CCT College Dublin

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning |
| **Assessment Title:** | CA1 Project |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Izaias de Oliveira Gomes Junior  Salvador Navarro |
| **Student Number:** | 20 |
|  |  |
|  |  |
|  |  |
|  |  |
| **Assessment Due Date:** | 26th Nov 2023 |
| **Date of Submission:** | 26th Nov 2023 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

## Introduction Error! Bookmark not defined.

Business description 3

Hypothesis Error! Bookmark not defined.

[General goal 3](#bookmark)

[Success criteria/indicators 4](#bookmark1)

[Technologies used 4](#bookmark2)

[Models and machine learning algorithms 4](#bookmark3)

[Libraries 4](#bookmark4)

Accomplishment 4

Data 4

[source 4](#bookmark5)

[Attributes 4](#bookmark6)

[Dimensions 5](#bookmark7)

[Descriptive statistics and Data visualization 5](#bookmark8)

[Correlation: Pair plot 11](#bookmark9)

[Data preparation and pre-processing 12](#bookmark10)

[Flowchart of Data Preparation and Modelling 13](#bookmark11)

[Feature engineering 15](#bookmark12)

[Models 15](#bookmark13)

[Challenges encountered 15](#bookmark14)

[Inclusion of strategies to overcome them 15](#bookmark15)

[Results and analysis 16](#bookmark16)

[Step 1 - Defining the X and y 16](#bookmark17)

[Conclusion 177](#bookmark18)

[References 188](#bookmark19)

## Introduction

For this pair-based project from our Machine Learning discipline we had to choose between three study areas: Covid-19, Transport and Crime. Transport was chosen not only because it is of great interested of both of us but because it is a central subject

The changes in the planet’s climate has been an important issue in our daily lives for decades, reaching alarming proportions in the past few years (REFERENCE). Among all the variables that impact these changes, transportation is one of the most important due the high consumption of fossil fuels and the number of individual cars (REFERENCE). Governments and companies have tried to find smart solutions to minimise the amount of pollution originating from means of transportation, such as: renewable fuels, more access to public transportations, individual means of transportation with less impact and others. The usage of bicycles in the daily life is one

# Business Understanding

We are going to predict the number of bicycles rented per hour in the city of Seoul. As this variable is a continuous one, we will implement Regression Models in our Machine Learning and try to find the one with the best results. It is important for any business to know all the variables that impact their revenue and how they influence it. In our particular case it is also important for the company to supply the right amount of bicycles throughout the day, and understand the variation and the demand in different hours.

# Data Understanding of Raw Data

The dataset used in this project was extracted in the UC Irvine Machine Learning Repository with DOI [10.24432/C5F62R](https://doi.org/10.24432/C5F62R). “This dataset is licensed under a [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/legalcode) (CC BY 4.0) license.” It contains 14 features, with 8 of them being related to the weather conditions, and 3 categorical being the Seasons, Holiday (if the day is a holiday or not) and Functional Day (if the day is a week or a weekend one). The number of observations is 8760 which is divided by hour of the day with the first day being 01/12/2017 and the last one 30/11/2018.

We will transform the data grouping it by month, reducing the number of instances, for the data visualisation and use the original data set (by hour) for the Machine Learning models. It is easier to visualise the data when grouped by month, because it is reduced from 24 hours to 12 months. However for the prediction the company will be more interested in the prediction by hour once the objective is to improve the operational planning by maximising customer satisfaction (having the expected number of bicycles at the busiest time) and minimising costs (doing the necessary maintenance in low demand periods when a large number of bikes are not in use).

# Business Description

# Hypothesis and General Goal

We are going to predict the number of bicycles rented per hour in the city of Seoul. Our hypothesis is that the weather has a significant correlation with the target variable as well the hour of the day. The aim of the predictions is to support a better understanding of the dynamics of the demand so the company can optimise its service, marketing strategy and customer experience.

# Success criteria/indicators

To evaluate the success of our predictions we are going to use: the Mean Squared Error (MSE) and its root, the R-squared and the Adjusted one, the Cross-Validation Score, the comparison between the MSE on the test and on the train data and Huber Loss. The first four are from the Sklearn library and the last one from Tensorflow.

The MSE is a method to calculate the error with the squared difference between the actual and the predicted value and it is common to take its root to simplify the comprehension putting it in the same unit as the original data. (M, 2021). The R-squared is the common measure to indicate the quality of the model, it is a proportion of the variance in the target variable that is explained by the features of the model (Frost, 2018). In the meantime the Adjusted R-squared takes into consideration the numbers of features of your model, measuring if adding more would be better and only increases if the additional features help to improve the model and penalising if it does not (Bhandari, 2020).

The Cross-Validation Score is a technique used to avoid overfitting by estimating how well the model will perform and generalise in unseen data and helping to reduce bias (Brownlee, 2018). Another way to avoid overfitting is comparing the MSE on the test and training data, the ideal is to have those two as close as possible (Bishop, 2006). The Huber Loss has the same goal as the MSE but it handles the presence of outliers in a better way (M, 2021).

## Technologies used

## Machine Learning Algorithms

We worked with a significant range of ML Algorithms to test which one would get the best results. Here is the list:

* Linear Regression, we experimented some variations of it with Polynomial for 2 and 3 degrees, and the Ridge and Lasso regularisation techniques;
* Random Forest;
* Artificial Neural Network;
* Support Vector Machine, we also experimented with the SVR adjusting the kernel for polynomial;
* K-Nearest Neighbours;
* Decision Tree.

## Libraries

## PCA

PCA is a handful method that reduces the dimension of the dataset but conserves the most significant variations in the pattern within the data. It is most useful to simplify the analysis once it is easier to do the exploring and faster for the ML algorithms to process. We tried to apply PCA in our dataset but the results were significantly worse outcomes than without it. A link with a separate Jupyter Notebook of the test is provided in the markdown of the main project to prove the experiment and the results will be shown at the end for comparison.

## Correlation

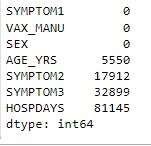
In regression problems it is important the right amount of variables so the R-squared is more trustworthy

In order to improve the model reducing the number of features by choosing the significant ones, a

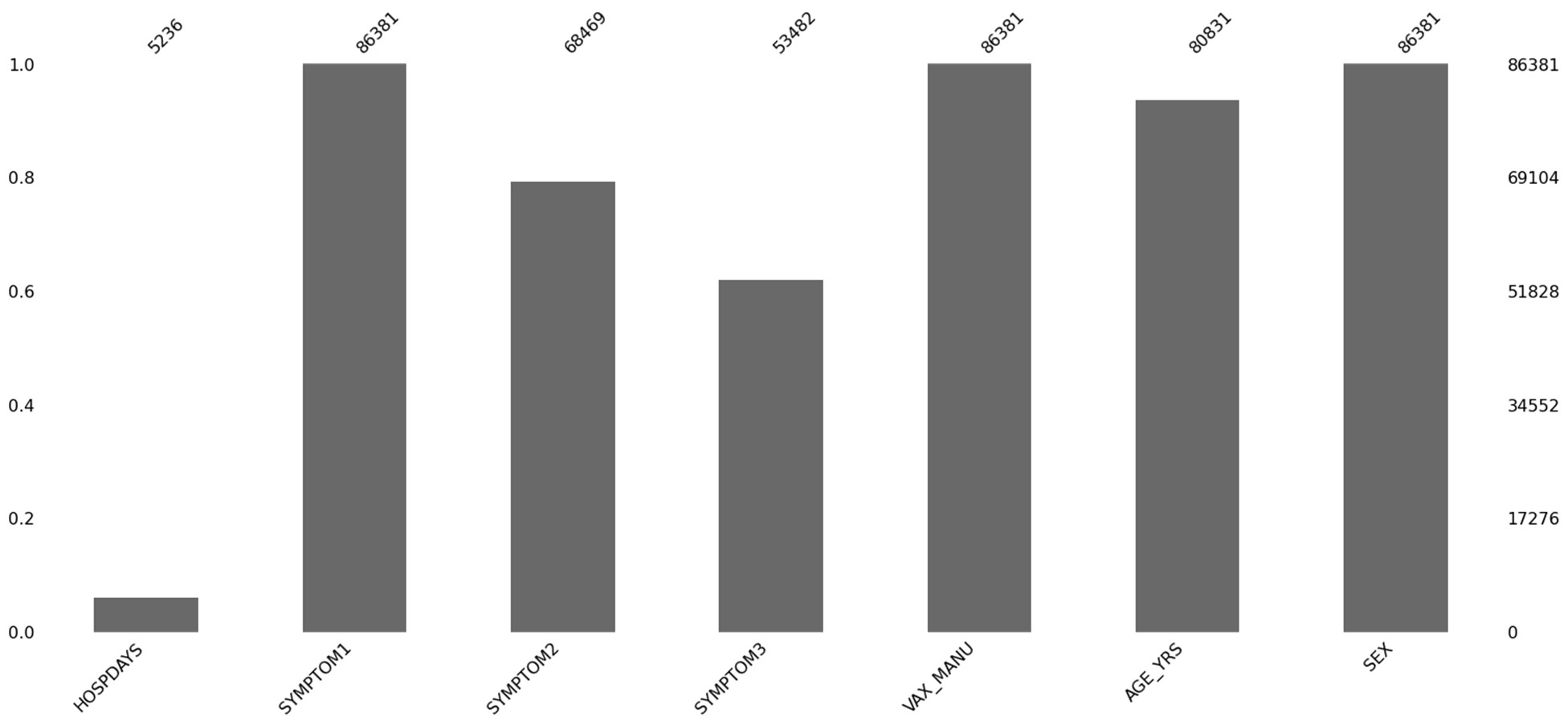
## Data preparation and pre processing

We have used different steps in data preparation such as cleaning the data, resizing of data, and feature engineering. We wanted to predict how many days a person will spend in hospital (if any days) after they advise they have had an adverse symptom to the Covid19 vaccine. From looking at the data dictionary, we dropped the features we believed were not necessary to complete our regression prediction problem i.e. we initially concentrated on the features we believed important for completing the regression problem. It is easier to choose the features we require than drop the features we don't require. This reduces the number of features from 52 columns to only 6 columns. Now, we have looked for missing values in our dataset and remove those values. There were tens of thousands of missing values in at least 4 of our selected features as shown in the below table and graph.

*Figure 9. Table of missing values*



*Graph 5: The data variables*



## Flowchart of Data Preparation and Modelling

Dataset check- understanding the raw data

Data Dictionary- check for quantitative and qualitative data

Analyze Features relevance/significance in regards to the project objectives

Decide the US state that we are analyzing

image15.png

Descriptive Statistics trough visualization Top 10 Symptoms

Count of how many died Top 10 allergies

Top covid vaccine manufacturers Distributions of Sex and age

Correlation analysis

Feature Selection- Initial removal/dropping of the unwanted columns to avoid noise and reduce the complexity of the EDA process

Text processing and using the top symptoms as recorded in the 3 symptoms features, reducing the dimension of the dataset

Feature Engineering:

* Change PFIZER\BIONTECH to PFIZER
* Remove any rows where the SEX is recorded as 'U' and the VAX\_MANU is is unknown
* Change the na values in HOSPDAYS to 0
* Change the values to the SEX feature to 1 and 0
* Change the AGE\_YRS and HOSPDAYS from floats to ints.
* Change all strings to lower caps.
* Drop any remaining null values.

Encoding the Symptoms and VAX\_Manu feature

Modelling using Random Forest Algorithm Defining the X and Y features

Splitting the dataset into the test and training set Creating and fitting the regressor to the training set

Calculating the accuracy of the training and the test set.

## Feature engineering

We have feature engineering analysis to improve the predictions of the results by extracting the information from features of raw text. We may do the following steps in order to extract the information. We have changed “PFIZER\BIONTECH” to “PFIZER”. Removed any rows where the SEX is recorded as 'U' and the VAX\_MANU is unknown. Changed the na values in HOSPDAYS to 0. Changed the values to the SEX feature to 1 and 0. Changed the AGE\_YRS and HOSPDAYS from floats to integers. Changed all strings to lower caps. Dropped any remaining null values.

## Models

We have used two different machine learning algorithms including linear regression and random forest algorithm. Both the models have focused on to resolve the regression and classification prediction of featured variable. We have done different steps to make the data useful. Such as removing unnecessary variables and making the dummy variables of symptoms using age of the patients. After encoding there were 4239 rows and 44 feature variables in this dataset.

## Challenges encountered

Different challenges have been encountered in working with this dataset and applying models on it. These challenges may include the big size of dataset, categorizing the descriptive text of symptoms. There were a lot of missing values, cleaning and pre-processing of the data, sparse data after encoding, finding the perfect model that may give greater accuracy.

## Inclusion of strategies to overcome them

All of the 52 features (variables) in the dataset were not be useful for analysis. So, in order to reduce the size of large dataset, we have removed unnecessary features by selecting the features of our choice and removing all-other features. After the removal process, we remained with only 6 features to work on. There was a lot of missing values in our selected features. With the library of missing numbers, we have removed the missing values.

## Results and analysis

We don't need to scale our data for the ensemble machine learning model Random Forest. We have set the number of estimators at 10 and used ' entropy ' as the criterion parameter. The overall accuracy score of the model is 99.8% when predicting the number of days trough classification (if any) a person would require if reporting an adverse effect to the vaccine if we are provided with the SYMPTOM1, SYMPTOM2, SYMPTOM3, VAX\_MANU, AGE\_YRS and SEX of the patient.

## Step 1 - Defining the X and y

In this step we may be dividing the testing and training dataset. Separating out the dependent or target variable

(y) and the independent variables (X) and convert them into arrays using values as ML algorithms require arrays (Menduni et al., 2022). I'm looking to predict the number of days a person will stay in hospital after reporting an adverse effect of the Covid19 vaccine so the 'y' variable will be HOSPDAYS. The 'X' features are the features that will feed into our target feature.

Firstly, we have imported Random Forest Classification from sklearn.ensemble. Then we have created the classifier variable and use n estimators to set the number of parameters i.e., it's the number of trees we want to add. It is recommended to work with ten decision trees and ‘entropy’ as the criterion, so we apply these as our parameter for our model. We then fit the classifier onto our X\_train and y\_train variables.

The results in this model have showed a very high accuracy of 99.8 percent. In the evaluation of the model, we have used Crisp DM model also as shown in figure. So, we must re-evaluate to see what else can be done to improve and make our prediction accurate classifying/ predicting the model.

*Figure 10: Crisp DM model*



## Conclusion

In this paper we have focused to analyse the Vaers dataset which explains different adverse symptoms of patients after their vaccination been done. We have tried to see and predict the results of how many days people may have to be hospitalized with adverse symptoms after the vaccination process. So the number of days were analysed using repressors that almost 10 to 20 days patients may have to hospitalize in this situation. Using the Random Forrest classifier we have been able to predict with 99% accuracy the amount of days someone with certain symptoms may have to stay in hospital. However our results may not be as accurate as they first appear. The data is severely imbalanced which may be causing the classifier to get such good results. We are now entering the evaluation stage of the CRISP DM model which means we may have to re-visit some of the earlier steps to obtain confidence in our predictions. As a team we feel we are on the right track and we have made great progress up to this point with a little more effort we are well on our way to achieving our goals.

## References

Classen, J.B., 2021. Review of COVID-19 vaccines and the risk of chronic adverse events including neurological degeneration. *Journal of Medical-Clinical Research and Reviews*, *5*(4), pp.1-7.

Gee, J., Marquez, P., Su, J., Calvert, G., Liu, R., Myers, T., Nair, N., Martin, S., Clark, T., Markowitz, L., Lindsey, N., Zhang, B., Licata, C., Jazwa, A., Sotir, M. and Shimabukuro, T., 2021. First Month of COVID-19 Vaccine Safety Monitoring — United States, December 14, 2020–January 13,

2021. *MMWR. Morbidity and Mortality Weekly Report*, 70(8), pp.283-288.

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N. and Shearer, S., 2018. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and electronics in agriculture*, *153*, pp.213-225.

Menduni, G., Zifarelli, A., Sampaolo, A., Patimisco, P., Giglio, M., Amoroso, N., Wu, H., Dong, L., Bellotti, R. and Spagnolo, V., 2022. High-concentration methane and ethane QEPAS detection employing partial least squares regression to filter out energy relaxation dependence on gas matrix composition. *Photoacoustics*, *26*, p.100349.

Moro, P.L., Haber, P. and McNeil, M.M., 2019. Challenges in evaluating post-licensure vaccine safety: observations from the Centers for Disease Control and Prevention. *Expert Review of*

*Vaccines*, *18*(10), pp.1091-1101.