**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA 2 – Capstone Project Proposal |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Arthur Speggiorin Verza |
| **Student Number:** | SBA23382 |
| **Assessment Due Date:** | 29/10/2023 |
| **Date of Submission:** | 28/10/2023 |

**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. |

**Introduction**

I was asked to find for my assignment selecting a dataset that I was interested. The task consists in creating a business plan to formulate a hypothesis and questions, to find the answers for it. I did a search using Kaggle on a few topics that I found interesting and I found this dataset about sleep health and lifestyle. I want to predict the correlation of quality of sleep with physical activity, weight, sleep duration, stress level and any sleeping disorder.

According to (Jansen, 2020), sleep is interdisciplinary because it touches every aspect of health and is essential for every person. Also in accordance with (Ramar et al., 2021), sleep is a biological necessity and in lower levels can cause disorders in health and well-being.

The Sleep health and Lifestyle Dataset contains around 400 rows and 13 columns, where I can find a wide range of information about sleep and people habits, including gender, age, occupation, sleep duration, stress level, quality of sleep, physical activity level, blood pressure, heart rate, daily steps and sleep disorders.

**Objectives**

The objective in this project is to predict sleep quality and stress for people and try to understand which are the biggest effects of a bad night of sleep on people's daily tasks. I am analysing and relating the occupation with stress level, and see the impact that it has in daily physical activities and quality and duration of a night of sleep. Also I am testing about what is the influence of any sleep disorder in the quality of sleep. For this analysis I will be using some Machine Learning models that I will be learning throughout course length and will be looking at the ones presented on the classes and will be pondering which would fit better on our project and dataset chosen.

For now I am looking to use a supervised learning because as stated in (Müller and Guido, 2016, p.25) Supervised learning is used when we want to predict a certain outcome from a given input and we have examples of input/output pairs, so as in our data set I have, I have decided to use the supervised learning.

Upon checking the dataset chosen I decided that It is important to consider all the information presented in order to understand the real impact of sleep quality in people's lives using as reference some factors and behaviours present on the daily life of an individual.

**Problem definition**

When looking into the chosen dataset I realised that some occupations interestingly have a relation to the level of stress and also some other factors such as quality of sleep, time spent in physical activity during the day and the sleep duration and if the person have or not any sleep disorder, all this together would affect the stress level of a person. So with my project I want to predict the correlation of quality of sleep with physical activity, occupation, sleep disorder and sleep duration with the stress level and in doing so, maybe propose some useful information to help people reduce their stress level.

**Scope**

In this project I am going to analyse one dataset about stress level and the correlation with sleep duration, sleep quality and physical activity. My dataset has 13 variables, but after a review of the dataset we decided to exclude some columns from the analysis.

I decided to not exclude any column from my dataset, because it may affect and have a positive or negative relation with the sleep quality. I will analyse some of the points, won’t analyse all because it will be so many variables and can be confused.

The stress level and the sleep quality are inversely related  and also looking into the other variables, like job occupation, sleep disorder. I will be looking at I can see that the level of stress changes accordingly depending on the job occupation and quality of sleep and sleep duration of a person on a night. According to my initial analysis of the dataset, people who have jobs where requires longer shifts and people with higher level of stress have worst sleep quality. Also, people with some sleep disorder tend to have worst sleep quality.

With this capstone project I will try to deliver a model where it will be possible to predict based on some variables the level of stress someone will have and then will be able to avoid high levels of stress providing support to improve the other variables that would affect it.

As the capstone project is a 2 semester project I need to have a proposed timeline so I can have a good project management plan.  I am thinking about the platform Monday.com to use as a good way of keeping track of the following phases. I have below some idea of how it will be divided by week duration, remembering that this is just an idea as we are still on the way of learning the topics to be used on the project.

* Dataset Analysis and preparation ( Approx. 3 weeks)
* Select Machine Learning Model to be used and start working on applying it  (Approx. 2 weeks)
* Apply Machine learning Model and do tests to see if will work as expected (Approx. 2 Weeks)
* Reports building will be done during the tests and selection of the model

**Data Sources**

The data selected was taken from Kaggle linked [here](https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data), I would like to make an observation about the dataset as this data set is a synthetic one that was created only for illustrative purposes and no real people information was used.

**Ethic**

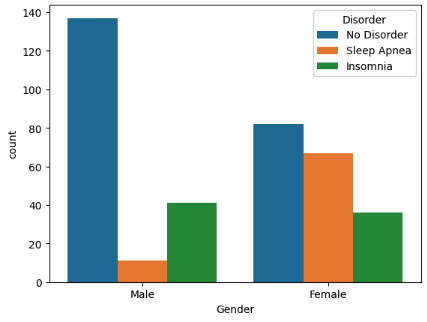
In my view as mentioned previously I did not have any issues with requesting permission to use the data set as it was publicly available and also no ethical issues due the dataset being a synthetic one and not personal information that could label a person was presented.

**Data processing and results**

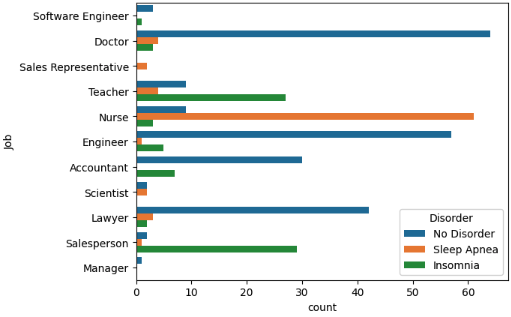
Initially, I started replacing the columns name for smaller names, I did this to be better for visualisation when I call ‘head’ or other commands. Also, checked the missing values and duplicated. When I checked missing values, I could see in sleep disorder I had 219 missing values, but in this case it means 219 people don’t have any sleep disorder, so I replace this missing value per ‘No Disorder’.

Checking the sleep disorder according gender it is possible to see male people have less disorder than female people(as per figure 1 below). Once I checked the sleep disorder based on occupation, it’s possible to see people where the working shift are longer or in general have payments lower, tend to have high number of sleep disorder, as per figure 2 below.

*Figure 1 – Gender x Sleep disorder*

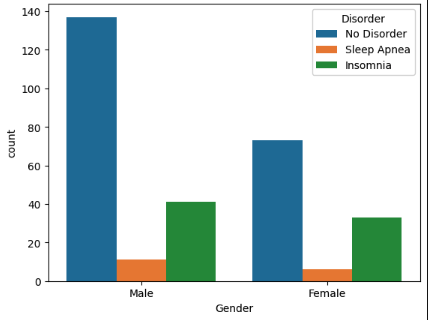


*Figure 2 – occupation x sleep disorder*



Also, it’s possible to see nurse is the group that have more sleep disorder for female and salesperson is the one for male. And after, I tested the sleep disorder excluding nurses to see and the result is totally different, showing a smaller number of women with any problem.

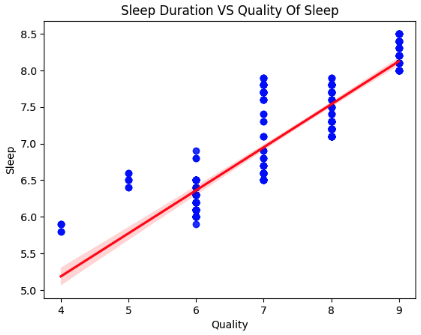
*Figure 3 – gender x sleep disorder excluding nurses*



Using heatmaps to test the corelation between my features, I could see the age and apnea are correlated, but insomnia no and could be a good predictor for the absence of a disorder. Also, the sleep duration and quality don’t have any relation with disorders, but the disorders affects the sleep quality and duration negatively, without a disorder, these values are higher.

Another test I did is the sleep duration versus quality, and this positive correlation shows people with more hours of sleep tend to have better quality, as show in the graph below.

*Figure 4 – sleep duration vs quality of sleep*



In opposite side, the relation between stress level and the sleep quality is negative, as per graph below. This means that the greater the stress, the lower the quality of sleep.

*Figure 5 – stress level vs quality of sleep*



Now testing the module for some of Machine Learning tests, I initially used the OLS Regression and got 0.791 accuracy. When I tested linear regression, I got the same value of 0.791. Those values are not a high accuracy, because it is under 80%.

The other test I did was kNN model, and according Prakash Shyam(2023), kNN model is a simple and effective machine learning algorithm for both classification and regression tasks. In KNN, the idea is to classify an unknown sample based on its distance to the K nearest samples in the training set.

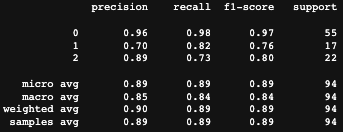
Using this test model, I got a table below and the results that I could extract from there were: in class 0 the precision was 96%, which is a good value. The recall was 98% and the F1-Score was 97%, which means the balance between Precision and Recall. The support for class 0 was 55, meaning the true instances for this class.

Classes with high precision ratio means there are many actually positive values and low false positive. In class 1 it is possible to see the precision was 70%, which is a lower value, not very accurate and the recall was 82% which is a bit better, but still not a satisfy result. The F1-Score for this class was 76% with 17 actual occurrences.

For class 2, the precision was a good result of 89%, but the recall was not accrue like the precision, getting a value of only 73%. F1-Score returns 80% and 22 actual occurrences, not a very satisfy value too.

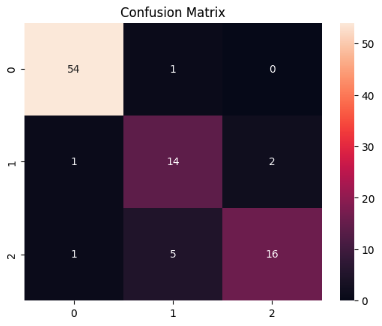
The other values, micro avg, macro avg, weighted avg and samples avg got values that can be considered ok, because it considered the three classes. Those values are around 89%, where is a good value, but if you look the class 0 are a bit lower, but if you compare with class 1 it is way higher.

*Figure 6 – table for kNN*



The confusion matrix have results similar with the kNN table, but with values bit lower than the previous test. The diagonal showing good values and the misclassified values are low, so it is satisfy result.

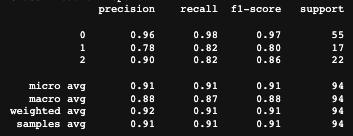
*Figure 7 – confusion matrix*



Testing the decision tree, which according SONG(2015), Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes.

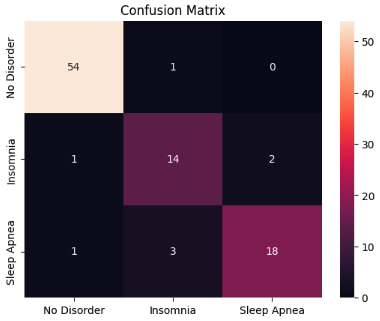
The classes 0 have the exactly same results of the kNN, but for class 1 and 2 they have increased a bit, which turns this more accrue than the previous test, as per table below.

*Figure 8 – decision tree table*



The other confusion matrix shows more accrue as well, showing a slide improvement if compared with the previous one. It seems a satisfactory result, with a small increase in accuracy compared to the previous test.

*Figure 9 – confusion matrix*



The final conclusion that I could get from this project was that some variables affect a lot in the quality of sleep, causing some good and clear results. The women who have nurse position or people in general where the occupation demands more hours of working, allowing less resting time seems to have higher level of stress, with worst sleep quality. In the opposite side, people with better positions, where the pressure is smaller if compared with the other group shows higher level of sleep quality and lower level of stress in general.

People overweight and with age higher seems to have more sleep disorders, which causes a bad night of sleep, increasing the stress level too. In general, healthier people seems to have a better live quality. This is something that serves as a lesson even for my personal life, to try to maintain healthier habits and have a better quality of life.

The link for my GitHub repository is [here](https://github.com/ArthurVerza/StrategicThinking.CA2)

**Reference list**

**Jansen, E. (2020). *Sleep 101: Why Sleep Is So Important to Your Health | The Pursuit | University of Michigan School of Public Health | Adolescent Health | Child Health | Chronic Disease | Epidemic | Mental Health | Obesity*. [online] sph.umich.edu. Available at: https://sph.umich.edu/pursuit/2020posts/why-sleep-is-so-important-to-your-health.html.**

**Müller, A.C. and Guido, S. (2016). *Introduction to machine learning with Python : a guide for data scientists*. First Edition ed. Beijing: O’reilly, p.25.**

Ramar, K., Malhotra, R.K., Carden, K.A., Martin, J.L., Abbasi-Feinberg, F., Aurora, R.N., Kapur, V.K., Olson, E.J., Rosen, C.L., Rowley, J.A., Shelgikar, A.V. and Trotti, L.M. (2021). Sleep is essential to health: an American Academy of Sleep Medicine position statement. Journal of Clinical Sleep Medicine, [online] 17(10). doi:https://doi.org/10.5664/jcsm.9476.

Prakash Shyam, Karuppiah . “K Nearest Neighbor - an Overview | ScienceDirect Topics.” www.sciencedirect.com, 2023, [www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/k-nearest-neighbor](http://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/k-nearest-neighbor).

Song, Yan-Yan, and Ying Lu. “Decision Tree Methods: Applications for Classification and Prediction.” Shanghai Archives of Psychiatry, vol. 27, no. 2, 2015, pp. 130–5, https://doi.org/10.11919/j.issn.1002-0829.215044.