Shape

Description automatically generated with medium confidence

**Diabetes Diagnosis Classification**

**(Data Mining Project)**

Master’s degree in Data Science & Engineering – FEUP

Introduction to Machine Learning and Data Mining

December 2022

Authors: Ian Karkles and Sónia Ferreira

A picture containing graphical user interface

Description automatically generated

# Table of Contents

[I. Table of Contents 2](#_Toc121127847)

[II. Introduction 3](#_Toc121127848)

[III. Materials and Methods 5](#_Toc121127849)

[IV. Data Source 5](#_Toc121127850)

[V. What is the CRISP-DM methodology? 6](#_Toc121127851)

[What are the six CRISP-DM phases? 6](#_Toc121127852)

[CRISP-DM phases & Diabetes Diagnosis Classification Mining Project phase 7](#_Toc121127853)

[VI. Business understanding (CRISP-DM - phase 1) 9](#_Toc121127854)

[VII. Data understanding (CRISP-DM - phase 2) 9](#_Toc121127855)

[VIII. Data preparation (CRISP-DM - phase 3) 9](#_Toc121127856)

[IX. Modeling (CRISP-DM - phase 4) 9](#_Toc121127857)

[X. Evaluation (CRISP-DM - phase 5) 10](#_Toc121127858)

[XI. Conclusions 10](#_Toc121127859)

[XII. Future Work 10](#_Toc121127860)

[XIII. Bibliography 10](#_Toc121127861)

# Introduction

*“A correct diagnosis is a three-fourths the remedy”* – Mahatma Gandi

Diabetes is considered one of the deadliest and most chronic diseases, either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces, which in turn makes the carbohydrate metabolism abnormal and raises the levels of glucose in the blood. In diabetes, a person generally suffers from high blood sugar. Intensify thirst, hunger, and frequent urination are some of the symptoms of the disease (Vijayan, V.V., Anjali, C., ).

Diabetes diagnosis is not only affected by several factors like height, weight, hereditary factors, and insulin, but the major reason considered is sugar concentration among all factors. The measure of sugar substances cannot be controlled, and this is what makes the process of identifying the disease morose and tedious.

In 2014, 8,5% of adults over 18 years old had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years. In addition, 460 000 kidney disease deaths were caused by diabetes and raised blood causes around 20% of cardiovascular deaths. Between 2000 and 2019, there was a 3% increase in age-standardized mortality rates from diabetes (WHO).

Many complications, like blindness, kidney failure, heart attacks, strokes, and lower limb amputation can occur if diabetes remains untreated or unidentified. A healthy diet combined with regular exercise, plus maintaining a normal body weight are some of the ways to prevent or delay to have diabetes. Consequently, early identification is the only remedy to stay away from complications. However, without the proper diagnosis, there isn’t much that can be done, and the occurrence of diabetes has been rising more rapidly in low-and middle-income countries than in high-income countries (WHO).

The process to identify the disease is wearisome and costly, as it requires the patient to visit a doctor or diagnostic center and run several exams. And it's here that the growth of Machine Learning approaches can solve this critical problem since it would not require a patient’s visit to a doctor or diagnosis center or to run several exams. Plus, a Machine Learning process could be applied worldwide at a cheaper cost than the traditional diagnosis of the disease.

Today, many researchers are conducting experiments into diagnosing diseases using various classification algorithms of machine learning approaches, since this type of algorithm was proven to work better for these cases. Classification strategies are broadly used for classifying data into different classes according to some constraints compared to an individual classifier.

Data mining and machine learning algorithms gain their strength due to their capabilities to manage large amounts of data, combine data from different sources, and integrate background information into the study.

Therefore, this research is performed on Pima Indians Diabetes Database (PIDD) which was sourced from “[Kaggle:](https://www.google.com/search?q=kaggle&oq=kaggle+&aqs=chrome..69i57j0i512j0i67j69i65j69i60l4.1656j0j7&sourceid=chrome&ie=UTF-8) Your Machine Learning and Data Science Community”. This database focuses on pregnant women, older than 21 years, and suffering from diabetes. Several machine learning classification algorithms, (e.g.: Decision Tree, KNN, and AdaBoost) are used to find the prediction of diabetes in a patient. The performance of all algorithms is then evaluated on various measures like precision, accuracy, f-measure, and recall. And these results are then verified using the Receiver Operating Characteristic (ROC) curves properly and systematically.

The remaining of this report is organized as follows: Section III briefs the materials and method used, Section IV introduces the data source used and its origin, Section V gives a summary of what CRISP-DM is and how it’s implemented in this research (its relation between report and technical solution), Section VI focus on understanding the research and project objectives and requirements from a business point of view, Section VII explores the raw data of the dataset, Section VIII details all activities to construct the final dataset from the raw data, Section IX will approach and discuss the evaluations done on the modeling techniques, Section X discusses and evaluates the results, Section XI determines the Conclusion of the research work and Section XII approach’s Future Work that can be done in order continue or improve this research.

# Materials and Methods

For this data mining project, the methodology used to approach its execution was CRISP-DM (phases one to five will be described in detail at the core of this report).

In this data mining project, Python programming language was used through Jupyter Notebook.

# Data Source

The present data mining project will use a dataset to diagnose diabetes, that was collected from the website “[Kaggle:](https://www.google.com/search?q=kaggle&oq=kaggle+&aqs=chrome..69i57j0i512j0i67j69i65j69i60l4.1656j0j7&sourceid=chrome&ie=UTF-8) Your Machine Learning and Data Science Community”.

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and its objective are to predict whether a patient has diabetes, based on certain diagnostic measurements that are included in the dataset. This dataset has the following constraints in place:

* All patients are female.
* They are at least 21 years old.
* They are of Pima Indian heritage.2.

# What is the CRISP-DM methodology?

CRISP-DM which stands for **Cross** **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining is an industry-proven modeling process that serves as a basis for data sciences processes that can be applied in any type of business without the dependency on software or service to be executed.

Published in 1999 to standardize data mining processes across industries, it has since become the most common methodology for data mining, analytics, and data science projects.

CRISP-DM consists of six sequential phases, to conceive a Data Mining project and these phases can have cycle iterations according to the developer's needs. The six sequential phases are:

1. Business understanding – What does the business need?
2. Data understanding – What data do we have/need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

## What are the six CRISP-DM phases?

The first phase, **Business Understanding** focuses on understanding the project objectives and requirements from a business perspective. The analyst formulates this knowledge as a data mining problem and develops a preliminary plan.

The second phase is **Data Understanding,** and its objective is to know what can be expected and achieved from the data. It checks the quality of the data, in several terms, such as data completeness, values distributions, and data governance compliance. In this phase, the analyst might also detect interesting subsets to form hypotheses for hidden information.

This phase is a crucial part of the project because it defines how viable and trustworthy can be the result. It can be necessary to step back and understand the business point of view and how that information can be beneficial.

The third phase, **Data Preparation** and involves the ELTs or ETLs process that will cover all activities to construct the final dataset, based on the initial raw data (select, clean, construct, integrate, and format/re-format data).

Sometimes data governance policies are not respected or set in on the organizations, human errors happen, and to give meaning to data, it becomes necessary to standardize the information. Likewise, some algorithms perform better under certain parameters, some do not accept numerical or non-numerical values or large values, then again is necessary to normalize this data.

Data Preparation is, therefore, one of the most phases of a data mining project and the step that consumes more time.

The fourth phase, **Modelling** is the core phase of any machine learning project. This phase will be responsible for the results that should satisfy or help satisfy the project goals.

This part of the project should however be the shortest phase of the project, as if everything previously is done correctly, there is little to adjust. In the cases that the results are improvable, the methodology is set to step back to data preparation and improve the available data.

The fifth phase, **Evaluation**, looks more broadly at which model best meets the business and what to do next. Here the analyst will evaluate if the results are valid and correct. If determined that the results are wrong, the methodology allows the review back to the first step, to understand why the results are incorrect.

The sixth phase (and last phase) is **Deployment** and depending on the requirements, the deployment phase can be as simple as creating a report or as complex as implementing a repeatable data mining process across the enterprise (CRISP-DM).

## CRISP-DM phases & Diabetes Diagnosis Classification Mining Project phases

The table below compares the CRISP-DM phases and the technical steps implemented in this data mining project.

|  |  |
| --- | --- |
| CRISP-DM Phases | Technical Project Phases |
| 1. Business understanding | 1. Understand and set goals from a business perspective for the project (Jupiter notebook chapter: Business Context / Data Context) |
| 1. Data understanding | 1. Check if the available data can meet the objectives of the project and its quality (Jupiter notebook chapter: Exploratory Data Analysis - EDA) |
| 1. Data Preparation | 1. The raw data is transformed, in the cases that are necessary for this data mining project (Jupiter notebook chapter: Pre-processing, Feature Engineering, Feature Selection) |
| 1. Modeling | 1. Execute the algorithms that satisfied the project objectives (Jupiter notebook chapter: Models) |
| 1. Evaluation | 1. The results are presented, analyzed, and evaluated (Jupiter notebook chapter: Evaluation) |
| 1. Deployment[[1]](#footnote-1) | 1. Not Applicable1 |

# Business Understanding (CRISP-DM - phase 1)

Any good project starts with a deep understanding of the customer’s needs. Data mining projects are no exception and CRISP-DM recognizes this. Thus, the first phase of this project focused on understanding the current context of the topic in the study: Diabetes disease, and understanding the project objectives and requirements from a business perspective.

Diabetes is a worldwide killer and ranks among the top causes of premature deaths. More than 500 million adults (20 – 79 years) are living with diabetes (nearly one in ten) and it's predicted that these numbers will keep ascending in the coming years. By 2030 is estimated that more than 600 million will suffer from diabetes and by 2045 more than 783 million (IDF).

Between 2000 and 2019, diabetes mortality rates by age increased by 3% (WHO). And just in 2021, diabetes was responsible for more than 6 million deaths (one every five seconds) (IDF).

A very concerning point is that three out of four people leaving with this disease, live in low- and middle-income countries that do not have the resources or money to map and treat rightfully this disease, which may affect the daily life of a person who is not aware of this illness. If diabetes is diagnosed correctly and timely, it can be treated, and its consequences are avoided or delayed with diet, physical activity, medication, and regular screening and treatment for complications (WHO).

Diabetes caused at least 966 billion dollars (approximately more than 900 million euros) in health expenditure – a 316% increase over the last 15 years (IDF).

There are several reasons, for the lack of the right diagnosis, including that the process is expensive and tedious as it includes several visits of the patient to a doctor or a diagnostic center, plus doing some exams. But one of the main reasons is the lack of technical skills to interpret the results of exams, especially in low and middle-income countries.

Nevertheless, if a machine learning approach could solve or improve this critical problem of diagnosing when a patient has or is prone to have diabetes, this would represent a win globally in the quality of health service and diagnosis, especially for low- and middle-income countries where the problem is more persistent, and they have less money and resources to invest. Another win, less related to the patient of diabetes would be that the resources (especially money) that are today spent to diagnose, prevent, and treat diabetes disease would be needed less and therefore could be allocated to other problems.

In this way, the main goal of this experiment with a classifier is to guide especially low and medium-income health professionals to classify a Diabetes case with simple exams in a person who most probably would be unaware of the disease until would be too late.

In the context of Data Classification, multiple classification methods can help to solve this type of problem and will be used in this experiment. For this type of project, the dataset was collected from the website “[Kaggle:](https://www.google.com/search?q=kaggle&oq=kaggle+&aqs=chrome..69i57j0i512j0i67j69i65j69i60l4.1656j0j7&sourceid=chrome&ie=UTF-8) Your Machine Learning and Data Science Community” and the experiments are performed on the [Pima Indians Diabetes Database](https://www.kaggle.com/uciml/pima-indians-diabetes-database), originally from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset has the following constraints in place:

* All patients are female.
* They are at least 21 years old.
* They are of Pima Indian heritage.2.

The dataset contains multiple features, described in the table below, that can help the classifier model to rightfully predict diabetes in a patient.

Table 1 - Dataset features

|  |  |
| --- | --- |
| Columns (features) | Description |
| Pregnancies | Number of times pregnant |
| Glucose | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| Blood pressure | Diastolic blood pressure (mm/Hg) |
| Skin thickness | The 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 |
| Age | Age of the Patient (in years) |
| Outcome | Class variable |

# Data Understanding (CRISP-DM - phase 2)

During phase 1 (Business Understanding), several questions related to the data that would be available in the dataset arose, therefore in this section (phase 2 – Data Understanding), the data is going to be analyzed and visualized for easy understanding and followed with appropriate statistical analysis for empirical findings within the data. And the following Business questions will be answered within this phase:

* From the total of patients in this study how many are diabetic and non-diabetic?
* Is there a difference in the average age of patients with diabetes and without?
* Do diabetic and non-diabetic patients have similar levels of glucose, insulin, blood pressure, and BMI?
* Are the skin thickness and the diabetes pedigree function similar between patients with and without diabetes?

Pima Indians Diabetes Dataset (PIDD), comprises 768 observations and 9 characteristics (features)[[2]](#footnote-2).

Out of which one is a dependent variable, and the remaining are independent:

* Dependent variable: outcome.
* Independent variable: all remaining features.

Apart from BMI and Diabetes Pedigree Function which are of type float, all remaining features are of type int64[[3]](#footnote-3).

The below figure shows a small extract of the raw dataset, where it can be observed the type of data and structure that is available in the dataset.

Table

Description automatically generated

Figure - Raw Dataset (extract)

The following conclusions were taken from analyzing the dataset:

* no missing values[[4]](#footnote-4) are present, as can be observed in figure 2.
* No redundant data exist[[5]](#footnote-5).
* As can be observed in figure 3, some parameters (features) have zero values, and they can be considered as NaN values classified as zero, for example in Skin Thickness (that represents the triceps skin fold thickness in mm) and Glucose.

Each feature was checked individually, to understand better the values present in each and see if there was some justifiable reason for the zeroes[[6]](#footnote-6).

* Outliers were detected in the dataset with the use of Z-Score and Winsorization. Both tests pointed out outliers and the results can be observed in figure 5 and figure 6, respectively, for Z-Score and Winsor.

Z-Score test was conducted with 3 standard deviations from the mean

* s

Chart

Description automatically generated

Figure - Dataframe Missing Values[[7]](#footnote-7)

Chart, box and whisker chart

Description automatically generated

Figure - Plots of individual variables of the raw dataset

Chart, diagram

Description automatically generated with medium confidence

Figure - Plots of scatter plots of individual variables of the raw dataset

Text, letter

Description automatically generated

Figure - Outliers: Z-Score result

Graphical user interface, text, application, email

Description automatically generated

Figure - Outliers: Winsorization result

Table

Description automatically generated

Figure - Statistics of the dataset

# Data Preparation (CRISP-DM - phase 3)

The data preparation phase covers all activities to construct the final dataset from the initial raw data.

# Modeling (CRISP-DM - phase 4)

The analyst evaluates, selects & applies the appropriate modeling techniques. Since some techniques like neural nets have specific requirements regarding the form of the data. There can be a loop back here to data prep.

In the modeling for this problem, multiple classification models were selected and compared in the foundation of what could be a Data Pipeline, with the generation of the model mainly through libraries in python like Sklearn and XGBoost as a Standalone Library of the process. Although was thought about the possibility to add Neural Networks in this pipelined test of models, due to the sake of the process and the project itself, the classification models used for this process were mainly traditional Machine Learning classifiers that were used for the process.

The classifiers used in this model were:

* Logistic Regression
* Linear Discriminant Analysis
* K-Neighbors Classifier
* Decision Tree Classifier
* Gaussian NB
* Support Vector Machine Classifier
* Adaboost Classifier
* Gradient Boosting Classifier
* Random Forest Classifier
* Extra Tree Classifiers
* Extreme Gradient Boost Classifier (XGBoost)

Of these, the most promising of them were the XGBoost Classifier, due to its track record of success in many machine Learning competition and experiments. However, this was not what exactly happened in the evaluation of each one of the respective models.

From this model, due to the sake of effectivity, all the models were run with their standard parameters.

# Evaluation (CRISP-DM - phase 5)

The analyst builds & chooses models that appear to have high quality based on loss functions that were selected. The analyst then tests them to ensure that they can generalize the models against unseen data. Subsequently, the analyst also validates that the models sufficiently cover all key business issues. The result is the selection of the champion model(s)

In the evaluation of the respective classification models, the metrics chosen by the team were Accuracy, F1-Score, and ROC/AUC Curves

The F1-Score is a measure of a model's accuracy that considers both the precision and the recall of the model. Precision is the proportion of correct positive predictions made by the model, while recall is the proportion of actual positive cases that the model was able to identify. The F1-Score is calculated as the harmonic mean of the precision and recall and is a useful metric for comparing the performance of different models.

The ROC, or receiver operating characteristic, is a plot that shows the relationship between the false positive rate and the true positive rate of a classification model. The false positive rate is the proportion of negative cases that are incorrectly classified as positive, while the true positive rate is the proportion of positive cases that are correctly classified as such. The ROC curve is a useful tool for evaluating a model's performance and comparing it to other models.

Finally, the accuracy of a classification model is simply the proportion of data points that the model correctly classifies. This is a commonly used metric, but it can be misleading, especially when the data is imbalanced (i.e., when there are unequal numbers of positive and negative cases). In such cases, the F1-Score and ROC may be more informative measures of a model's performance.

In the case of this specific study, we had a large bunch of models that performed almost the same for most of them. As the following values returned for the respective models in mean accuracy after the 10-fold process and its standard deviation between parentheses.

The accuracies for each one of the models were the following:

|  |  |  |
| --- | --- | --- |
| Model Acronym | Mean Accuracy | Mean Standard Deviation |
| LR | 0.764035 | 0.060710 |
| LDA | 0.764065 | 0.058034 |
| KNN | 0.718784 | 0.087651 |
| CART | 0.692650 | 0.061552 |
| NB | 0.758863 | 0.067307 |
| SVM | 0.769147 | 0.072132 |
| AB | 0.736207 | 0.063703 |
| GBM | 0.741319 | 0.075488 |
| RF | 0.730853 | 0.075273 |
| ET | 0.729310 | 0.070007 |
| XGB | 0.723956 | 0.065987 |

In this result, we can confirm the information returned from the notebook where most of them had a similar result. In other aspects, it’s possible to notice that some models have slightly better results in the outcome of the models. The highlights over here were the following ones:

* The Gradient Boost Classifier was a better classification in a ROC metric (0.757), mostly due to its good tolerance for imbalanced data, and its ability to combine multiple weak classifications into a strong one, which is the case in this dataset, which the strongest one of the classifiers correlate with the Outcome variable of 54%
* The Linear Discriminant Analysis also was a Highlight due to its good accuracy and F1-Score concerning other classifiers (0.7640 and 0.78 respectively). This slightly better accuracy has a foundation in the logic behind the classifier itself. This specific classifier works with a linear method of resolution where you can generate linear methods between the classes and ensemble them afterward. The Classification also works with a Generative model, that stacks multiple linear regressors into a classifier, different than a method like an SVC, that utilizes discriminative bounds to split the classes among the model.

# Conclusions

In this project was possible to evaluate the importance of Data Science and Data Mining in real-world problems and how they could directly affect day-to-day basis world decisions.

In our specific problem for dataset classification based on health and physical parameters, we could get through testing multiple classifiers a reasonably good accuracy of 75% of the cases, mostly in denying a case of Diabetes and then confirming it, due to its imbalanced dataset.

However even if this imbalanced dataset we managed to understand how the preprocessing steps work and are done and how it’s possible to evaluate machine learning models for future problems.

Even with this, the team also can have a more critical view of future classification problems that they face in the future and are possible to counsel and directional the efforts of the team for a better outcome even before the models are applicable.

# Future Work

For future works of this project, one of the main targets would be to enhance each one of the specific models with hyperparameters more suitable for them and mostly create step processes in the notebook where we evaluate the results and outcomes of the models in each step of the pre-processing process, in the intent to understand how each of the steps affects in the outcome and which steps just add complexity for the final output without any gain on accuracy or performance.

# Bibliography

CRISP-DM. n.d. <https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf>.

Vijayan, V.V., Anjali, C., "Prediction and diagnosis of diabetes mellitus A machine learning approach." *IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. 122–127doi:10.1109/RAICS.2015.7488400 vols. 2015.

IDF, International Diabetes Federation. *IDF Diabetes Atlas 10th Edition 2021*. 2021. <https://diabetesatlas.org/>.

WHO. *World Health Organization*. 2022. <https://www.who.int/news-room/fact-sheets/detail/diabetes>.

# Appendix

Text

Description automatically generated

Appendix - Figure 1 - Dataframe Shape

Table

Description automatically generated

Appendix - Figure 2 - Dataframe columns and corresponding data

Table

Description automatically generated with medium confidence

Appendix - Figure - Missing Values

Graphical user interface

Description automatically generated

Appendix - Figure – Missing Values Matrix

A picture containing text

Description automatically generated

Appendix - Figure - Redundant Data

Chart, box and whisker chart

Description automatically generated

Appendix - Figure – Plots of individual variables of the raw dataset

Chart, diagram

Description automatically generated with medium confidence

Appendix - Figure - Plots of scatter plots of individual variables of the raw dataset

Table

Description automatically generated

Appendix - Figure – Glucose 0 values

Table

Description automatically generated

Appendix - Figure – Blood Pressure 0 values

Table

Description automatically generated

Appendix - Figure – Skin Thickness 0 values

Table

Description automatically generated

Appendix - Figure - Insulin 0 values

Graphical user interface

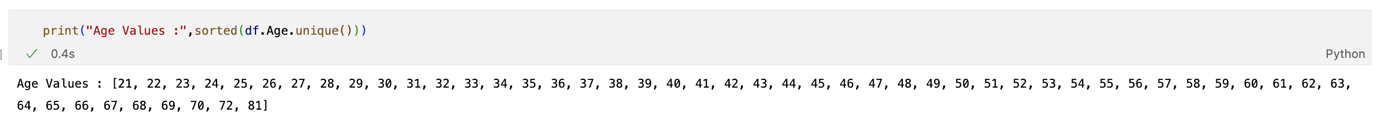
Description automatically generated with medium confidence

Appendix - Figure - BMI 0 values

Text

Description automatically generated

Appendix - Figure – Diabetes Pedigree Function values



Appendix - Figure - Age values

Table

Description automatically generated

Appendix - Figure - Statistics of the dataset

Chart, treemap chart

Description automatically generated

Appendix - Figure - Correlation of the variables

Text

Description automatically generated with medium confidence

Appendix - Figure – Outliers detection (Z-Score and Winsorization)

1. Phase six was not required for the conclusion of this project [↑](#footnote-ref-1)
2. Appendix - Figure 1 - Dataframe Shape [↑](#footnote-ref-2)
3. Appendix - Figure 2 - Dataframe columns and corresponding data [↑](#footnote-ref-3)
4. Appendix - Figure 3 - Missing Values [↑](#footnote-ref-4)
5. Appendix - Figure 5 - Redundant Data [↑](#footnote-ref-5)
6. Appendix - Figure 8 – Glucose 0 values; Appendix - Figure 9 – Blood Pressure 0 values; Appendix - Figure 10 – Skin Thickness 0 values; Appendix - Figure 11 - Insulin 0 values; Appendix - Figure 12 - BMI 0 values; Appendix - Figure 13 – Diabetes Pedrigree Function values; Appendix - Figure 14 - Age values [↑](#footnote-ref-6)
7. Appendix - Figure 4 – Missing Values Matrix [↑](#footnote-ref-7)