# Capstone Project

**Sebastian Jimenez**

Machine Learning Engineer Nanodegree

May 4, 2021

# Octopus: A Machine Learning Workflow

## Definition

### Proposal

One task of a Machine Learning Engineer is to design and build applications that automate the execution of predictive models. The goal of this project is to implement a machine learning workflow to increase the efficiency in the supervised learning specifically in **classification problems**.

### Problem Statement

When Data Scientists and Data Analysts are tackling a Machine Learning problem, there are some steps that they usually must do. Steps like: exploration data analysis, missing values analysis, outlier detection, correlation analysis, data transformation, feature engineering, choose the best metric for the problem, feature selection, hyperparameter tuning, compare models and finally put the best model in production. These can take a lot of time and it can reduce the time placed in to find insights. Some of these steps can be automated to help them to do their work easier and faster. That's why this project looks to build a workflow for ML projects specifically in **classification problems**. I hope that it brings a good baseline and encourages them to obtain better metrics.

I didn't choose to work directly with a use case, because I think that this Nanodegree is a good opportunity to improve my programming skills instead to work in a Data Science project itself. I want to finish by mentioning the following quote "a range of Auto-ML tools will need to be developed to support varying user goals such as simplicity, reproducibility, and reliability." (Xin, et al. 2008)

### Datasets

The datasets used to test the workflow will be the Titanic[1] and the Diabetes[2] Datasets. The first one is useful to test the workflow in a no-complex problem and the second one adds more complexity. The Titanic dataset has *891* records and *10*

[1] https://www.kaggle.com/c/titanic/data

[2] https://www.kaggle.com/c/diabetes-hospital-readmission/data

features. Its class is relatively balanced with 38% / 62%. For the other side, the Diabetes dataset has 101766 records and 50 features. Its class is imbalanced 9% / 91%, here we will need to use balancing methods.

## Solution Statement

The *Image 1* shows the solution proposal. This is until suggesting a model, putting the model in production is out of this scope.

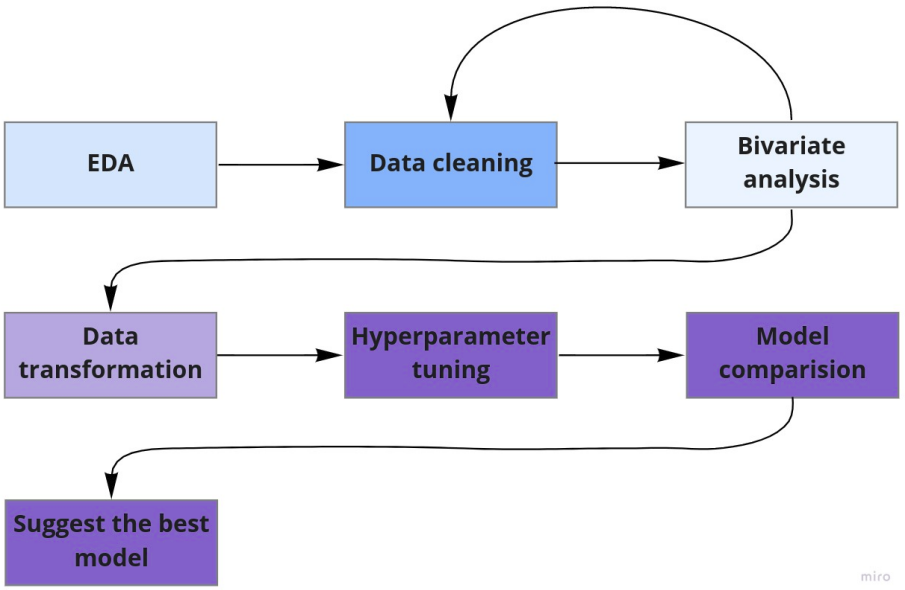


*Image 1. Mockup: workflow proposal*

## Evaluation Metrics

The idea is that the Data analyst or Data scientist can analyze several metrics around a problem that they are working on. Therefore, the workflow should return different metrics. But, in this particular case, the metric for Titanic dataset is the **Accuracy** and for Diabetes readmission dataset the **AUC ROC** will be used, with the goal to compare with the kaggle's competition.

## Benchmark Model

In this case the benchmark is going to be the kaggle's competition, therefore, the accuracy benchmark will be 86%, and the AUC ROC will be 69%.

## Project Design

The application proposal is something like *Image 1*.

First, I'm going to define the EDA structure for any dataset, this EDA will be univariate.
Second, data cleaning will be applied, I propose to handle the missing values from some rules, I mean for instance remove features with high missing values. For quantitative variables, I propose to use tools like, LOF (Breuning, et al. 2000), Adjust Boxplot (Hubert, et al. 2008) and Isolation Forest (Liu, et al. 2008) to detect outliers. For categorical variables, I will remove features with many categories.
Third, I'm going to do a module that allows bivariate analysis, comparing the two classes using statistical inference (Bootstrap and Chi-square test).
Then, feature scaling will be done, we could use Standard Scaler and Robust Scaler tools.
Model to use: Logistic regression with and without regularization, Random Forest Support Vector Machine and XGBoost.
Finally, the results will be shown, further the best model will be suggested.

## References

XIN, Doris, et al. Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. arXiv preprint arXiv:2101.04834, 2021.

BREUNIG, Markus M., et al. LOF: identifying density-based local outliers. En Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000. p. 93-104.

HUBERT, Mia; VANDERVIEREN, Ellen. An adjusted boxplot for skewed distributions. Computational statistics & data analysis, 2008, vol. 52, no 12, p. 5186-5201.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. Isolation forest. En 2008 eighth ieee international conference on data mining. IEEE, 2008. p. 413-422.