

**UNIVERSIDAD DE MONTERREY
SCHOOL OF ENGINEERING AND TECHNOLOGIES
DEPARTMENT OF ENGINEERING
ARTIFICIAL INTELLIGENCE**

LOGISTIC CLASSIFICATION REPORT

Mauricio A. De León Cárdenas. 505597

M/TH 2:30 P.M

Dr. Andrés Hernández Gutiérrez

San Pedro Garza García, N.L. a 19 de abril de 2020



Introduction:

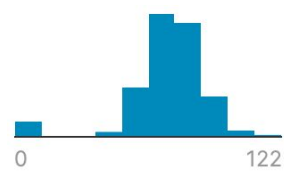
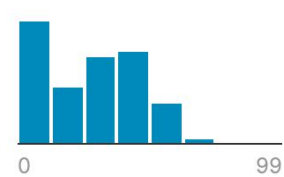
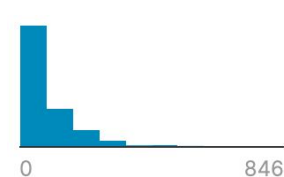
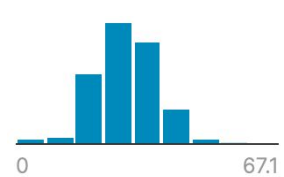
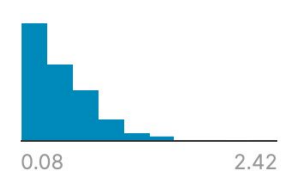

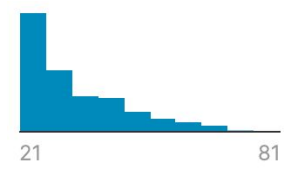
Diabetes is a disease that occurs when the blood glucose count in a person is too high. Approximately, just in the U.S, 30.3 million people have diabetes as of 2015 but nearly 7.2 million are undiagnosed. One reason for this might be that diabetes symptoms can be difficult to identify since these come on slowly and can easily be mistaken individually for other diseases. Not identifying the disease on time could lead to serious health problems such as an increase of probability in strokes, blindness, kidney failure, or even worst, death.

Although diabetes doesn't have a cure, scientists are using machine learning algorithms to help identify easier whether a person has or not the disease. A common algorithm used to classify the condition of a person is using the logistic regression algorithm, this is a statistical method for predicting binary classes (has diabetes | doesn't have diabetes). By nature this algorithm makes use of the sigmoid function which can take any real value number and position it into a value between 0 and 1, this allows the algorithm to predict whether someone has diabetes (1) or not (0).

Data exploration stage:

For the training of the algorithm a dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases was used. This was downloaded from kaggle.com which contains 768 data entries (or patients tested) using the following format:

Label	Description	Distribution	Comments
Pregnancies	Number of times pregnant		This doesn't seem to be a significant data since there seems to be a big tendency to be 0 or almost 0 so it could easily be discarded through feature selection.
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test		No comment

Blood Pressure	Diastolic blood pressure (mm Hg)		No comment
Skin Thickness	Triceps skin fold thickness (mm)		No comment
Insulin	2-Hour serum insulin (mu U/ml)		No comment
BMI	Body Mass Index (weight in kg / (height in m)^2)		No comment
DiabetesPedigreeFunction	Function that represents how likely a patient is to get the disease by extrapolating from their ancestors history		Observing the outcome this data doesn't seem to affect significantly, because patients with low likelihood still get diabetes, so it could easily be discarded through feature selection.
Age	Age of patient in years		No comment
Outcome	Class variables (0 or 1)		No comment

Although two features could be left unused through feature selection, all features were used for the sake of evaluating if the dataset, as it is, would accurately predict whether a person has diabetes or not using logistic regression.

In comparison to other activities done throughout the course, this was the first to involve only one dataset, which meant implementing a function that splits this in two parts, 80% for training and the rest (20%) for testing. Once the data was splitted accordingly, these were scaled using feature scaling, this decision was taken because without it the algorithm would take an immense amount of time to converge, for testing the code was left running in the background training for almost 10 minutes without converging, in comparison, with scaling the complete code takes approximately 1.47 seconds to run.

Performance metrics:

- **Confusion matrix:**

```
-----
Confusion matrix
-----
|                                     | Has diabetes (1): | Doesn't have diabetes (0): |
|-----+-----+-----|
| Has diabetes (1):                | 22 | 6 |
| Doesn't have diabetes (0):       | 32 | 93 |
|-----+-----+-----|
```

- **Accuracy:** 0.7516339869281046
- **Precision:** 0.7857142857142857
- **Recal:** 0.407407407407407
- **Specificity:** 0.9393939393939394
- **F1 Score:** 0.5365853658536585

As it can be seen, even though the results are not as high as expected the algorithm does a good prediction and classification of whether a person has diabetes or not. It is thought that these value could improve if feature selection was done, since the algorithm would discard those that are not representative, improving computational time and calculus.

These results were compared to those obtained in the tutorial from the first reference in the *references* section and these were almost identical which leads to believe that some tuples of the dataset were incorrectly on purpose to improve the

students abilities to detect these and delete them, do feature scaling, feature selection, amongst other methods to improve the results.

References:

- Navlani, A. (2019). *Understanding Logistic Regression in Python*. Retrieved on april 19th 2020 from: <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>
- National Institute of Diabetes and Digestive and Kidney Diseases. (2018). *Pima Indians Diabetes Database*. Retrieved on april 19th 2020 from: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
- Nall, R. (2020). *An overview of diabetes types and treatments*. Retrieved on april 19th 2020 from: <https://www.medicalnewstoday.com/articles/323627>
- World Health Organization. (2018). *Diabetes*. Retrieved on april 19th 2020 from: <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- National Institute of Diabetes and Digestive and Kidney Diseases. (2016). *What is Diabetes?*. Retrieved on april 19th 2020 from: <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>
- Miller, K. (2017). *Nearly a Quarter of People with Diabetes Don't Know They Have It*. Retrieved on april 19th 2020 from: <https://www.self.com/story/undiagnosed-diabetes>

**The code implementation of the datacamp tutorial is included in the same folder where the owner's (Mauricio) code is also included.

I hereby declare that I have worked in this activity with academic integrity