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

[Table of Contents 2](#_Toc121082380)

[Introduction 3](#_Toc121082381)

[Materials and Methods 4](#_Toc121082382)

[What is the CRISP-DM methodology? 4](#_Toc121082383)

[What are the six CRISP-DM phases? 4](#_Toc121082384)

[CRISP-DM phases & Diabetes Diagnosis Classification Mining Project phases 5](#_Toc121082385)

[Business understanding (CRISP-DM - phase 1) 7](#_Toc121082386)

[Data understanding (CRISP-DM - phase 2) 7](#_Toc121082387)

[Data preparation (CRISP-DM - phase 3) 7](#_Toc121082388)

[Modeling (CRISP-DM - phase 4) 7](#_Toc121082389)

[Evaluation (CRISP-DM - phase 5) 7](#_Toc121082390)

[Conclusions 8](#_Toc121082391)

[Future Work 8](#_Toc121082392)

[Bibliography 8](#_Toc121082393)

# Introduction

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

Diabetes is considered as one of the deadliest and most chronic diseases, that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces.

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 (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. But without the proper diagnosis, there isn’t much that can be done. Prevalence has been rising more rapidly in low-and middle-income countries than in high-income countries.

But the process to identify the disease is wearisome, 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 this would not require the “visit” of a patient to a doctor or diagnosis center or to run several exams.

Therefore, the purpose of this study is to design a model that can predict if a patient can have diabetes with maximum accuracy. The experiments are 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”.

For the purpose of this experiment, several classification algorithms, (e.g.: Decision Tree, KNN, and AdaBoost) are used. 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 in a proper and systematic manner.

# 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 is 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 **CR**oss **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, to understand the business point of view and how that piece of 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 below table offers a comparison between 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 |
| 1. Data understanding | 1. Check if the available data can meet the objectives of the project and its quality. |
| 1. Data preparation | 1. The raw data is transformed, in the cases that is necessary for this data mining project |
| 1. Modeling | 1. Execute the algorithms that satisfied the project objectives |
| 1. Evaluation | 1. The results are presented, analyzed and evaluated |
| 1. Deployment[[1]](#footnote-1) | 1. Not Applicable1 |

# Business understanding (CRISP-DM - phase 1)

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.

**Any good project starts with a deep understanding of the customer’s needs. Data mining projects are no exception and CRISP-DM recognizes this.**

# Data understanding (CRISP-DM - phase 2)

Starting with initial data collection, the analyst proceeds with activities to get familiar with the data, identify data quality problems & discover first insights into the data. In this phase, the analyst might also detect interesting subsets to form hypotheses for hidden information.

# 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.

# 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)

# Conclusions

# Future Work

# Bibliography

**CRISP-DM** [Online]. - https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf.

**WHO** World Health Organization [Online] // World Health Organization. - https://www.who.int/news-room/fact-sheets/detail/diabetes.

1. Phase six was not required for the conclusion of this project [↑](#footnote-ref-1)