Matteo Shafer, Ashton Alonge, Oli Lane
Professor Stanchev
CSC 369
3 June 2024

Sleeping Disorder Prediction Report

Introduction

In this project we used Scala/Spark to create a Logistic Regression model to predict sleep disorders. The dataset we used contains 12 different health and lifestyle measurements for 374 different individuals. The columns we focused on for developing our model were sleep duration, sleep quality, physical activity, resting heart rate, daily steps, stress level, and occupation.

Algorithm

In our dataset, a sleep disorder is recorded as either None, Sleep Apnea, or Insomnia. Since sleep disorder is a categorical variable, a classification algorithm is necessary. We chose Logistic Regression because it is a common classification algorithm and is supported by Sparks ML library and thus has great documentation. Logistic regression models the probability of an outcome using a logistic function, and classifies a target variable based on the most likely outcome.

Methodology

1. Data Processing: Read in dataset csv and process as a Spark Dataset using Spark Structs
2. Feature Selection: Select 3 groups of features to find the optimal features for predicting a sleep disorder
    - Sleep Model: Sleep Duration, Sleep Quality
    - Physical Activity Model: Physical Activity, Heart Rate, Daily Steps
    - Stress Model: Stress Level, Occupation
3. Model Creation
    - Using Spark's ML library, create a logistic regression model for each feature set
    - Convert categorical variables into numerical (Sleeping Disorder, Occupation)
    - Hyperparameter tuning: test different values of regularization parameter to balance between underfitting and overfitting
    - Cross-Validation: 5-fold cross-validation to test accuracy of models on data independent from training data

Results

Sleep Model Accuracy: 0.580
Physical Activity Model Accuracy: 0.630
Stress Model Accuracy: 0.889

Conclusions

The optimal features for predicting a sleep disorder are stress level and occupation, correctly predicting the sleep disorder 88.9% of the time. This is consistent with the studies we looked at while researching sleep disorders to select high quality feature sets, as stress has been shown to inhibit melatonin production which can lead to sleep disorders.