**Diabetes analysis**

By

Leopoldo Rojo Romero (Student ID: 2023355)

Higher Diploma in Science in Data Analytics for Business

Data Preparation & Machine Learning

David McQuaid & Muhammad Iqbal

CCT College

Dublin, Ireland

**Table of contents**

[**Introduction** 2](#_Toc155124982)

[**Problem definition** 3](#_Toc155124983)

[**Data source** 3](#_Toc155124984)

[**Ethical considerations** 3](#_Toc155124985)

[**Rationale and data exploration** 3](#_Toc155124986)

[**Machine Learning model** 12](#_Toc155124987)

[**Github** 16](#_Toc155124988)

[**References** 17](#_Toc155124989)

# **Introduction**

In the present report, we seek to explore and analyse a dataset containing data from individuals that were studied to see if they have diabetes disease or not, or if the individual has a predisposition to develop diabetes.

Our analysis will be focused on data from the United States of America (USA) with data gathered by the Centers for Disease Control and Prevention (CDC). According to the American Diabetes Association (ADA), in 2021, 38.4 million Americans (11.6% of the population) has diabetes, in addition, every year 1.2 million Americans are being diagnosed with diabetes.

Mapa

Descripción generada automáticamente

Figure 1: County level prevalence of diagnosed diabetes, adults aged 20 years or older, USA 2004 and 2021.

Types of diabetes:

**Type 1 diabetes:** The body attacks itself by mistake, stopping the body from making insulin, approximately 5% to 10% of diabetic individuals have type 1 diabetes.

**Type 2 diabetes:** The body does not use insulin well and can not keep blood sugar at normal levels, approximately 90% to 95% of diabetic individuals have type 2 diabetes.

**Gestational diabetes:** This is developed in pregnant women who have never had diabetes, after the baby is born, gestational diabetes usually goes away, but it increases the risk of developing type 2 diabetes later in life for both the mother and the baby.

We believe that this analysis is of high importance owing to the fact that diabetes was the eighth leading cause of death in the USA in 2021, this comes with high medical expenditures among individuals with diabetes, being 2.6 times higher than what expenditures would be in the absence of diabetes (American Diabetes Association, 2023). Also, in the USA, around 98 million adults (more than 1 in 3 adults) have prediabetes and more than 8 in 10 of them do not know about they having it.

Through exploratory data analysis, we aim to find if there is any correlation between major factors correlated with diabetes. Our analysis aims to predict based on independent features (factors correlated with diabetes) if an individual has diabetes or not, or if the individual has a predisposition to develop diabetes.

# **Problem definition**

As previously mentioned, the chronic disease that we are addressing in this analysis is diabetes.

Diabetes occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood glucose; raised blood glucose or raised blood sugar, is a regular sequel of uncontrolled diabetes and over time leads to serious damage to many of the body’s systems, mainly the nerves and blood vessels (World Health Organization, 2023).

Key facts (World Health Organization, 2023):

* Diabetes is a major cause of blindness, kidney failure. Heart attacks, stroke and lower limb amputation.
* In 2019, diabetes and kidney disease due to diabetes caused an estimated 2 million deaths worldwide.

We believe that the analysis of this chronic disease is of high significance due to the major consequences of not treating and / or preventing it, according to the WHO, diabetes can be treated, and its consequences avoided or delayed with diet, physical activity and medication (World Health Organization, 2023).

# **Data source**

We will utilize the NHANES dataset, it was created to assess the health and nutritional status of adults and children in the United States. This dataset is funded by the Centers for Disease Control and Prevention (CDC), specifically through its National Center for Health Statistics (NCHS). The data is based on the survey respondents throughout the United States. Data was gathered through interviews, physical examinations, and laboratory tests.

The dataset is publicly available in the following link:

[National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+(nhanes)

# **Ethical considerations**

We considered the social importance of this project topic, as this can influence public action or inactions, for example people deciding to take on an extreme diet or physical activity due to the fear of developing diabetes.

After consideration, we do not seek to make any medical conclusions and / or recommendations, the objective of this analysis is just to analyse and report our findings for capstone research purpose only.

Also, to prevent breach of anonymity of population from which the data was gathered, the dataset that we will utilize does not have any sensitive data.

# **Rationale and data exploration**

We will approach this case aiming to fit a Random Forest Classifier Machine Learning model, due to its multiple benefits like low risk of overfitting, easy to determine feature importance and not needing considerable pre-processing work, to name a few; this model tackles classification problems and can handle complex datasets.

Let’s have a first look at our data, stored in the ‘df’ variable:

Tabla

Descripción generada automáticamente

Figure 2: NHANES dataset from the CDC.

***Dataset dictionary***

SEQN: ID, Respondent Sequence Number.

age\_group: Respondent’s Age Group (senior (65 or older) / non-senior).

RIDAGEYR: Respondent’s Age.

RIAGENDR: Respondent’s Gender, a 1 represents Male and 2 represents Female.

PAQ605: A 1 represents that the respondent takes part in weekly moderate or vigorous intensity physical activity and a 2 represents that they do not.

BMXBMI: Respondent’s Body Mass Index.

LBXGLU: Respondent’s Blood Glucose after fasting.

DIQ010: If the Respondent’s is diabetic, 1 represents that they are, 2 represents that they are not and a 3 represents that they are about to be.

LBXGLT: Respondent’s Glucose value.

LBXIN: Respondent’s Blood Insulin Levels.

This data set has 2278 observations (rows) and 10 features (columns), we now need to search for null values in our data due to that could make noise and errors when we fit our machine learning model.

Tabla

Descripción generada automáticamente

Figure 3: Percentage of missing data on each feature.

We can observe that the dataset does not has null values from the above ‘isnull()’ method.

Following that we know that we do not have null values, we require to manipulate our dataset in a way that will enable us to further explore and analyse the data.

To help us visualize our data, we will perform feature engineering by creating new columns with the data of our current ones, these new columns will hold the following data:

* 'gender' holds 'RIAGENDR'
* 'exercise' holds 'PAQ605'
* 'senior' holds 'age\_group'
* 'diabetic' holds 'DIQ010' - **our target feature**

Imagen que contiene Tabla

Descripción generada automáticamente

Figure 4: Creating new features that holds data from existing ones.

After creating our new features, we can now remove the previous ones, alongside with the ‘SEQN’ and ‘RIDAGEYR’, ‘SEQN’ due to its data does not add any value to our analysis being just ID’s and ‘RIDAGEYR’ holds the ‘age\_group’ data in a numeric form.

Texto

Descripción generada automáticamente con confianza media

Figure 5: Removing features that we no longer require.

To easily recognise each feature and for visualization purposes, we will rename our features.

Tabla

Descripción generada automáticamente

Figure 6: Renamed dataset.

Now that we have our dataset after feature engineering, we require to visualise our data to further explore and obtain more information that graphs can give us.

Imagen que contiene Gráfico de dispersión

Descripción generada automáticamente

Figure 7: Visualization of our data.

From the above pairplot, we can notice some key factors:

* We do not have a clear pattern
* Imbalanced data (looking at the last plot in the bottom right)
* Skewed data

Following our pairplot, we need to confirm the key factors that we found, this information will define our next steps in our data preparation phase.

Gráfico, Histograma

Descripción generada automáticamente

Figure 8: Visualization to confirm if we have balanced data or not.

As we can observe, we have an imbalanced data, being 96.5% of the data from non diabetic individuals, consequently, for our Machine Learning model we will need to use the balanced random forest classifier that works better than a normal random forest classifier for imbalanced data, this finding is extremely important for the reason that with so few diabetic individuals or individuals with predisposition to diabetes, our machine learning model will spend most of its time learning on non diabetic individuals and will not learn enough from diabetic individuals or individuals with predisposition to diabetes.

Thereafter, we will visualise the data distribution of our body mass index, glucose after fasting, glucose value and insulin level features to confirm if we have skewed data or not.

* Mean value represented by a vertical red line.
* Median value represented by a vertical yellow line.

Gráfico, Histograma

Descripción generada automáticamente

Figure 9: Visualization of our 'body\_mass\_index' feature.

Gráfico, Histograma

Descripción generada automáticamente

Figure 10: Visualization of our 'glucose\_after\_fasting' feature.

Gráfico, Histograma

Descripción generada automáticamente

Figure 11: Visualization of our 'glucose\_value' feature.

Gráfico, Histograma

Descripción generada automáticamente

Figure 12: Visualization of our 'insulin\_level' feature.

As shown, we are able to say that we have right-skewed (positive skewed) data, since our median value is on the left side of our mean value, therefore, we will scale our data using min-max scaler to bring all of our values into a 0 to 1 range, we perform a scaling technique due to skew data can negatively impact the statistical performance of a machine learning model.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figure 13: Scaled dataset.

Succeeding the scaling of our dataset, we will explore the possibility of dimensional reduction; knowing that at this stage of our analysis we only have 7 independent features and 2278 observations, we might not need to perform dimensional reduction due to the fact that it aids with data compression and analysing complex data, these benefits are of huge advantage when we work with massive data sets.

Let’s explore the possibility of using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA):

PCA:

If we have 7 independent features, how many independent features do we need to preserve the 90% and 100% of the pattern in our data?

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 14: Number of components needed to preserve the pattern in our data.

As we can see, we would need our 7 features to preserve 100% of our pattern in our data and only 2 features to preserve above 90% of our pattern in our data.

Let’s observe how well PCA can find the pattern that divide if an individual has diabetes or not, or if an individual has a predisposition of developing diabetes.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 15: PCA scatter plot.

LDA:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 16: LDA scatter plot.

As we can notice, LDA works better than PCA to solve multi class classification problems, but it is not working effectively, this could be due to our imbalanced data, since we have 96.5% of the data from non diabetic individuals and only few observations of diabetic individuals and individuals with predisposition to develop diabetes.

# **Machine Learning model**

As previously mentioned after noticing that we have imbalance data, we will use Balanced Random Forest Classifier from imbalanced learn as our machine learning model for this analysis.

To be able to obtain the best results possible for our dataset, we will use a tool called Grid Search CV from sklearn, this tool will help us to iterate multiple values of hyperparameters of our model searching for the best combination, we will be measuring them using Precision and Accuracy scores, when bout values are at their highest score, those values will become the best parameters, the reason of using a combination of scores to measure our parameters is that accuracy by itself does not do well when working with imbalanced data like ours.

In addition, our scores will be gotten with Stratified K fold, that is a cross-validation technique to ensure the authenticity of those scores.

**Precision:** Percentage of correct positive predictions (true positives) out of the total positive predictions.

**Accuracy:** Percentage of all observations that are correctly classified.

**K fold cross validation:** Randomly divides the dataset into k groups (folds), repeating this task k times using different set of values each time.

For the sake of science, we will train our model in 3 different proportions of our data to compare them, the results were the following:

Grid Search CV:

Best Hyperparameters (**90% training**): {'max\_depth': 3, 'max\_features': 'sqrt', 'max\_leaf\_nodes': 6, 'min\_samples\_leaf': 3, 'n\_estimators': 20, 'random\_state': 4, 'replacement': False, 'sampling\_strategy': 'auto'}

Best Score (**90% training**): 0.741951219512195

Best Hyperparameters (**80% training**): {'max\_depth': 3, 'max\_features': None, 'max\_leaf\_nodes': 3, 'min\_samples\_leaf': 3, 'n\_estimators': 30, 'random\_state': 4, 'replacement': False, 'sampling\_strategy': 'all'}

Best Score (**80% training**): 0.6992533493903357

Best Hyperparameters (**70% training**): {'max\_depth': 3, 'max\_features': 'sqrt', 'max\_leaf\_nodes': 3, 'min\_samples\_leaf': 4, 'n\_estimators': 20, 'random\_state': 4, 'replacement': False, 'sampling\_strategy': 'all'}

Best Score (**70% training**): 0.7603359555213817

After obtaining our parameters for each one of our machine learning models, we require to compare their performance to select the best model and carry out further analysis.

90% training – 10% testing:

* Accuracy (**training**): 0.71
* Precision (**training**): 0.71
* Accuracy (**test**): 0.75
* Precision (**test**): 0.75
* Cross validation score: 0.70

80% training – 20% testing:

* Accuracy (**training**): 0.78
* Precision (**training**): 0.78
* Accuracy (**test**): 0.82
* Precision (**test**): 0.82
* Cross validation score: 0.78

70% training – 30% testing:

* Accuracy (**training**): 0.62
* Precision (**training**): 0.62
* Accuracy (**test**): 0.67
* Precision (**test**): 0.67
* Cross validation score: 0.73

As we can observe, our 80% training - 20% testing machine learning model is the one that is more consistent throughout every score, thus, we will take a closer look to it, observing its predictions with the help of a classification report and a confusion matrix and what features are more relevant for the model when predicting if an individual has diabetes or not, or if the individual has predisposition to develop diabetes.

Classification report features:

**Precision:** Percentage of correct positive predictions (true positives) out of the total positive predictions.

**Recall:** Percentage of correct positive predictions (true positives) out of the actual positives.

**F1 Score:** Harmonic mean of precision and recall, the closer to 1, the better the performance of our model.

**Support:** These values are the observations that belong to each class, 0, 1 and 2 for this analysis.

Tabla

Descripción generada automáticamente

Figure 17: Classification report of our 80% training model.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figure 18: Confusion matrix of our 80% training model.

In our classification report we can appreciate the imbalance data in our dataset, since our model is predicting effectively non diabetic individuals and we can visually notice this in our confusion matrix.

80% training – 20% testing:

* body mass index: 0.093
* glucose after fasting: 0.233
* glucose value: 0.466
* insulin level: 0.115
* gender: 0.0
* exercise: 0.0
* senior: 0.092

We can observe that our glucose value feature is the main determining factor in our data, nevertheless, knowing that diabetes is caused by multiple factors, we are not able to fully conclude that this feature is the most important one due to we only have 7 independent features (attributes) , however, we understand why our model is taking our glucose value and insulin level features as the main features to predict whether an individual has diabetes or not, or if an individual has a predisposition to develop diabetes.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 19: Feature importance of our 80% training model.

**Conclusions**

For this analysis, we can say that we would need more data from diabetic individuals to be able to train our model with enough data to effectively predict individuals with diabetes from independent features.

As shown in *fig 19*, glucose and insulin are the more important features when predicting if an individual has diabetes or not, whilst gender and exercise are irrelevant in this case scenario, this result is quite interesting, knowing that one of the key facts of treating or avoiding diabetes is physical activity, however, this could be happening due to our imbalanced data and it might gain more relevance as we get more data of diabetic individuals.

According to the Centers for Disease Control and Prevention (CDC), there are more features that we could analyse to be able to improve the performance of our machine learning model in regards to predicting if an individual has diabetes or not, or if an individual has a predisposition to develop diabetes.

(assuming that we got enough data):

* Urinate a lot, often at night
* Are very thirsty
* Lose weight without trying
* Are very hungry
* Have blurry vision
* Have numb or tingling hands or feet
* Feel very tired
* Have very dry skin
* Have sores that heal slowly
* Have more infections than usual

If we could obtain more data with more features of individuals with diabetes and with predisposition to diabetes, we would be able to train our model more effectively in regards to predicting if an individual has diabetes or not, or if an individual has a predisposition to develop diabetes, further analyse the trends in the data and observe how feature importance would change in our machine learning model.

# **Github**

<https://github.com/LeopoldoCCT/HDIP_DPrep_ML_CA2>

# **References**

American Diabetes Association (2023). *Statistics About Diabetes | ADA*. [online] diabetes.org. Available at: https://diabetes.org/about-diabetes/statistics/about-diabetes [Accessed 2 Jan. 2024].

Centers for Disease Control and Prevention (2023). *Diabetes Symptoms*. [online] CDC. Available at: https://www.cdc.gov/diabetes/basics/symptoms.html.

imbalanced-learn (n.d.). *BalancedRandomForestClassifier — Version 0.10.1*. [online] imbalanced-learn.org. Available at: https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedRandomForestClassifier.html.

Indeed (2023). *What Is Skewed Data in Statistics? (With Definition and Example)*. [online] Indeed Career Guide. Available at: https://www.indeed.com/career-advice/career-development/skewed-data.

scikit-learn (2019). *sklearn.model\_selection.GridSearchCV — scikit-learn 0.22 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html.

World Health Organization (2023). *Diabetes*. [online] World Health Organisation. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed 1 Jan. 2024].

Zach (2022a). *How to Interpret the Classification Report in sklearn (With Example)*. [online] Statology. Available at: https://www.statology.org/sklearn-classification-report/.

Zach (2022b). *What is a ‘Good’ Accuracy for Machine Learning Models?* [online] Statology. Available at: https://www.statology.org/good-accuracy-machine-learning/.