A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Lorrany Inacio Cardoso dos Santos |
| *Student Number* | 2024062 |
| *Module Title* | Data Preparation and Machine Learning |
| *Assessment Title* | Integrated CA2 |
| *Assessment Due Date* | 24th May 2024 |
| *Date of Submission* | 29th May 2024 |

**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.

SUMMARY

[Introduction 1](#_Toc167883266)

[Objective 1](#_Toc167883267)

[Development 2](#_Toc167883268)

[Data Preparation 2](#_Toc167883269)

[Implementing PCA 5](#_Toc167883270)

[Machine Learning 6](#_Toc167883271)

[References 10](#_Toc167883272)

# Introduction

Nick states that in the field of data science, a project manager is responsible for ensuring the successful completion of advanced analytics and AI/Machine Learning projects. While their role shares similarities with traditional IT project management, it is specifically focused on the complexities of data science applications. The main duties include: outlining detailed project roadmaps, coordinating and managing the daily activities and workflows of project teams, interacting with stakeholders, defining project tasks to align with the overall project vision, overseeing and documenting the project scope, and identifying and gathering the necessary datasets for project execution. Additionally, some positions may require technical skills, where the data science project manager is expected to be proficient in Python and visualization tools (HOTZ,2024).

According to Ndung data preparation requires understanding the desired output from the machine learning model and identifying the necessary data attributes to achieve that output. Knowing the output helps in collecting relevant data and defining its quality and value (NJERI,2022). Ndung also says that data preparation is the process of converting raw data through pre-processing before being used in fitting and evaluating machine learning predictive systems. Machine learning models are particular to their data source, and hence the credibility of the data source and utility of the data collected is essential. It is plausible for a machine learning model to be high end model but training it with the wrong data yields the wrong information. Standardizing your data entry point ensures the right information is attained at the end result. For these reasons, data collection remains an imperative part of data preparation. Data preparation ascertains minimal errors in your data, and allows for data monitoring of any future errors. This will eventual ensure the machine learning is trained with the correct data and hence the output will be accurate. Data exploration analysis will provide a summary of your data set, and allow for necessary changes or formatting to be done (NJERI,2022).

# Objective

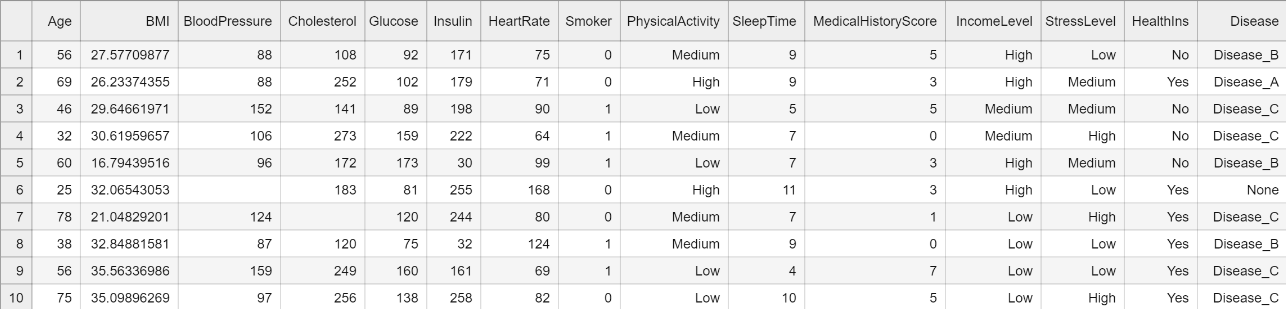
The main objective is to use patient data to improve treatment results and operational efficiency. With this in mind, an analysis objective will be formulated to guide the study and subsequent analyses. "Identify patterns and factors that positively influence treatment outcomes and operational efficiency within the healthcare organization, utilizing patient demographics, health metrics, treatment details, and follow-up outcomes, to provide recommendations that drive continuous improvements in patient care and internal operations."

# Development

## Data Preparation

As maintained by IBM website data understanding phase focuses on closely examining the available data for mining. This step is crucial for preventing unforeseen issues during the subsequent data preparation phase, which is often the most time-consuming part of a project. During data understanding, you access and explore the data using tables and graphics. These can be organized in with the CRISP-DM (Cross-Industry Standard Process for Data Mining) project tool. This process helps you assess the data's quality and document the findings in the project records (UNKNOWN, 2021)

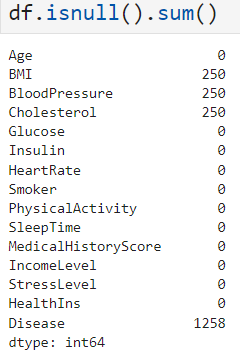
The common code to understand the data is “.head()”, which prints all the columns and rows as shown in Table 1 below.



**Table 1 (Jupyter Notebook, 2024)**

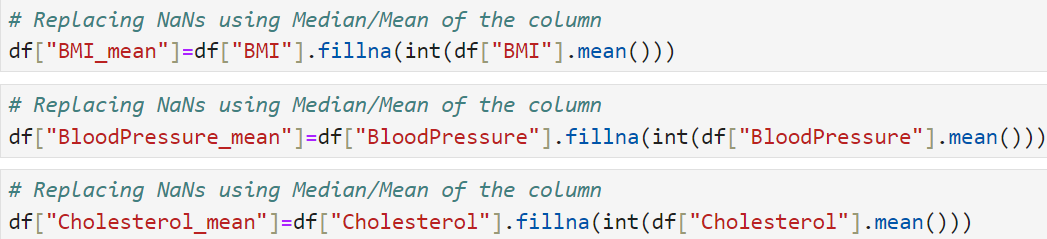
Data pre-processing involves several steps:

* Data Cleaning: First you identify the noise, missing, or inconsistent data and treat it later.

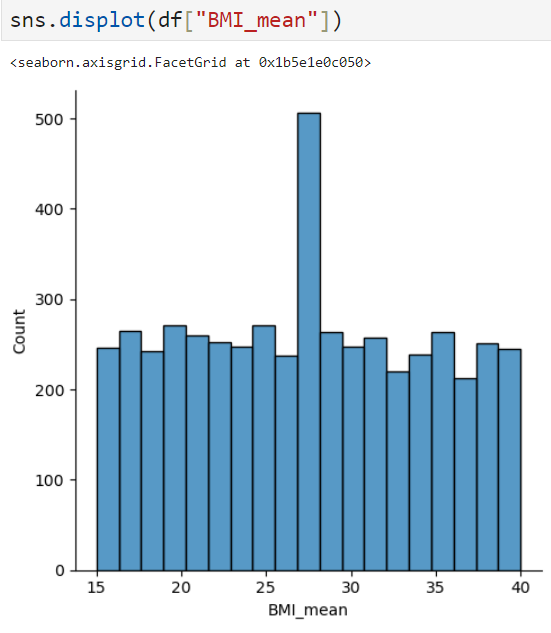


**Figure 1 (Jupyter Notebook, 2024)**

Generally, when you have missing values, you need to understand their functionality in the dataset. Because just removing them sometimes harms the analysis. We have a dataset with 5000 rows, 2008 is considered missing values ​​if you do not analyze it carefully because when you compare it with the dictionary, you understand that 1258 are patients without disease, which does not define that is missing data. In this case, we will not delete any missing values ​ but rather treat them. The following codes were used.



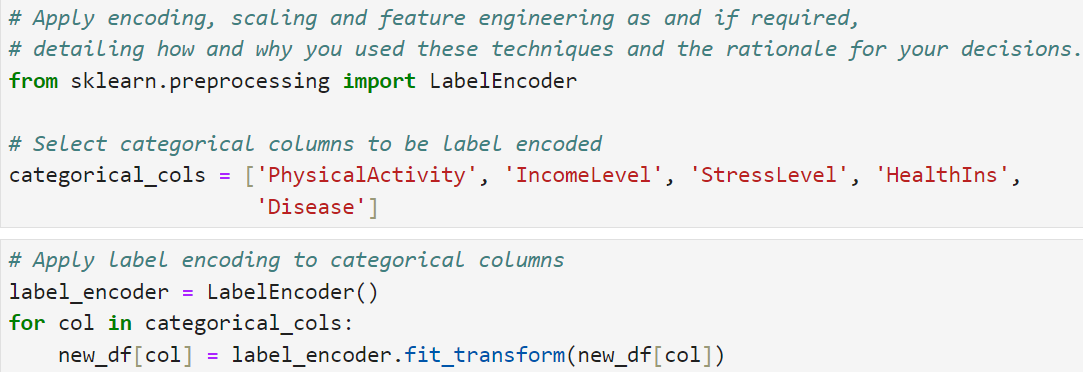
**Figure 2 (Jupyter Notebook, 2024)**



**Graphic 1 (Jupyter Notebook, 2024)**

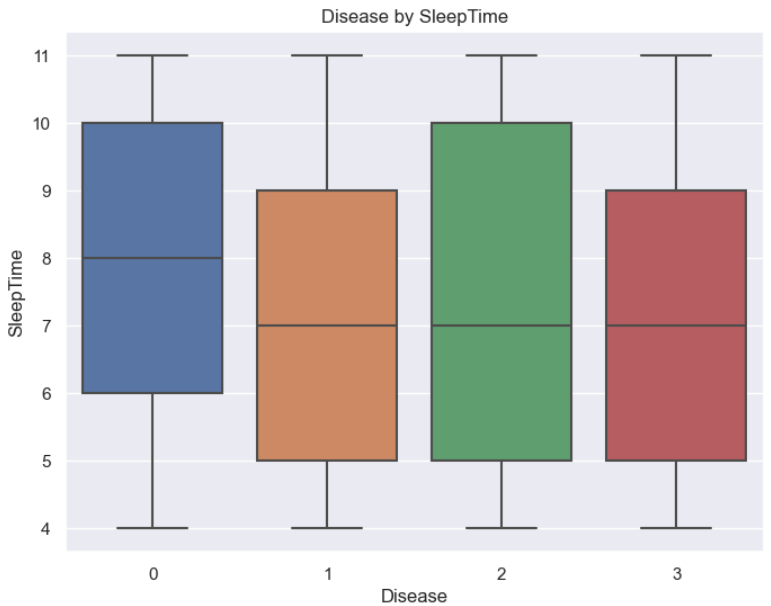
The histogram shows that the distribution of BMI\_mean is somewhat uniform, except for the significant spike at the mean value used to fill NaNs. This visualization highlights the impact of filling missing values with the mean, demonstrating that it can introduce a prominent peak in the data distribution.

Another code used that is also very important in the preparation stage is called Label Encoder, which transforms all categorical data into numerical data at once.



**Figure 3 (Jupyter Notebook, 2024)**

Ndung says that exploratory data analysis is mainly done on statistical manipulation software. The graphical techniques allow for understanding the distribution of the data set, and the statistical summary of all attributes. The provided Figure is a box plot that visualizes the distribution of sleep time (SleepTime) for different disease categories (Disease) (NJERI,2022).

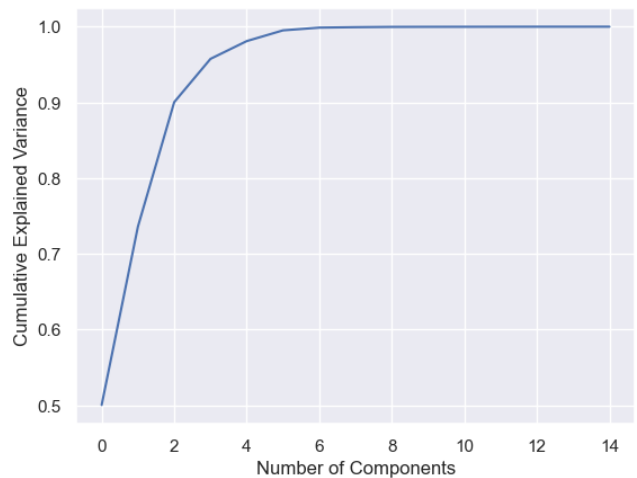


**Graphic 2 (Jupyter Notebook, 2024)**

The graphic previously suggests differences in sleep time among patients with different diseases. Understanding this information can be crucial for healthcare organizations aiming to improve treatment outcomes by considering patients' sleep patterns as part of their overall health assessment. For example, patients with Disease 1 typically sleep less compared to other categories, with a concentration of around 6 to 8 hours.

### Implementing PCA

Conforming to IBM website, principal component analysis (PCA) simplifies large datasets by reducing the number of dimensions to a few key components that keep most of the original information. It does this by converting potentially related variables into a smaller set of uncorrelated variables, called principal components. This helps reduce model complexity because adding more features can worsen model performance, a problem known as the “curse of dimensionality.” By mapping a high-dimensional dataset into a smaller space, PCA also helps avoid issues like multicollinearity and overfitting (UNKNOWN, 2022).



**Graphic 3 (Jupyter Notebook, 2024)**

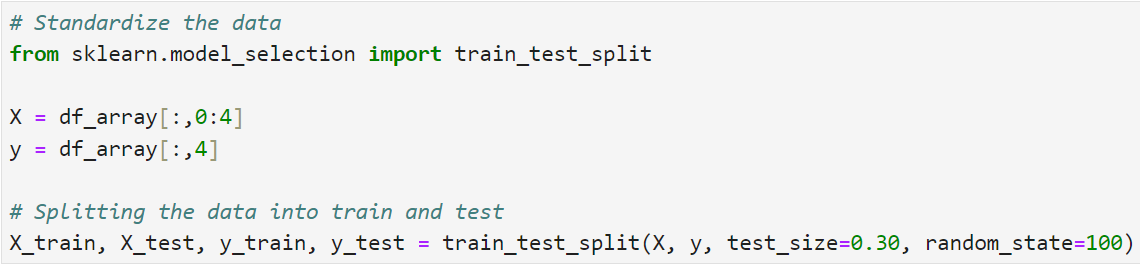
This elbow plot before suggests that using around three to four principal components would be sufficient to capture nearly all the variance in the data, providing an effective dimensionality reduction while preserving the dataset's essential information.



**Table 2 (Jupyter Notebook, 2024)**

The previous table shows what the dataset looks like after implementing PCA, reducing 14 features into just 4.

Following Ndung in his studies, any data source in machine learning is divided into both the training and the test data, and the technique of this division is achieved during data preparation.



**Figure 4 (Jupyter Notebook, 2024)**

# Machine Learning

Ravi says that choosing the appropriate model for a particular problem can be difficult in the rapidly developing field of machine learning. Optimal model performance can be achieved by following a methodical strategy, especially with the abundance of available algorithms and methodologies. We will go over important factors and methods in this post to assist you in selecting the ideal machine learning model for your project. It is important to have a clear idea of the problem you are trying to solve before you start choosing a model. Clearly define the work at hand, the expected results, and the issue statement (e.g., classification, regression, clustering). This will make things clearer and help you choose the best machine learning models (SINGH, 2023).

The basis of our decision is that my healthcare organization aims to use patients’ data to improve treatment outcomes and operational efficiency. Then thinking this was decided modelling Random Forest.

In accordance with the article on the IBM website Random Forest is a popular type of machine learning method that combines many decision trees to make predictions. Decision trees work by asking a series of questions to classify or predict things. For example, a decision tree might decide whether to go surfing based on factors like wave size and wind direction. Each question splits the data into smaller groups until a final decision is made. However, a single decision tree can sometimes make mistakes because it focuses too much on the training data. Random forests solve this problem by using many different trees and averaging their predictions, which helps to reduce errors (UNKNOWN, 2024).

Benefits of using random forests include:

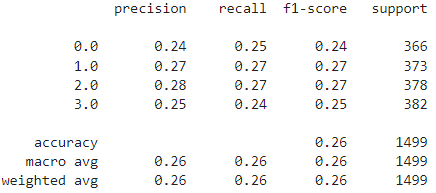
* Lower risk of making mistakes: Since random forests use many trees, they're less likely to overfit the data and make overly specific predictions.
* Versatility: They can handle different types of tasks, like predicting categories or numbers, with high accuracy.
* Easy to understand feature importance: Random forests can show which factors are most important for making predictions.

Challenges of using random forests include:

* Time-consuming: Building and using random forests can take a long time because they involve processing a lot of data.
* Resource-intensive: They require more computer memory and processing power because they work with large datasets.
* Complexity: Interpreting the predictions of a random forest can be more difficult than understanding a single decision tree's prediction.

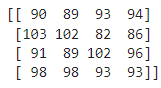
**Understanding the results**

* Classification Report: The model has low precision, recall, and F1-scores across all classes, indicating it struggles to make accurate predictions. The overall accuracy of 26% suggests that the model's predictions are only better than random guessing for this dataset. The similar values for macro and weighted averages indicate that the class distribution is relatively balanced.



**Figure 5 (Jupyter Notebook, 2024)**

* Learning on Science Direct website we understand that prediction summary is represented in matrix form by the confusion matrix. It displays the proportion of accurate and inaccurate forecasts for each class. It aids in comprehending the classes that the model is confusing with respect to other classes.

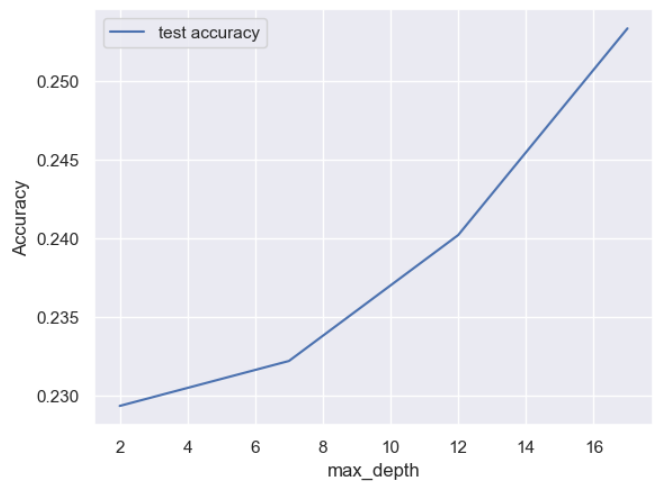


**Figure 6 (Jupyter Notebook, 2024)**

A sizable percentage of all classes have been misclassified, suggesting that the model has trouble telling one class from another. There are a significant number of cases in each class that are incorrectly assigned to other classes, indicating that the attributes the model uses to distinguish between the classes may not be unique enough.

**Regularization and Hyperparameter Tuning:**

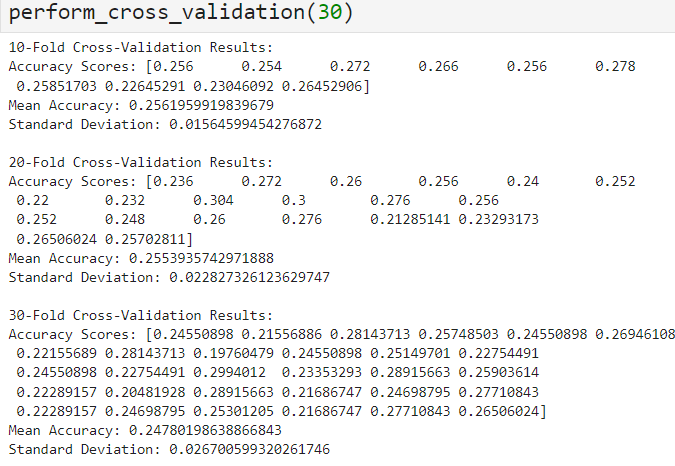
According to Alice regularization hyperparameter regulates the model's capacity, or how adaptable and how many degrees of freedom it has while fitting the data. Knowing the total number of trees is necessary for both boosted decision trees and random forests (although this might also be categorized as a regularization hyperparameter). It is also necessary to adjust them to sensible values so that the training process can identify a suitable model. Hyperparameter configurations may significantly affect the trained model's prediction accuracy. various datasets frequently have various optimal hyperparameter values. As a result, each dataset should be adjusted. There must be a meta-process that fine-tunes the hyperparameters because the training process does not set them (ZHENG,2024).



**Graphic 4 (Jupyter Notebook, 2024)**

When trying different settings ("hyperparameters") for models, like the C setting for an SVM, there is a risk of overfitting the test set. This means that you might adjust the settings to make the model perform well only on the test set, which can cause the model to learn details specific to that set instead of generalizing well to new data. To avoid this, a separate "validation set" can be used. The model is trained on the training set, evaluated on the validation set, and then, if successful, finally tested on the test set. However, splitting the data into three parts (training, validation, and test sets) reduces the amount of data available for training the model. This can also lead to results that depend heavily on how the data is split.

A solution is a method called cross-validation (CV). With CV, you still keep a separate test set for the final evaluation, but you don't need a validation set. In basic k-fold CV, the training set is divided into k smaller sets. This allows the model to be trained and validated multiple times on different data splits, ensuring a more reliable evaluation.

****

**Figure 6 (Jupyter Notebook, 2024)**

# Conclusion

The cross-validation results suggest that your random forest model is not performing well, with a mean accuracy of around 25% regardless of the number of folds. To improve the model's performance, some ideas were considered like enhancing the features used for training the model, optimizing the hyperparameters of the random forest, ensuring that the data is clean and well-preprocessed, trying different machine learning algorithms or ensembles to see if they perform better and if there is a class imbalance, use techniques like oversampling, undersampling, or class weighting.

# References

Nick Hotz. “The Project Manager in Data Science”. Data Science Process Alliance. <https://www.datascience-pm.com/data-science-project-manager/>

Njeri, N. (2022), ‘Data Preparation for Machine Learning Modelling', International Journal of Computer Applications Technology and Research, Volume 11–Issue 06, 231-235, 2022.

Unknown. “Understanding Data Overview”. IBM Website. <https://www.ibm.com/docs/en/spss-modeler/saas?topic=understanding-data-overview>

Unknown. “DM CRISP”. IBM Website. <https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview#crisp_overview>

Unknown. “What is random forest?”. IBM Website. <https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,Decision%20trees>

Singh, R. (2023). “Choosing the Best ML Model: A Guide to Model Selection”. Linkedin. <https://www.linkedin.com/pulse/choosing-best-ml-model-guide-selection-ravi-singh/>

Kulkarni, A. (2020). “Confusion Matrix”. Science Direct Website. <https://www.sciencedirect.com/topics/engineering/confusionmatrix#:~:text=A%20confusion%20matrix%20represents%20the,by%20model%20as%20other%20class>.

Zheng, A. “Evaluating Machine Learning Models”. <https://www.oreilly.com/library/view/evaluating-machinelearning/9781492048756/ch04.html#:~:text=A%20regularization%20hyperparameter%20controls%20the,has%20in%20fitting%20the%20data>.