Sleep\_health\_and\_lifestyle\_dataset

#### Notebook link: [ML\_Sleep\_health\_lifestyle.ipynb](https://colab.research.google.com/drive/1zehfC6KZkAQ0TjSphM7aMWOyMj8oijRP?usp=sharing)

### **Brief description of each dataset and tasks**

* ***Description***: This dataset is about how lifestyle and health affect sleep. The data includes information such as: Gender, Age, Occupation, Sleep, Quality of Sleep, Physical Activity Level, Stress Level, BMI Category, Blood Pressure, Heart Rate, and Daily Steps.
* ***Tasks***: Our task is to build a model to predict Sleep disorders based on the provided features. There are 3 types of disorders in these datasets: None, Sleep Apnea, and Insomnia.

### **Summary of model architectures and training strategies**

* 1. ***Model architecture:***
* The model architectures I used were 2 **Relu** layers, 1 **Dropout** layer, and 1 **softmax** layer.
* The reason why I used this model architecture is that:
  + RELU: Because it is fast and safe
  + Dropout: As mentioned in class, Dropout might make the learning process more efficient by creating more difficulties for the model
  + Softmax: Because our label is categorical, softmax is the best choice
  1. ***Training Strategies:***
* My approach was to clean all the data, followed by splitting the train and the test set. Then I did the preprocessing process before actually training the model, and finally ended with validating and testing the model. Along the way, I did add EarlyStopping and ReduceLearningRateOnPlateau to make sure the learning process was ‘safe’.

### **Comparative analysis of performance and feature importance**

1. ***Analysis of performance:***

* The model stopped at epoch 31, with accuracy: **0.9109** - loss: 0.3050 - val\_accuracy: **0.9600** - val\_loss: *0.1407* - learning\_rate: 0.0010.
* The test accuracy was 96%

**Confusion Matrix:**

[[13 0 2]

[ 0 44 0]

[ 1 0 15]]

**Classification Report:**

precision recall f1-score support

Insomnia 0.93 0.87 0.90 15

None 1.00 1.00 1.00 44

Sleep Apnea 0.88 0.94 0.91 16

accuracy 0.96 75

macro avg 0.94 0.93 0.94 75

weighted avg 0.96 0.96 0.96 75

* Not a good accuracy for such a small dataset like this, so there might be some problems. Moreover, it seems to be sensitive to outliers (not a non-sleep disorder).

1. ***Feature Importance:***

|  | **Percentage** |
| --- | --- |
| **Occupation\_Manager** | 14.408125 |
| **Occupation\_Teacher** | 13.747177 |
| **Department\_Law** | 13.576189 |
| **Quality of Sleep** | 13.539274 |
| **Occupation\_Software Engineer** | 13.468002 |
| **Department\_Health** | 13.415280 |
| **Diastolic** | 13.387852 |
| **Department\_STEM** | 13.370339 |
| **Age** | 13.294774 |
| **BMI Category** | 13.085119 |
| **Occupation\_Sales Representative** | 13.081861 |
| **Occupation\_Scientist** | 13.077158 |
| **Occupation\_Doctor** | 13.063266 |
| **Heart Rate** | 12.960947 |
| **Gender** | 12.930976 |
| **Department\_Business** | 12.909225 |
| **Rate Pressure Product** | 12.880842 |
| **Occupation\_Lawyer** | 12.822402 |
| **Physical Activity Level** | 12.759039 |
| **Sleep Duration** | 12.687552 |
| **Systolic** | 12.656175 |
| **Stress Level** | 12.610388 |
| **Occupation\_Nurse** | 12.335144 |
| **Occupation\_Accountant** | 12.307756 |
| **Department\_Education** | 12.255411 |
| **Daily Steps** | 12.163538 |
| **Pulse Pressure** | 11.812254 |
| **Occupation\_Engineer** | 10.797251 |

* The features are somewhat similar in terms of importance in this model.

### **Insights into what you discovered in your experiments**

* The correlation in the EDA process does not always tell you everything about the importance of the features. For example, in my case, I did some feature engineering (created features such as Pulse Pressure and Rate Pressure Product from Diastolic and Systolic ~ taken from Blood Pressure). Those variables have high correlation with the labels, but turn out not to be the most important features. And on the opposite side, features that are not too correlated, quality of sleep, do play a vital role.
* The width and depth are dependent on the data. I mistakenly used too many nodes for my RELU layers, which makes my model too complicated. But somehow, it really saves me from being underfitted in this situation since I noticed I made some mistakes in the preprocessing process.
* Feature Engineering or Extraction is not a must in deep learning. Deep Learning can learn interactions between features, which helps a lot in learning the pattern. But it does not mean we will never do it in Deep Learning. As for some information, for example, the Department, the model can not learn it without declaring it, so we should strike a balance in creating or not creating new features for the model to learn.
* Preprocessing incorrectly can affect your model. In my case, I should handle the Gender and BMI Category differently. Since I handled it similarly with numerical features, those features got Standard Scaler.