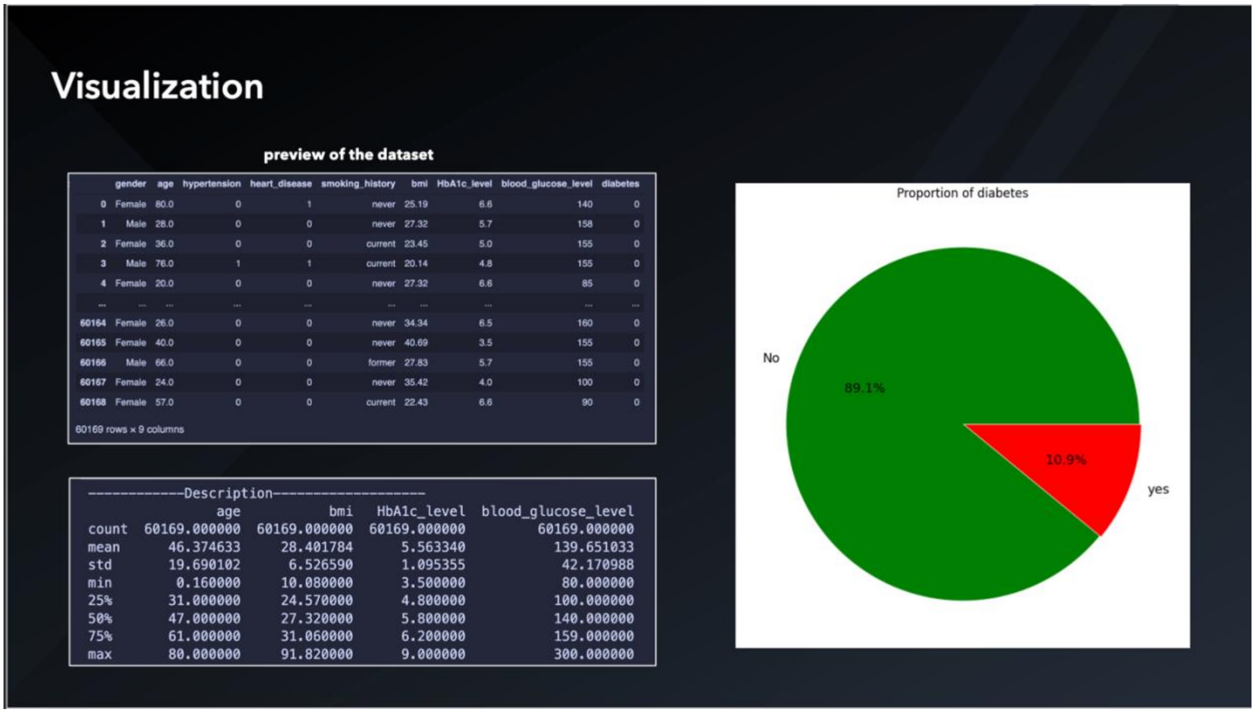
In this project, our main goal was to find the reasons why people develop diabetes. Answering this question required us to follow a series of steps.

3.1Exploratory Data Analysis (EDA)

The first thing to do was to find a dataset on which to conduct our study. After extensive research and consideration, we found a dataset on Kaggle. The dataset contained 9 features including the predictors and 60,000 rows where each row represents a patient. The numerical features were age measured in years, BMI which is the body mass index of the patient, blood glucose level (in mg/dL) which is measure percentage of glucose in the blood at a given time, and HbA1c level which is the average blood glucose level over 2 to 3 months. The blood glucose level and HbA1c levels are direct medical indicators of diabetes. A normal blood glucose level is usually less than 140 mg/dL and a value higher than 200 mg/dL is diabetes. Similarly, an HbA1c level below 5.7% is normal and a level above 6.5% is diabetes. Additionally, the BMI which is a ratio of the weight over the height squared has a normal value between 18.5 and 25. A value lower than 18.5 means that person is underweight, a value larger than 25 means that the person is overweight, and a value greater than 30 is condition called obesity. In addition to the numerical feature, the dataset contained categorical ones that were Gender (Male, Female, Other), Hypertension (1 if the patient has hypertension, 0 otherwise), heart disease (1 if the patient has heart disease, 0 otherwise), and smoking history (Never, Former, Current, Ever, No Info). Finally, the dataset had diabetes which is a categorical column indicating if the patient has diabetes (1 if diabetes, 0 if not). We will use this feature as the response variable to answer our research questions.

Before diving into the EDA, we had to clean the dataset. This was straight forward because we did not have any missing data. Therefore, we corrected inconsistencies in the features. This was done by dropping “Ever” and “No Info” from the smoking history columns and “others” from the gender column. This was done because there was no details about the meaning of “Ever”, “No Info” does not explain the smoking habit, and the class “others” in gender represented 0.02% of the data. We then combined “Not Current” and “former” smokers in the former column. Next, we excluded the HbA1c and blood sugar level from the analysis because they are direct medical measures to detect diabetes. Finally, we randomly split the dataset in two sets. 70% of the data will be used for EDA and training, and the remaining 30% will be used has testing. We will also use two versions of the training set. The first one is the training set as it is and the second is an oversampled one where we oversample the instances of diabetes to avoid class imbalance. The oversampling was done using ROSE. The ROSE (Random OverSampling Examples**)** package in R addresses class imbalance by generating synthetic data points through a smoothed kernel density estimate of the feature space, rather than duplicating existing samples. It creates a balanced dataset by oversampling the minority class and, if needed, undersampling the majority class. This approach generates diverse synthetic examples that preserve the dataset's structure, reducing the risk of overfitting compared to traditional methods. With that, the data cleaning was done, and we could focus on the analysis of the dataset containing 6 features and 1 target. Note that the gender and smoking history columns will be automatically transformed using onehot encoding during model building in R.

The dataset includes both numerical and categorical features. As we can see on Figure 1 below, for the numerical features, the age of individuals ranges from 18 to 80 years, with a mean of 45.2 years and a standard deviation of 15.6 years. The BMI ranges from 18.5 to 50.4, with a mean of 28.7 and a standard deviation of 5.4.



Visualization

preview of the dataset

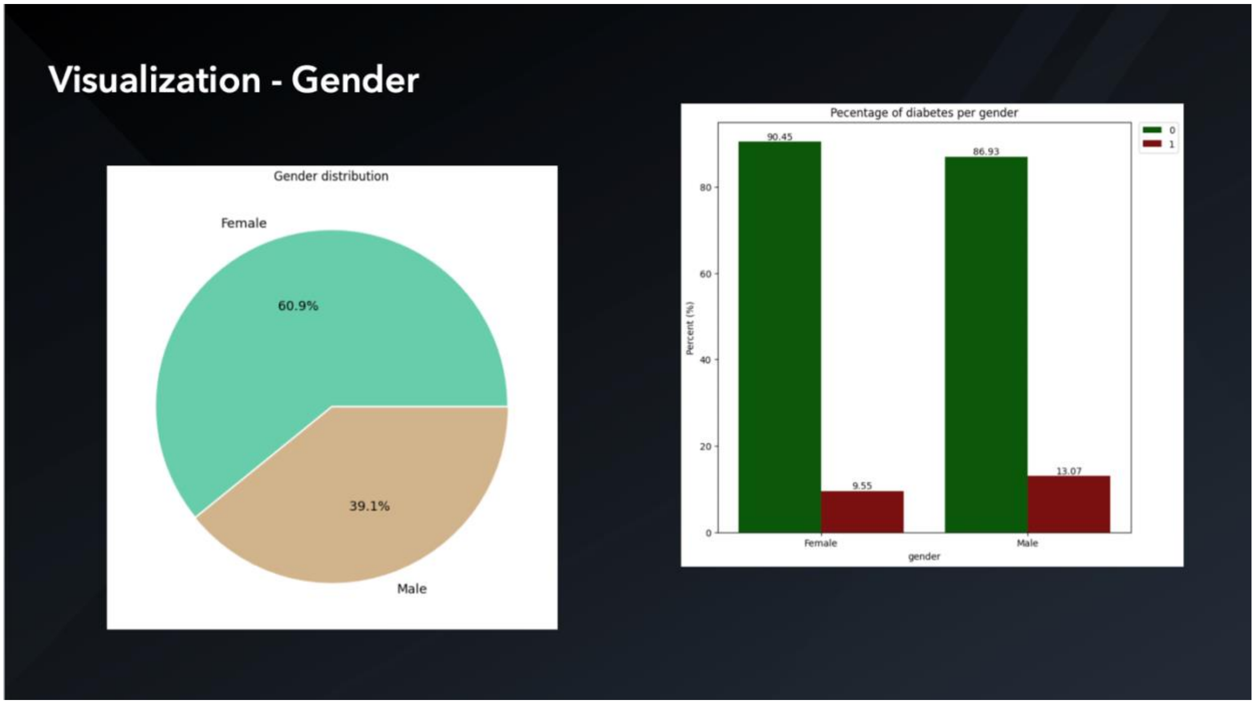
	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Male	28.0	0	0	never	27.32	5.7	158	0
2	Female	36.0	0	0	current	23.45	5.0	155	0
3	Male	76.0	1	1	current	20.14	4.8	155	0
4	Female	20.0	0	0	never	27.32	6.6	85	0
...	...	...	...	...	...	...	...	...	...
60164	Female	26.0	0	0	never	34.34	6.5	160	0
60165	Female	40.0	0	0	never	40.69	3.5	155	0
60166	Male	66.0	0	0	former	27.83	5.7	155	0
60167	Female	24.0	0	0	never	35.42	4.0	100	0
60168	Female	57.0	0	0	current	22.43	6.6	90	0

60169 rows × 9 columns

```
------------Description--------------------
             age          bmi  HbA1c_level  blood_glucose_level
count  60169.000000 60169.000000 60169.000000         60169.000000
mean      46.374633    28.401784     5.563340           139.651033
std       19.690102     6.526590     1.095355            42.170988
min        0.160000    10.080000     3.500000            80.000000
25%       31.000000    24.570000     4.800000           100.000000
50%       47.000000    27.320000     5.800000           140.000000
75%       61.000000    31.060000     6.200000           159.000000
max       80.000000    91.820000     9.000000           300.000000
```

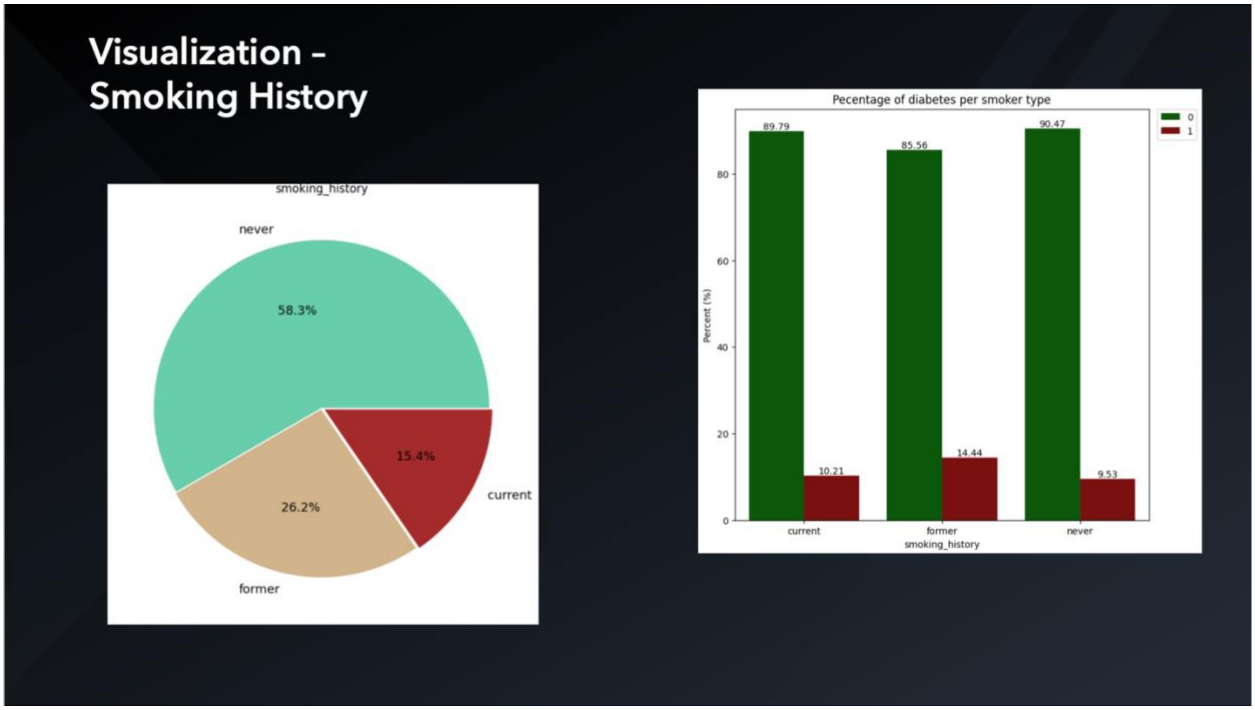
Figure 1: Descriptive statistics of the dataset

For the categorical features, 39% of participants are male, and 61% are female (Figure 2). A higher percentage of male have diabetes than woman. Regarding smoking history, 58% of participants have never smoked, 26% are former smokers, and 16% are current smokers (Figure 3). People who never smoked have the lowest percentage of diabetes, followed by those who stopped smoking. Additionally, individuals without hypertension or heart disease are less likely to have diabetes (Figure 4). Moreover, diabetes prevalence tends to increase with age and higher body mass index (Figure 5, 6, and 7). On figure 7, we can also see that when you have blood glucose level and HbA1c, you can tell if someone has diabetes and the higher those values, the higher the chance of having diabetes. For the univariate analysis, the age distribution is skewed towards middle age 45-54 years, and we can see that more older people have diabetes (Figure 5). The BMI distribution follows a normal distribution with a slight right skewness. Knowing these from the dataset gave us insight of the relation between diabetes and the predictors. So, the next to use machine learning to validate our assumptions.



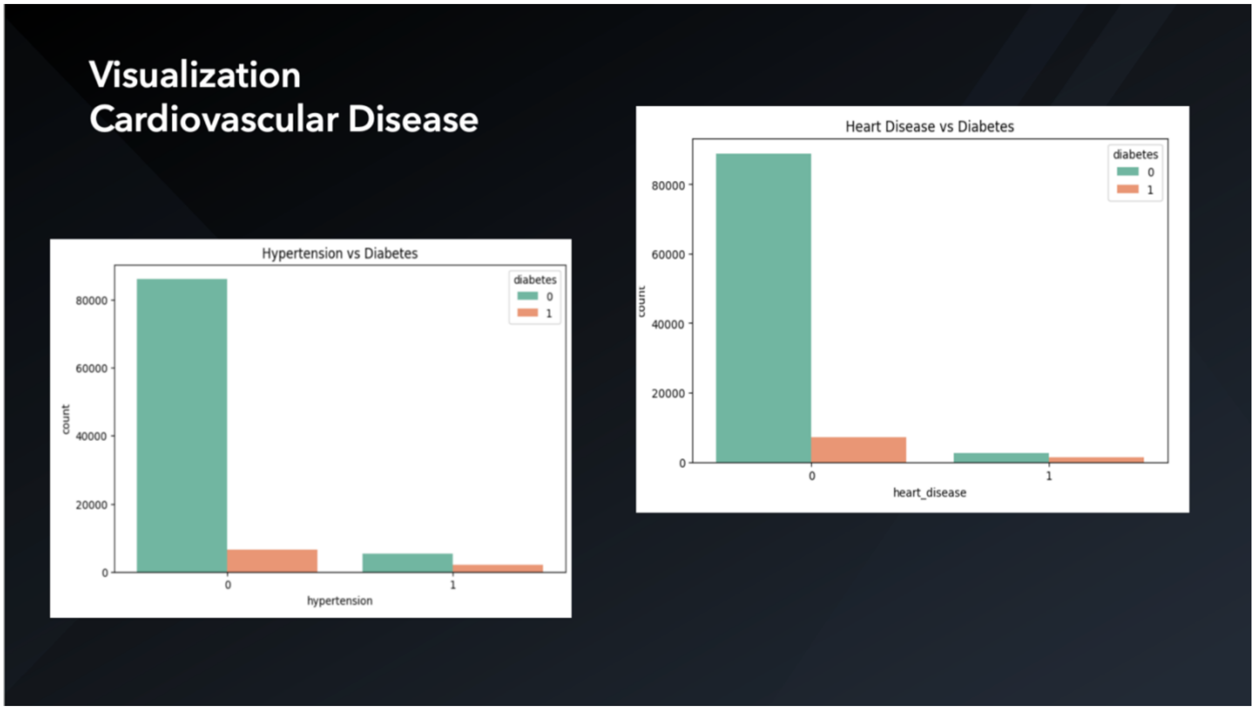
Visualization - Gender

Figure 2: Visualization of gender



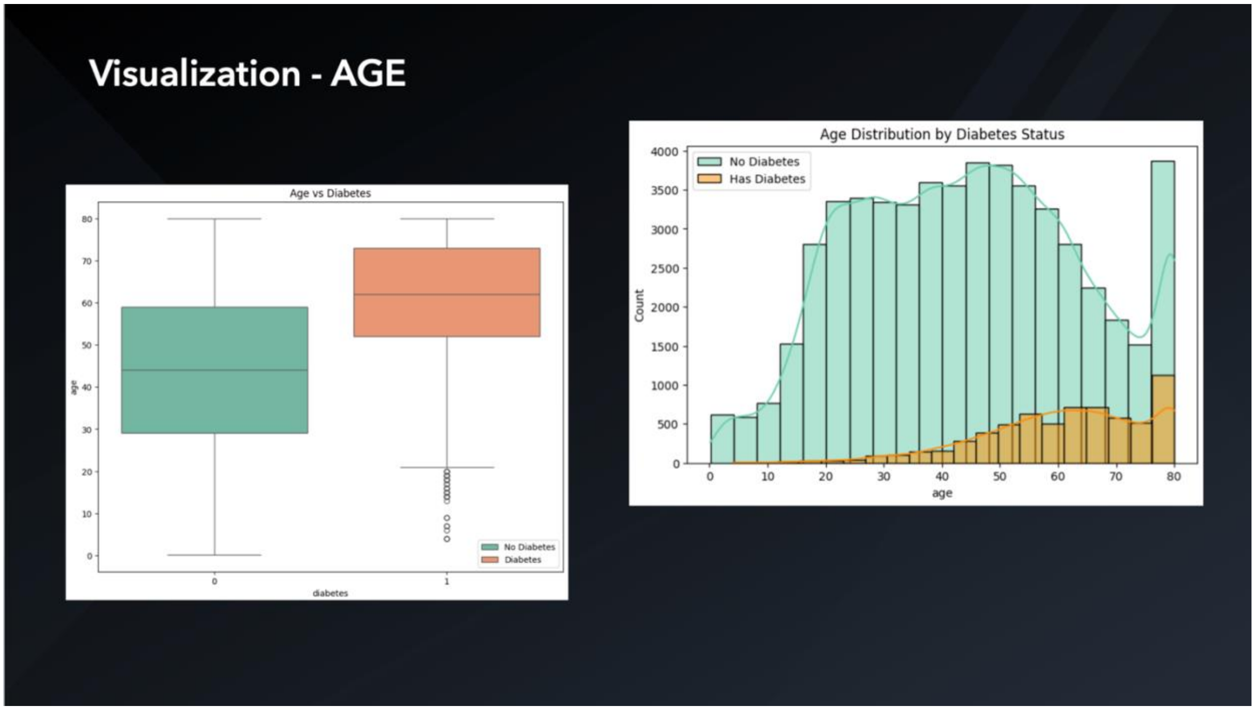
Visualization – Smoking History

Figure 3: Visualization of Smoking History



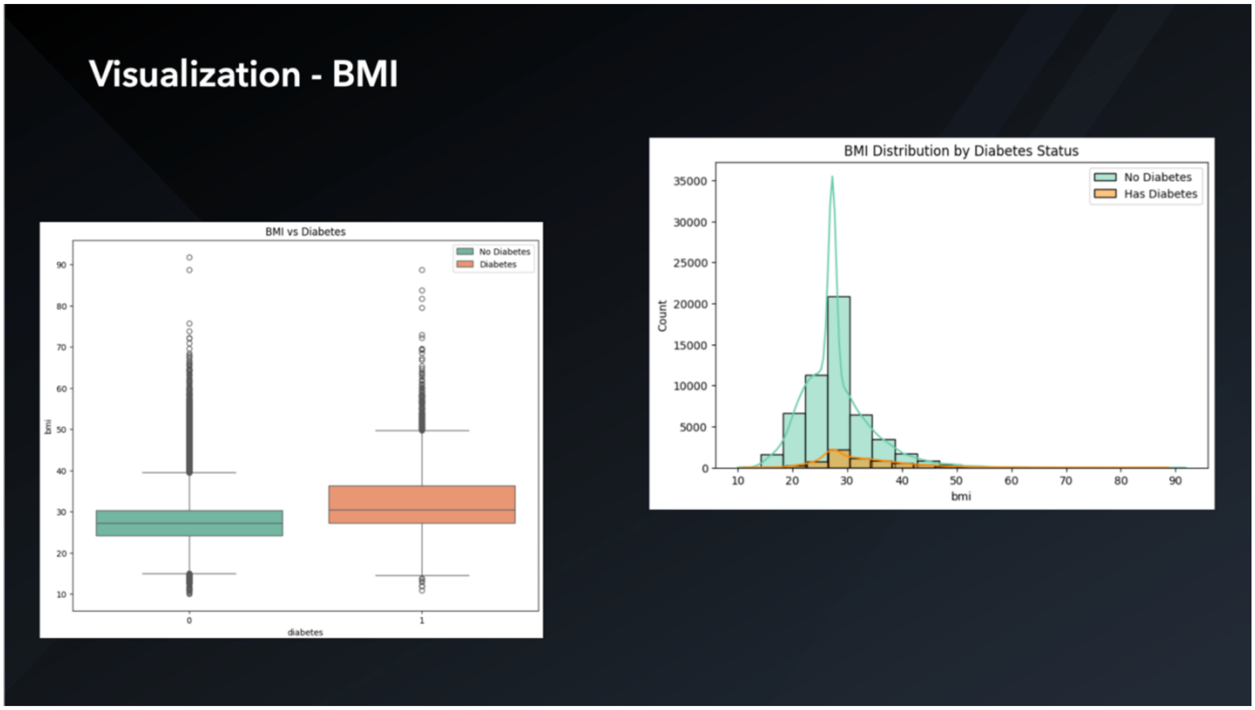
Visualization Cardiovascular Disease

Figure 4: Visualization of hypertension and heart disease



Visualization - AGE

Figure 5: Visualization of Age



Visualization - BMI

Figure 6: Visualization of BMI

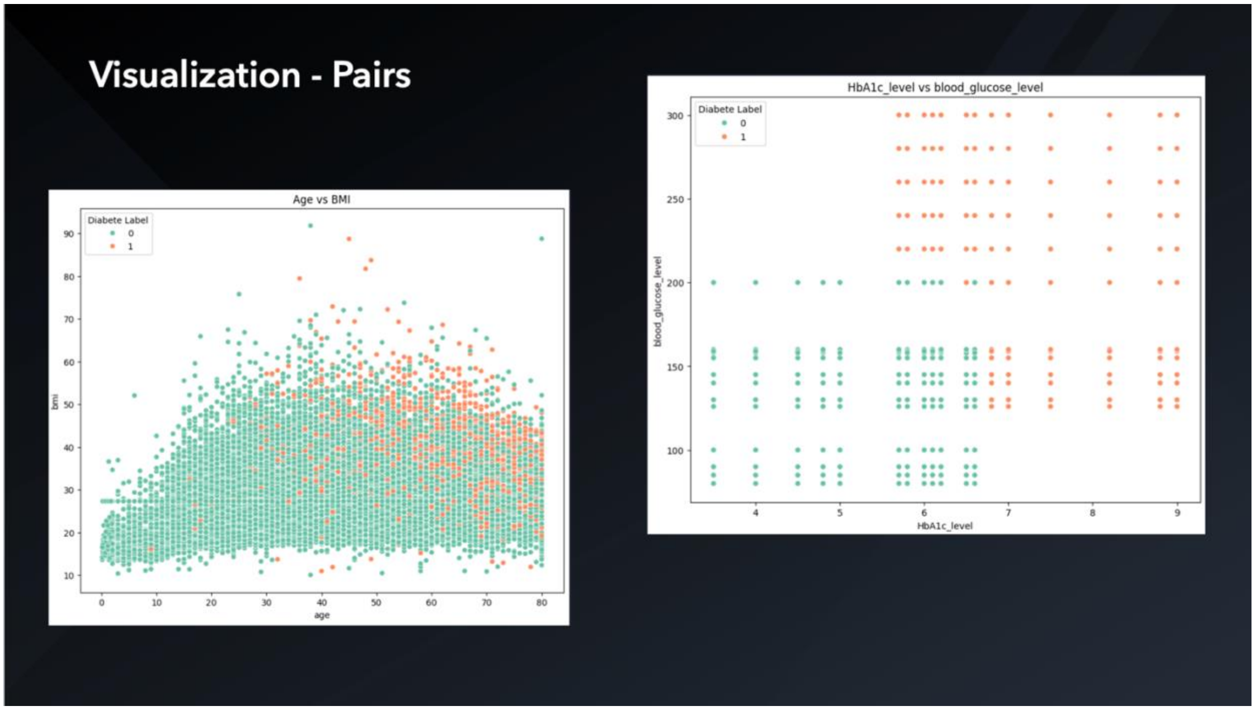
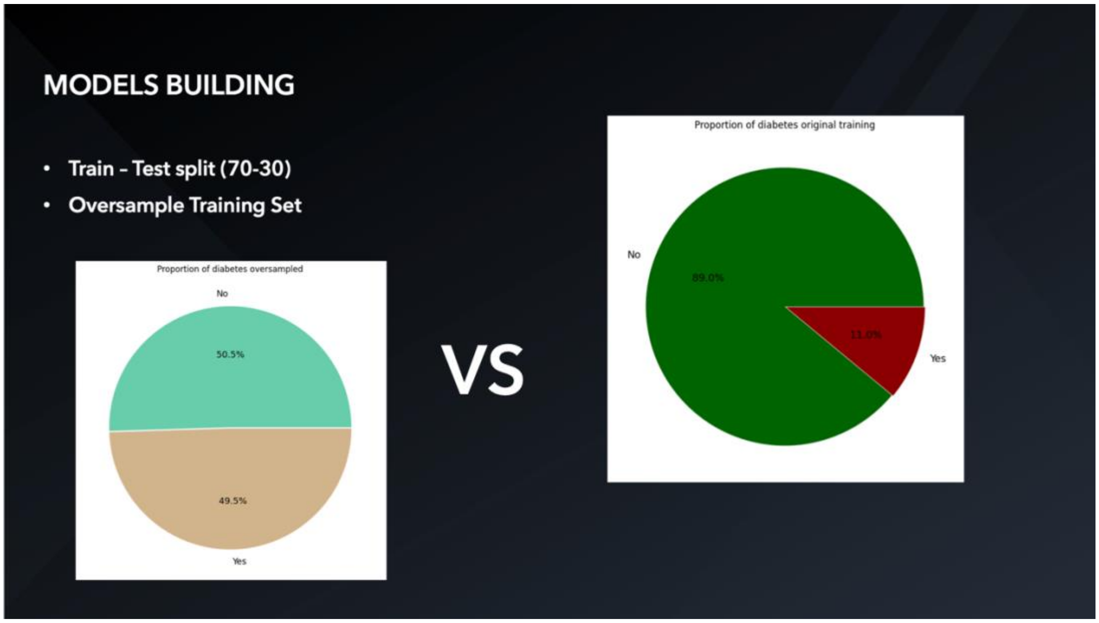

Visualization - Pairs

Figure 7: Supervised scatterplot of numerical features

3.2 Model Building

The dataset initially displayed an imbalanced composition regarding the presence of diabetes (89% with diabetes; 11% without). This imbalance could cause problems in machine learning models where they would give more importance to the majority class. To avoid this, we will also use an oversampled training set in our model building (section 3.1). A correlation analysis showed that there is low positive correlation among the numerical features. Focusing on age and BMI, this could help avoid multicollinearity of the data (Figure 9).


MODELS BUILDING

- Train - Test split (70-30)
- Oversample Training Set

Figure 8: training set vs oversampled training set

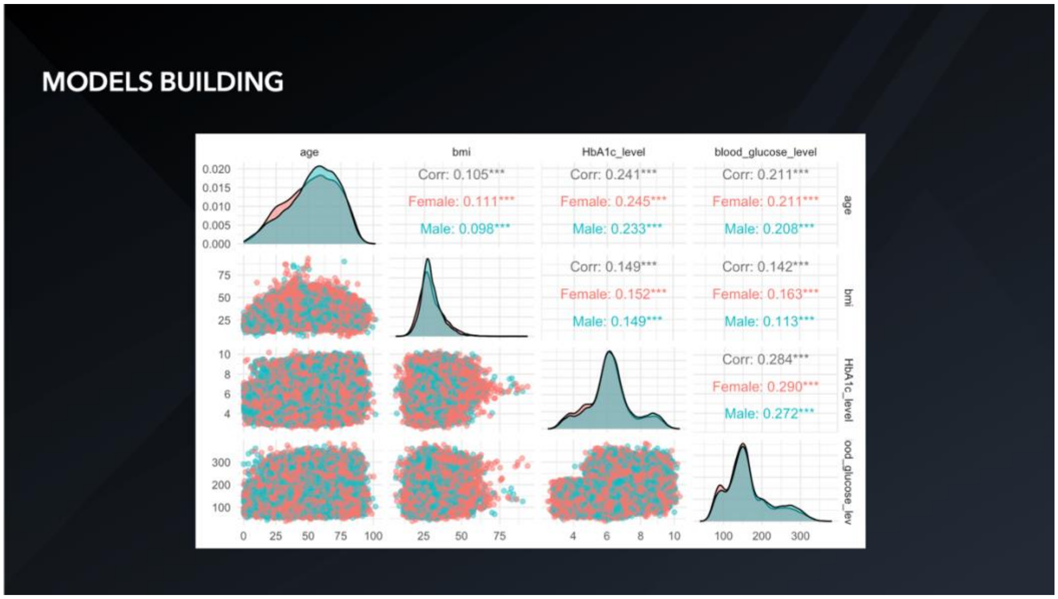

MODELS BUILDING

Figure 9: Supervised Correlation matrix

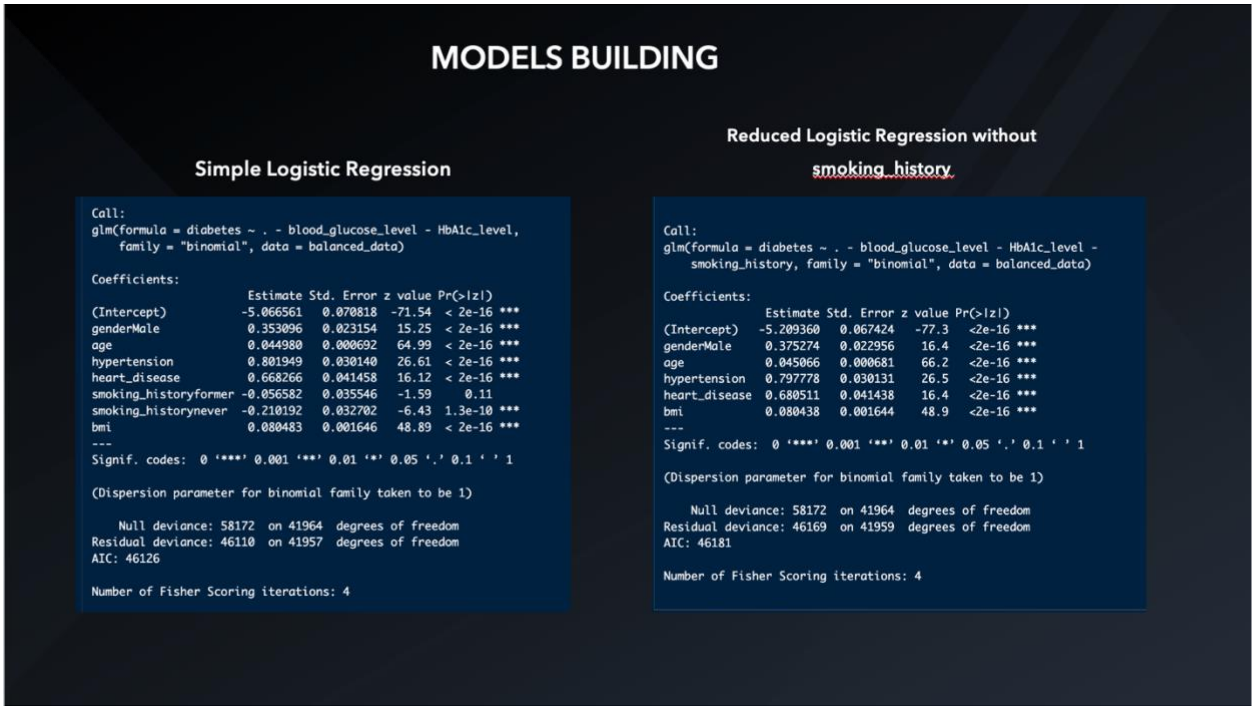
Following the correlation analysis, we started building the models. We started by building a multilinear logistic regression model because our target, diabetes, is a categorical target. A linear regression would not be suitable because a linear model has output values that goes from negative infinity to infinity. Since those values will not have any meaning when the only values our target has is 0 or 1, it is better to use the logistic model which uses the sigmoid function to make the predictions fall between 0 and 1. Thus, the closer the values are to 0 or 1, the more confident our model is that these predictions belong to closer class. The model was first built on the normal training set and the oversampled training set. We then used the z-test to validate the usefulness of each feature in the model (Figure 10). The tests concluded that being a former smoker (one hot encoded column) is not important in the model when the other variables are included. An Anova test reinforced by the chi-squared test and the BIC concluded that adding the smoking history feature provides a little more variability explanation. Although we can drop this column at this point, we decided to keep it because it can still help us quantify how smoking affects diabetes. This is because smoker and never smoker are important in the model (Figure 10 and 11). Following this we wanted to check if the model meets the assumptions of logistic regression.

Multiple linear regression will not work for this dataset because the goal is to predict a binary response variable on if a person has diabetes or not. If multiple linear regression is implemented then the range of predictions will be outside of {0,1} and many of the model assumptions would be violated. The following are the model assumptions for logistic regression that will be explored. The response variable should be binary, taking values of 0 or 1, and each observation should be independent of others. A sufficiently large sample size is required to allow for reliable estimation and robust model performance. The relationship between the independent variables and the log-odds of the outcome must be linear, a condition known as the linearity of the logit. Predictors should not exhibit perfect multicollinearity, as this can distort the model's coefficients and reduce interpretability. Additionally, high-leverage points, or observations with extreme predictor values, should be monitored to prevent undue influence on the model.

The dataset satisfies most of the assumptions for logistic regression. The binary outcome assumption is met as the response variable indicates whether a person has diabetes (1) or does not (0). Independence of observations is also satisfied since each observation represents a unique individual. With over 60,000 observations, the dataset has a sufficiently large sample size for reliable model estimation.

To evaluate the linearity of the logit, the quantitative independent variables were analyzed using log-odds graphs and the Box-Tidwell test. While the graphs suggest linear relationships, the Box-Tidwell test indicates that age meets the linearity assumption, but BMI does not, with a p-value less than 0.05. To address this, polynomial transformations of BMI will be incorporated into the model (Figure 12)

The absence of perfect multicollinearity was confirmed using Variance Inflation Factors (VIFs), all of which were close to 1, indicating no correlated predictor variables (Figure 13). Lastly, Cook’s distance was used to check for influential observations, revealing one potential outlier. However, its Cook’s distance was very small, allowing us to conclude that this assumption is also satisfied (Figure 13).



MODELS BUILDING

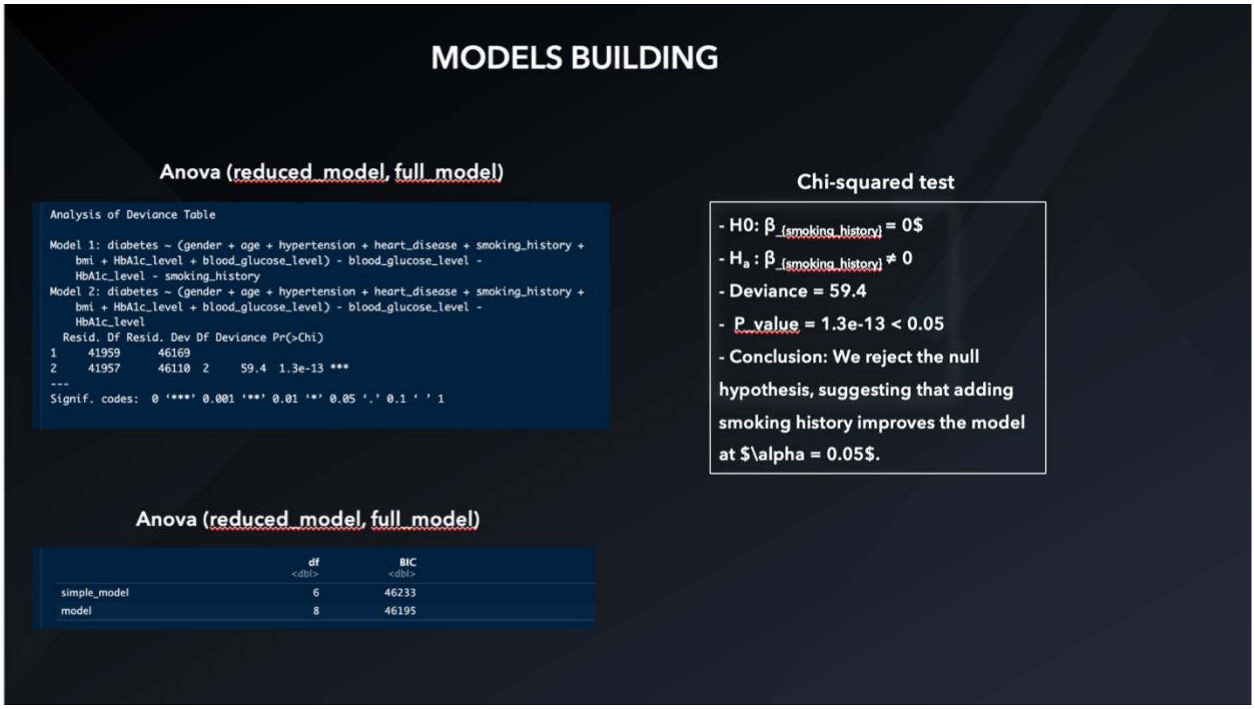
Simple Logistic Regression

```
Call:
glm(formula = diabetes ~ . - blood_glucose_level - HbA1c_level,
    family = "binomial", data = balanced_data)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -5.066561   0.070818  -71.54  < 2e-16 ***
genderMale             0.353096   0.023154   15.25  < 2e-16 ***
age                    0.044980   0.000692   64.99  < 2e-16 ***
hypertension           0.801949   0.030140   26.61  < 2e-16 ***
heart_disease          0.668266   0.041458   16.12  < 2e-16 ***
smoking_historyformer -0.056582   0.035546   -1.59     0.11
smoking_historynever  -0.210192   0.032702   -6.43  1.3e-10 ***
bmi                    0.080483   0.001646   48.89  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58172  on 41964  degrees of freedom
Residual deviance: 46110  on 41957  degrees of freedom
AIC: 46126

Number of Fisher Scoring iterations: 4
```

Reduced Logistic Regression without smoking_history

```
Call:
glm(formula = diabetes ~ . - blood_glucose_level - HbA1c_level -
    smoking_history, family = "binomial", data = balanced_data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.209360   0.067424   -77.3   <2e-16 ***
genderMale     0.375274   0.022956    16.4   <2e-16 ***
age            0.045066   0.000681    66.2   <2e-16 ***
hypertension   0.797778   0.030131    26.5   <2e-16 ***
heart_disease  0.680511   0.041438    16.4   <2e-16 ***
bmi            0.080438   0.001644    48.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58172  on 41964  degrees of freedom
Residual deviance: 46169  on 41959  degrees of freedom
AIC: 46181

Number of Fisher Scoring iterations: 4
```

Figure 10: Simple logistic regression on oversampled training set



MODELS BUILDING

Anova (reduced_model, full_model)

```
Analysis of Deviance Table

Model 1: diabetes ~ (gender + age + hypertension + heart_disease + smoking_history +
    bmi + HbA1c_level + blood_glucose_level) - blood_glucose_level -
    HbA1c_level - smoking_history
Model 2: diabetes ~ (gender + age + hypertension + heart_disease + smoking_history +
    bmi + HbA1c_level + blood_glucose_level) - blood_glucose_level -
    HbA1c_level
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     41959      46169
2     41957      46110  2     59.4  1.3e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

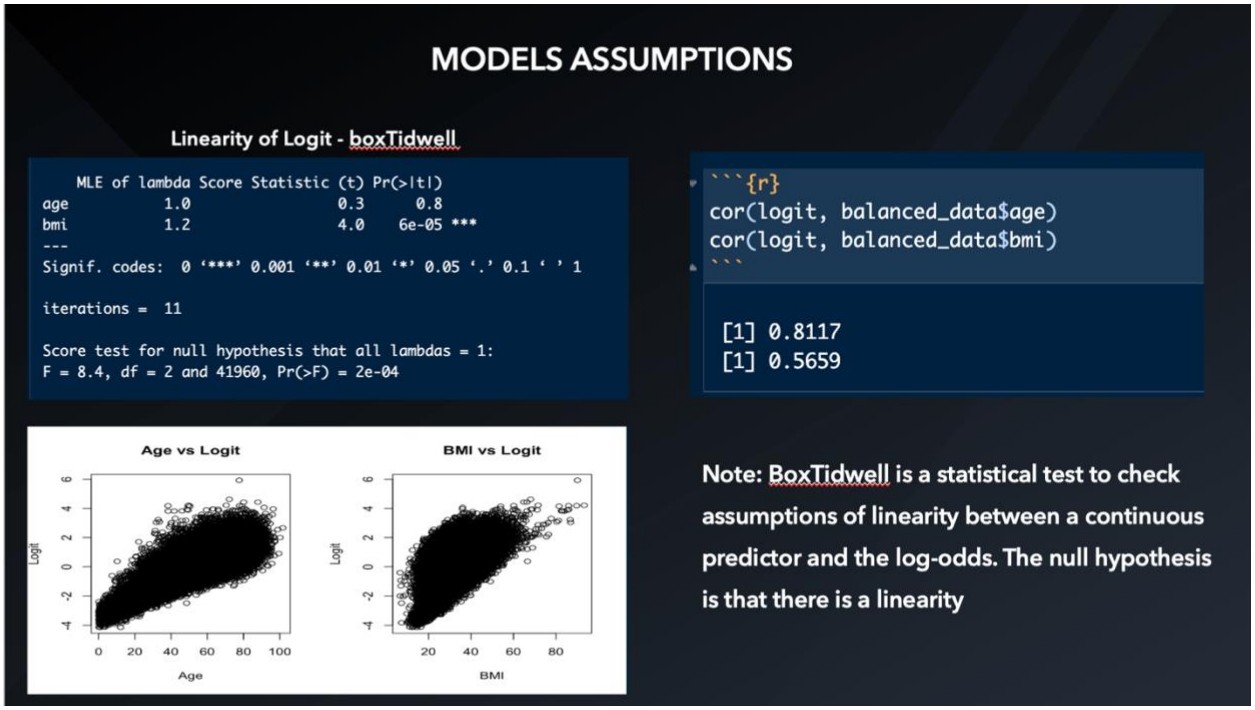
Chi-squared test

- H0: $\beta_{\_\{smoking\_history\}}$ = 0$
- H_a : $\beta_{\_\{smoking\_history\}} \neq 0$
- Deviance = 59.4
- P_value = 1.3e-13 < 0.05
- Conclusion: We reject the null hypothesis, suggesting that adding smoking history improves the model at $\alpha = 0.05$.

Anova (reduced_model, full_model)

	df <dbl>	BIC <dbl>
simple_model	6	46233
model	8	46195

Figure 11: Test if smoking history can be dropped



MODELS ASSUMPTIONS

Linearity of Logit - boxTidwell

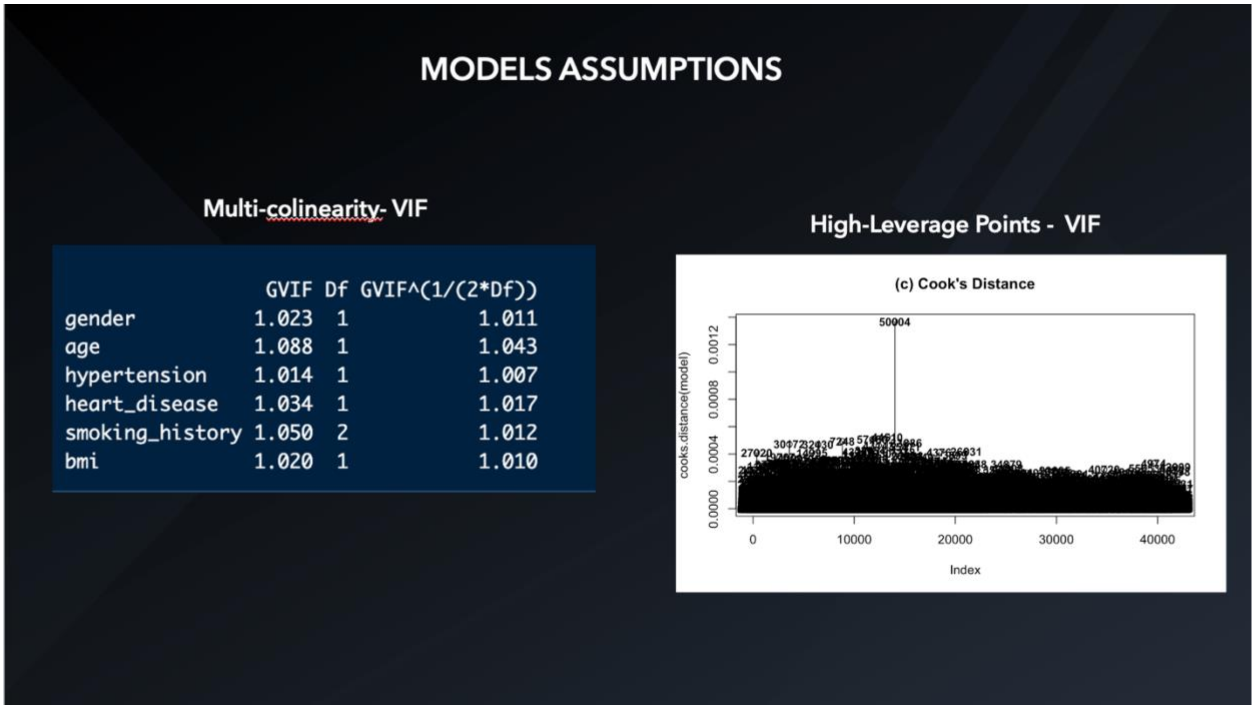
```
    MLE of lambda Score Statistic (t) Pr(>|t|)
age           1.0                  0.3      0.8
bmi           1.2                  4.0    6e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  11

Score test for null hypothesis that all lambdas = 1:
F = 8.4, df = 2 and 41960, Pr(>F) = 2e-04
```

```{r}
cor(logit, balanced_data$age)
cor(logit, balanced_data$bmi)
```

```
[1] 0.8117
[1] 0.5659
```

Note: BoxTidwell is a statistical test to check assumptions of linearity between a continuous predictor and the log-odds. The null hypothesis is that there is a linearity

Figure 12: Linearity of logit assumption



MODELS ASSUMPTIONS

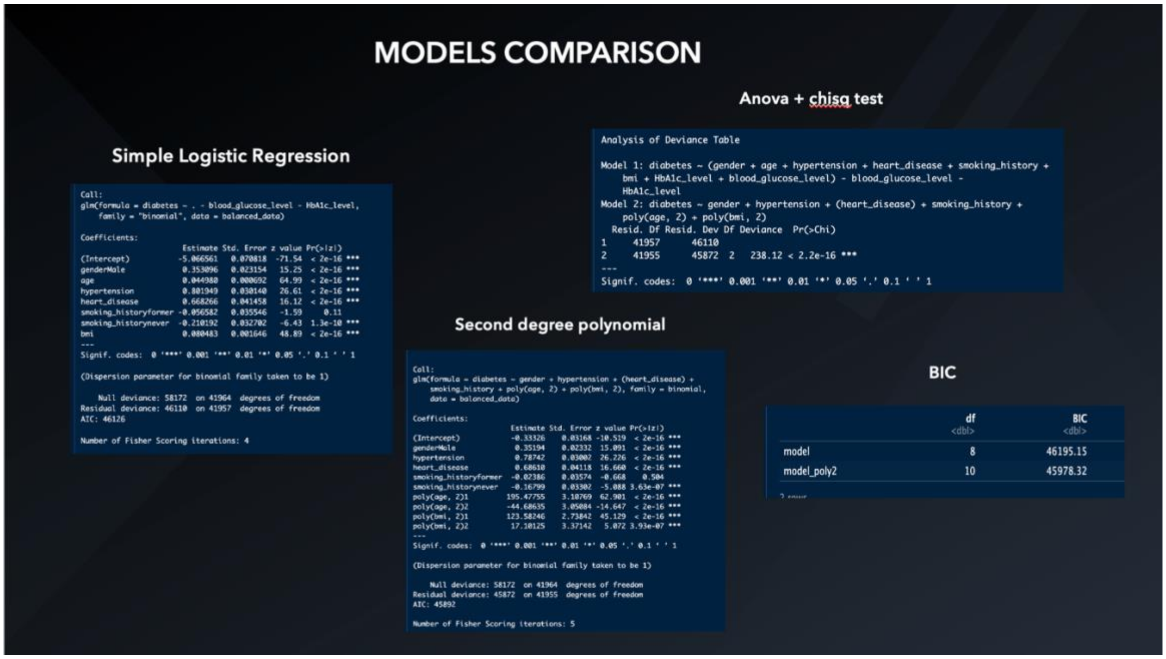
Multi-colinearity- VIF

```
                 GVIF Df GVIF^(1/(2*Df))
gender          1.023  1           1.011
age             1.088  1           1.043
hypertension    1.014  1           1.007
heart_disease   1.034  1           1.017
smoking_history 1.050  2           1.012
bmi             1.020  1           1.010
```

High-Leverage Points - VIF

Figure 13: Multi-collinearity and outlying assumptions

Since the tests were not too confident about the assumption of linearity of the logit, we soften the boundaries of the logistic model to see if a more complex model would be better to explain the variabilities in the response. We built a second and third order logistic regression and used the Anova test, the chi-squared test, and the BIC metric to check if adding higher polynomial terms were useful. In summary, the results showed that adding a second terms is slightly better but adding a third polynomial term was not as important (Figure 14 and 15). The BIC for the simple and second-degree models were close, so to check if the added variability is significant, we tested the models on the testing set. The results showed that the two models performed about the same making the simple model the better one. Furthermore, we can see that the simple model trained on the oversampled dataset predicts correctly more instances of diabetes than the one trained on the original training set (Figure 16). At this point we were confident in our model's ability in predicting and explaining diabetes and used it to answer our research questions (Figure 17).



MODELS COMPARISON

Simple Logistic Regression

```
Call:
glm(formula = diabetes ~ . - blood_glucose_level - HbA1c_level,
    family = "binomial", data = balanced_data)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -5.066561   0.070818  -71.54  < 2e-16 ***
genderMale             0.353096   0.023154   15.25  < 2e-16 ***
age                    0.044980   0.000692   64.99  < 2e-16 ***
hypertension           0.801949   0.030140   26.61  < 2e-16 ***
heart_disease          0.668266   0.041458   16.12  < 2e-16 ***
smoking_historyformer -0.056582   0.035546   -1.59     0.11
smoking_historynever  -0.210192   0.032702   -6.43  1.3e-10 ***
bmi                    0.080483   0.001646   48.89  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58172  on 41964  degrees of freedom
Residual deviance: 46110  on 41957  degrees of freedom
AIC: 46126

Number of Fisher Scoring iterations: 4
```

Anova + chisq test

```
Analysis of Deviance Table

Model 1: diabetes ~ (gender + age + hypertension + heart_disease + smoking_history +
    bmi + HbA1c_level + blood_glucose_level) - blood_glucose_level -
    HbA1c_level
Model 2: diabetes ~ gender + hypertension + (heart_disease) + smoking_history +
    poly(age, 2) + poly(bmi, 2)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     41957      46110
2     41955      45872  2   238.12 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

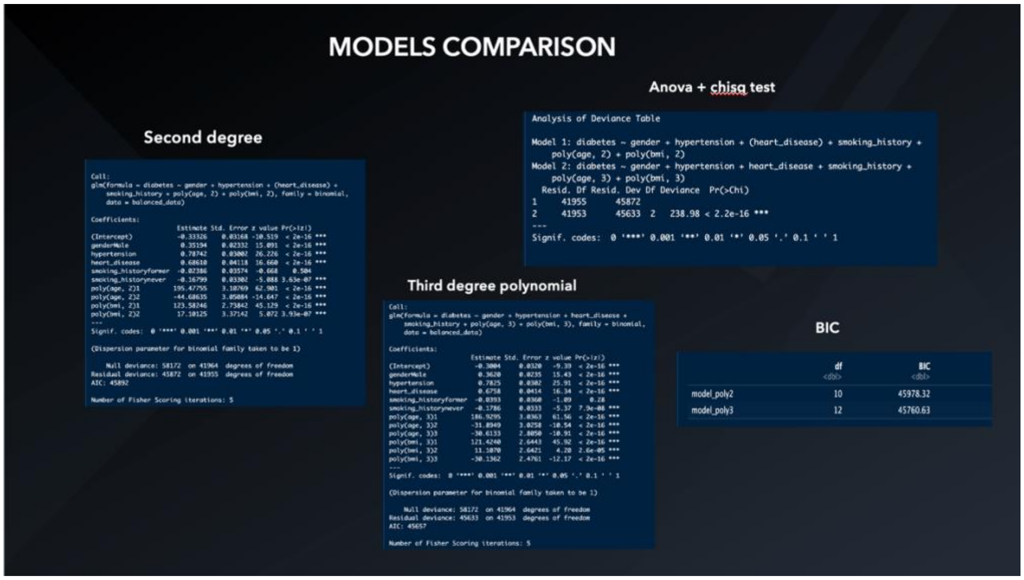
Second degree polynomial

```
Call:
glm(formula = diabetes ~ gender + hypertension + (heart_disease) +
    smoking_history + poly(age, 2) + poly(bmi, 2), family = binomial,
    data = balanced_data)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.33326    0.03168 -10.519  < 2e-16 ***
genderMale              0.35194    0.02332  15.091  < 2e-16 ***
hypertension            0.78742    0.03002  26.226  < 2e-16 ***
heart_disease           0.68610    0.04118  16.660  < 2e-16 ***
smoking_historyformer  -0.02386    0.03574  -0.668    0.504
smoking_historynever   -0.16799    0.03302  -5.088 3.63e-07 ***
poly(age, 2)1         195.47755    3.10769  62.901  < 2e-16 ***
poly(age, 2)2         -44.68635    3.05084 -14.647  < 2e-16 ***
poly(bmi, 2)1         123.58246    2.73842  45.129  < 2e-16 ***
poly(bmi, 2)2          17.10125    3.37142   5.072 3.93e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58172  on 41964  degrees of freedom
Residual deviance: 45872  on 41955  degrees of freedom
AIC: 45892

Number of Fisher Scoring iterations: 5
```

BIC

	df <dbl>	BIC <dbl>
model	8	46195.15
model_poly2	10	45978.32

Figure 14: Simple vs Second degree model



MODELS COMPARISON

Second degree

```
Call:
glm(formula = diabetes ~ gender + hypertension + (heart_disease) +
    smoking_history + poly(age, 2) + poly(bmi, 2), family = binomial,
    data = balanced_data)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.33326    0.03168 -10.519  < 2e-16 ***
genderMale              0.35194    0.02332  15.091  < 2e-16 ***
hypertension            0.78742    0.03002  26.226  < 2e-16 ***
heart_disease           0.68610    0.04118  16.660  < 2e-16 ***
smoking_historyformer  -0.02386    0.03574  -0.668    0.504
smoking_historynever   -0.16799    0.03302  -5.088 3.63e-07 ***
poly(age, 2)1         195.47755    3.10769  62.901  < 2e-16 ***
poly(age, 2)2         -44.68635    3.05084 -14.647  < 2e-16 ***
poly(bmi, 2)1         123.58246    2.73842  45.129  < 2e-16 ***
poly(bmi, 2)2          17.10125    3.37142   5.072 3.93e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58172  on 41964  degrees of freedom
Residual deviance: 45872  on 41955  degrees of freedom
AIC: 45892

Number of Fisher Scoring iterations: 5
```

Anova + chisq test

```
Analysis of Deviance Table

Model 1: diabetes ~ gender + hypertension + (heart_disease) + smoking_history +
    poly(age, 2) + poly(bmi, 2)
Model 2: diabetes ~ gender + hypertension + heart_disease + smoking_history +
    poly(age, 3) + poly(bmi, 3)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     41955      45872
2     41953      45633  2   238.98 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

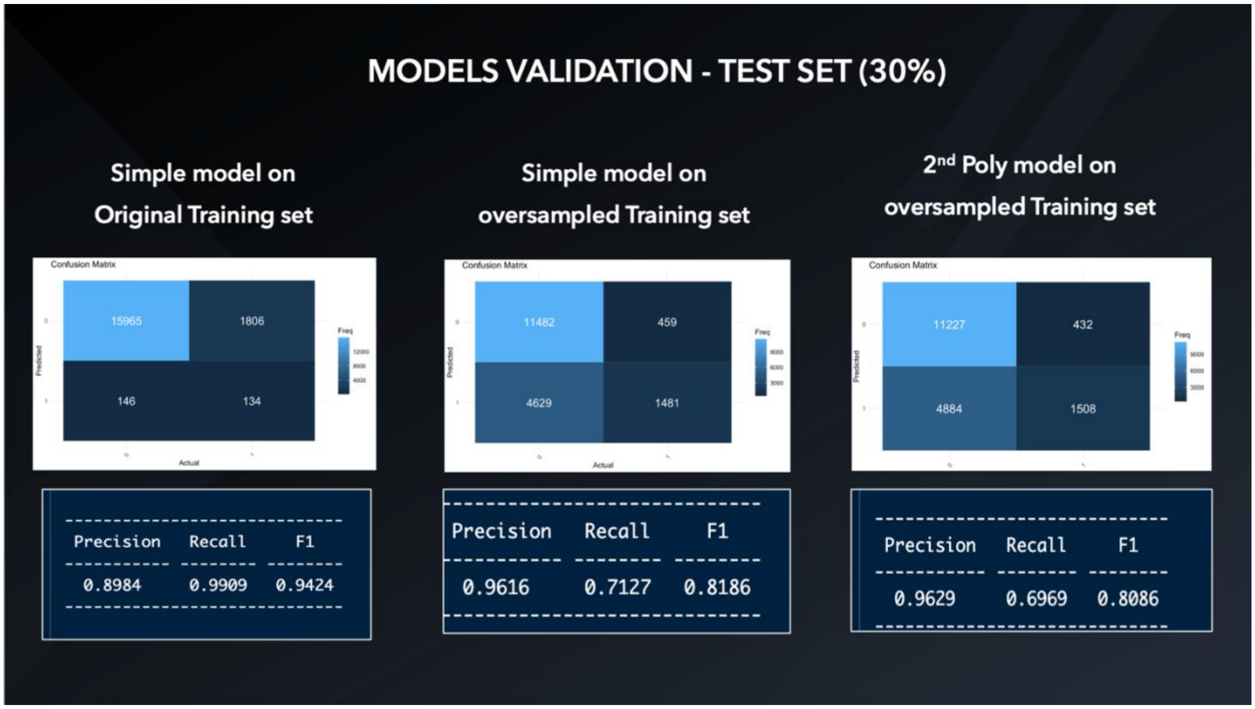
Third degree polynomial

```
Call:
glm(formula = diabetes ~ gender + hypertension + heart_disease +
    smoking_history + poly(age, 3) + poly(bmi, 3), family = binomial,
    data = balanced_data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.3004     0.0320   -9.39  < 2e-16 ***
genderMale              0.3620     0.0235   15.43  < 2e-16 ***
hypertension            0.7825     0.0302   25.91  < 2e-16 ***
heart_disease           0.6758     0.0414   16.34  < 2e-16 ***
smoking_historyformer  -0.0393     0.0360   -1.09     0.28
smoking_historynever   -0.1786     0.0333   -5.37  7.9e-08 ***
poly(age, 3)1         186.9295     3.0363   61.56  < 2e-16 ***
poly(age, 3)2         -31.8949     3.0258  -10.54  < 2e-16 ***
poly(age, 3)3         -30.6133     2.8050  -10.91  < 2e-16 ***
poly(bmi, 3)1         121.4240     2.6443   45.92  < 2e-16 ***
poly(bmi, 3)2          11.1070     2.6421    4.20  2.6e-05 ***
poly(bmi, 3)3         -30.1362     2.4761  -12.17  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58172  on 41964  degrees of freedom
Residual deviance: 45633  on 41953  degrees of freedom
AIC: 45657

Number of Fisher Scoring iterations: 5
```

BIC

	df <dbl>	BIC <dbl>
model_poly2	10	45978.32
model_poly3	12	45760.63

Figure 15: Second VS Third degree model



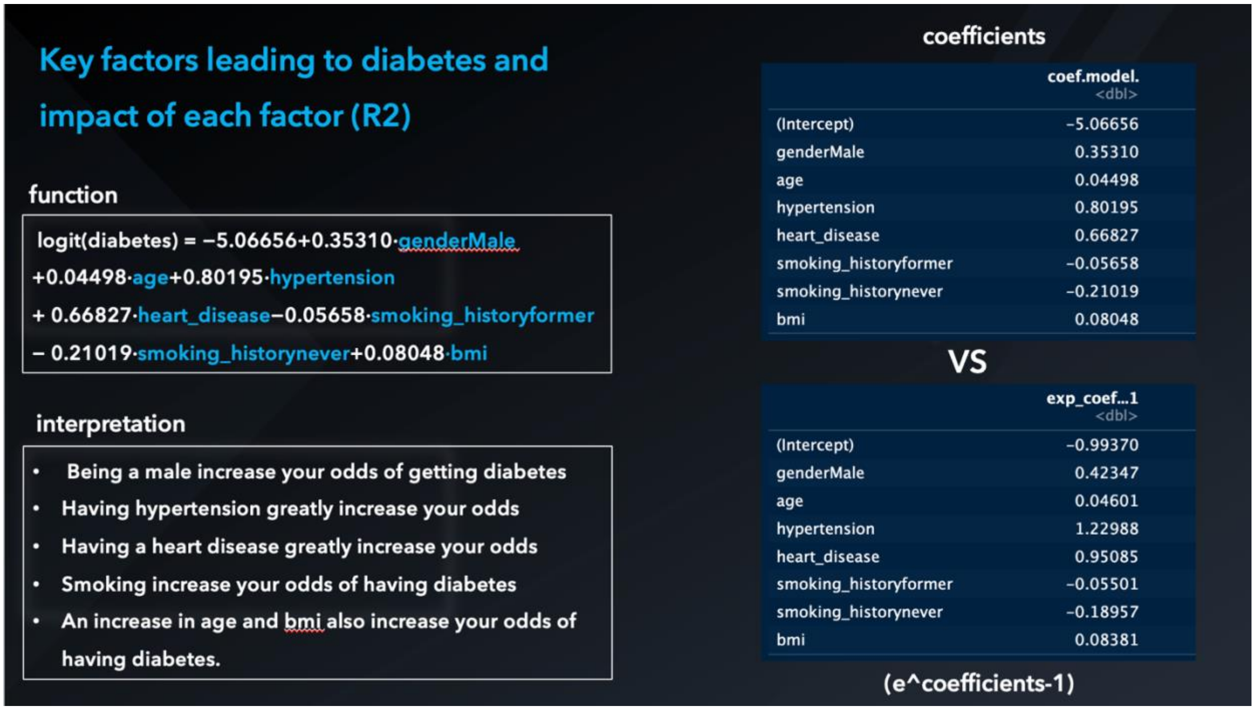
MODELS VALIDATION - TEST SET (30%)

```
----------------------------
 Precision   Recall     F1
----------- -------- --------
  0.8984     0.9909   0.9424
----------------------------
```

```
----------------------------
Precision   Recall     F1
----------- -------- --------
  0.9616     0.7127   0.8186
----------------------------
```

```
----------------------------
 Precision   Recall     F1
----------- -------- --------
  0.9629     0.6969   0.8086
----------------------------
```

Figure 16: Models' performance on the testing set.



Key factors leading to diabetes and impact of each factor (R2)

function

logit(diabetes) = −5.06656+0.35310·genderMale +0.04498·age+0.80195·hypertension + 0.66827·heart_disease−0.05658·smoking_historyformer − 0.21019·smoking_historynever+0.08048·bmi

interpretation

- Being a male increase your odds of getting diabetes
- Having hypertension greatly increase your odds
- Having a heart disease greatly increase your odds
- Smoking increase your odds of having diabetes
- An increase in age and bmi also increase your odds of having diabetes.

coefficients

	coef.model. <dbl>
(Intercept)	-5.06656
genderMale	0.35310
age	0.04498
hypertension	0.80195
heart_disease	0.66827
smoking_historyformer	-0.05658
smoking_historynever	-0.21019
bmi	0.08048

VS

	exp_coef...1 <dbl>
(Intercept)	-0.99370
genderMale	0.42347
age	0.04601
hypertension	1.22988
heart_disease	0.95085
smoking_historyformer	-0.05501
smoking_historynever	-0.18957
bmi	0.08381

(e^coefficients-1)

Figure 17: Explanation of the model

4. Conclusion

In conclusion, this project successfully demonstrated the use of machine learning techniques, specifically logistic regression, in predicting the likelihood of diabetes and identifying key risk factors. Through exploratory data analysis, we found significant relationships between diabetes and factors such as hypertension, heart disease, age, gender, smoking history, and BMI. The model built on an oversampled dataset yielded the best performance, providing valuable insights into the predictors of diabetes. Additionally, the study highlighted the importance of lifestyle changes, such as regular exercise, healthy eating, and smoking cessation, in preventing diabetes. These findings can serve as a foundation for future research and interventions aimed at reducing the prevalence of this chronic disease.