

Sleep Disorder Prediction and Visualization using Spark's Random Forest Classification



Mata Kuliah Big Data Processing - COMP6579001

Oleh

2602071783 - Darius Felix Lummawie

2602071662 - Devin Fristian Setiawan

2602069236 - Khresna Sariyanto

2602071202 - Sandy Wirananda

BAB I: LATAR BELAKANG DAN PENGENALAN DATA SET

Kesehatan tidur adalah komponen penting dari kesejahteraan manusia yang seringkali dianggap remeh. Pola tidur yang tidak sehat tidak hanya mempengaruhi kehidupan sehari-hari, tetapi juga mempengaruhi kesehatan fisik dan mental. Dalam era modern ini, gaya hidup yang serba cepat, stres dengan pekerjaan, dan pola aktivitas fisik yang berlebihan dapat mengganggu kualitas tidur. Dampak yang didapatkan dari gangguan tidur bisa terlihat dari aspek kesehatan seperti, tekanan darah, detak jantung, dan resiko gangguan tidur seperti insomnia dan sleep apnea.

Penelitian dan analisis mengenai kesehatan tidur dan faktor yang mempengaruhinya menjadi semakin penting untuk membantu individu dan ahli kesehatan memahami kondisi ini dengan lebih baik. Dengan demikian, data yang komprehensif mengenai berbagai aspek yang berhubungan dengan tidur dan gaya dapat memberikan wawasan berharga untuk memperbaiki kualitas hidup dan intervensi kesehatan yang lebih tepat.

Dataset yang digunakan berjudul “Sleep Health and Lifestyle Dataset”, dataset yang dipublikasikan di Kaggle oleh Laksika Tharmalingam. Dataset ini terdiri dari 400 baris dan 13 kolom yang mencakup berbagai variabel yang berkaitan dengan tidur dan kebiasaan sehari-hari. Dataset ini berfungsi sebagai sumber data untuk menganalisis hubungan antara pola tidur dan gaya hidup.

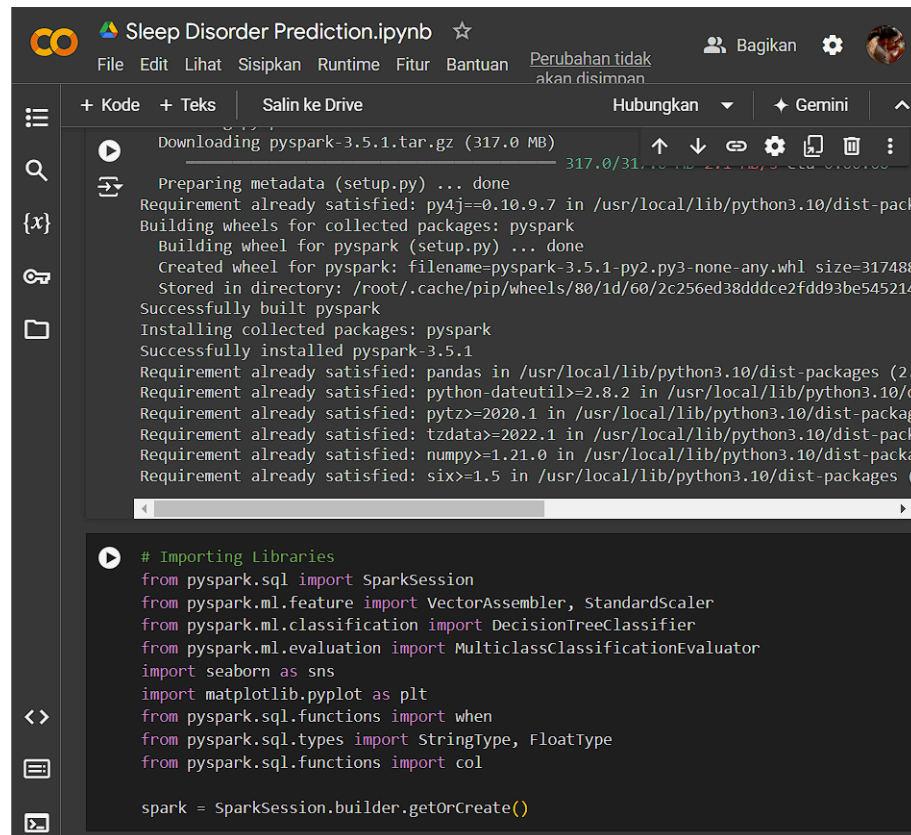
BAB II: METODOLOGI

Berikut adalah tahapan - tahapan yang kami lakukan untuk mengolah dataset sleep health and lifestyle menjadi sebuah model prediksi :

[Link Code](#)

1. Preparation

- a. Tahap pertama memasukan module dan library yang dibutuhkan :
 - SparkSession dari pyspark: Digunakan untuk membaca data dari berbagai sumber
 - VectorAssembler dan StandardScaler dari pyspark: Digunakan untuk menggabungkan berbagai kolom fitur menjadi satu kolom
 - StandardScaler dari pyspark: Digunakan untuk menormalisasi data
 - DecisionTreeClassifier dari pyspark: untuk klasifikasi yang membangun model dalam bentuk Decision Tree Classifier
 - MulticlassClassificationEvaluator: Untuk mengevaluasi kinerja model klasifikasi dengan menghitung metrik
 - Seaborn untuk membuat plot statistik
 - Matplotlib untuk memvisualisasi data
- b. Membaca file .csv dan menampilkannya menggunakan spark



```
Download pyspark-3.5.1.tar.gz (317.0 MB)
Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-pack
Building wheels for collected packages: pyspark
Building wheel for pyspark (setup.py) ... done
Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488
Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38ddce2fdd93be545214
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (2.
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/c
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: numpy>=1.21.0 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (

# Importing Libraries
from pyspark.sql import SparkSession
from pyspark.ml.feature import VectorAssembler, StandardScaler
from pyspark.ml.classification import DecisionTreeClassifier
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
import seaborn as sns
import matplotlib.pyplot as plt
from pyspark.sql.functions import when
from pyspark.sql.types import StringType, FloatType
from pyspark.sql.functions import col

spark = SparkSession.builder.getOrCreate()
```

2. Pre-Processing dan Feature Engineering

- Select fitur - fitur penting (7 dari 11)
- Menghilangkan null values
- Convert data kategorikal menjadi numerik agar lebih mudah untuk diproses
- Split data menjadi dua subset yaitu training Data (80%) dan testing Data (20%)

```
cols = list(SleepHealth_df.columns)
cols.remove("Sleep Disorder")

assembler = VectorAssembler(inputCols=cols, outputCol="features")
SleepHealth_df = assembler.transform(SleepHealth_df)

# Split data (80% for training, 20% for testing)
(trainingData, testingData) = SleepHealth_df.randomSplit([0.8, 0.2])

trainingData.show(10)
testingData.show(10)
```

Berikut hasil data untuk training dan testing secara berurutan:

	Gender	Age	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Daily Steps	Sleep Disorder	features
	0	29	5	40	7	0	4000	1	[0.0,29.0,5.0,40....]
	0	30	5	35	7	0	4100	1	[0.0,30.0,5.0,35....]
	0	30	5	35	7	0	4100	2	[0.0,30.0,5.0,35....]
	0	31	8	75	4	0	6800	0	[0.0,31.0,8.0,75....]
	0	33	6	50	6	1	5500	0	[0.0,33.0,6.0,50....]
	0	33	6	50	6	1	5500	0	[0.0,33.0,6.0,50....]
	0	34	4	32	8	1	5200	2	[0.0,34.0,4.0,32....]
	0	34	4	32	8	1	5200	2	[0.0,34.0,4.0,32....]
	0	35	8	60	4	0	7000	0	[0.0,35.0,8.0,60....]
	0	36	8	60	4	0	7000	0	[0.0,36.0,8.0,60....]
only showing top 10 rows									
	Gender	Age	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Daily Steps	Sleep Disorder	features
	0	29	5	40	7	0	4000	2	[0.0,29.0,5.0,40....]
	0	36	8	60	4	0	7000	0	[0.0,36.0,8.0,60....]
	0	37	8	60	4	0	7000	0	[0.0,37.0,8.0,60....]
	0	37	8	60	4	0	7000	0	[0.0,37.0,8.0,60....]
	0	37	8	60	4	0	7000	0	[0.0,37.0,8.0,60....]
	0	37	8	60	4	0	8000	0	[0.0,37.0,8.0,60....]
	0	38	8	60	4	0	7000	0	[0.0,38.0,8.0,60....]
	0	39	9	80	3	0	7500	0	[0.0,39.0,9.0,80....]
	0	44	7	45	4	1	6000	1	[0.0,44.0,7.0,45....]
	0	45	7	45	4	1	6000	1	[0.0,45.0,7.0,45....]
only showing top 10 rows									

3. Generate Model & Evaluation

- Memakai Decision Tree untuk mempelajari hubungan antara fitur-fitur medis dan demografis dengan diabetes
- Normalisasi fitur - fitur agar skala semua fitur sama dan stabil
- Melatih model dengan menggunakan metode fit dengan dataset training
- Menghasilkan prediksi dengan menggunakan dataset testing
- Mengevaluasi akurasi yang dihasilkan model

- Akurasi model: 92.21%

```
Modeling

[ ] from pyspark.ml.classification import RandomForestClassifier

    rf = RandomForestClassifier(featuresCol='scaledFeatures', labelCol='Sleep Disorder')
    rf_model = rf.fit(trainingData)
    rf_model = rf.fit(trainingData)

    #Gradient Boosting
    rf_predictions = rf_model.transform(testingData)
    #dt_predictions.show()

    #Accuracy
    accuracy = evaluator.evaluate(rf_predictions)
    print("Random Forest Accuracy: {:.2f}%".format(accuracy * 100))

Random Forest Accuracy: 92.21%
```

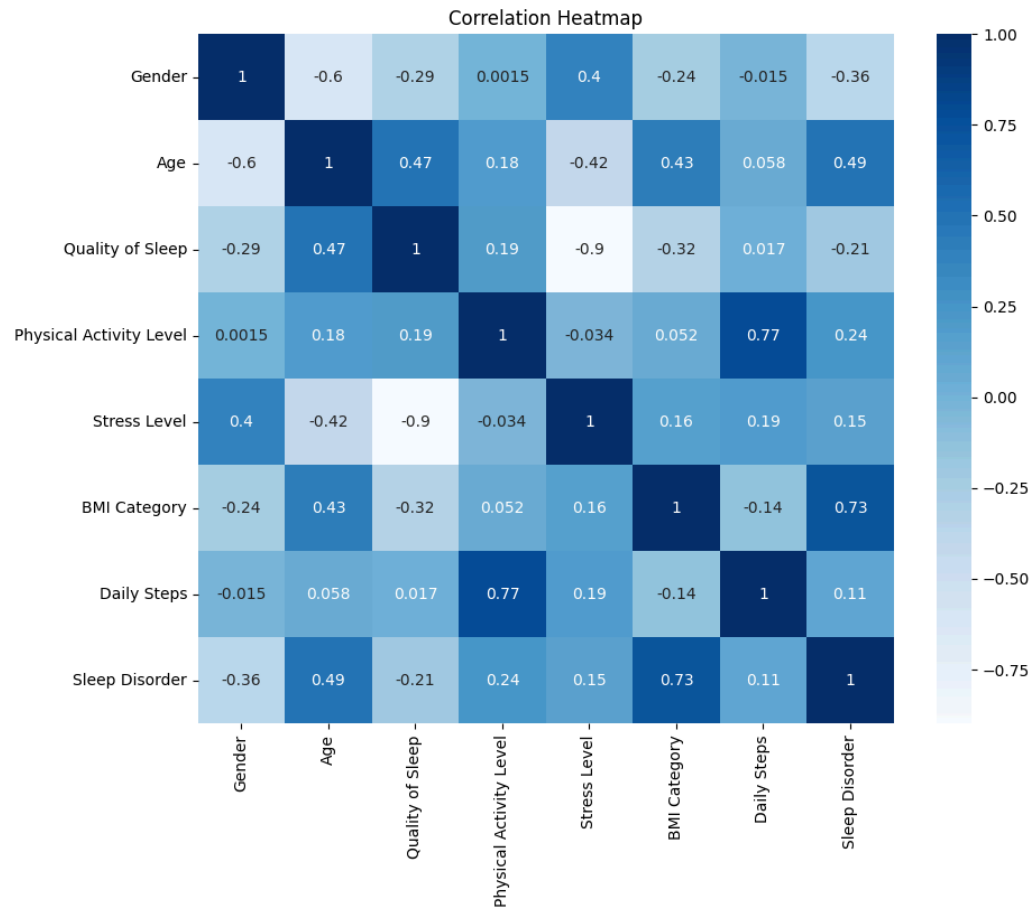
Berikut persentase akurasi model:

```
Random Forest Accuracy: 92.21%
```

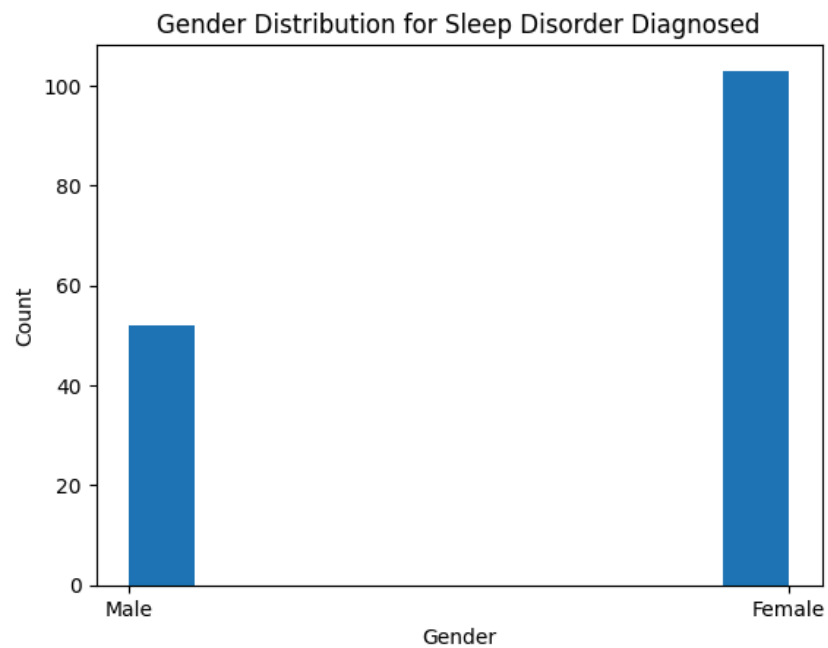
4. Visualization

- Menggunakan matplotlib untuk memahami lebih lanjut isi dari data dan kaitannya dengan memprediksi diabetes
- Selain plot, heat map juga digunakan untuk mengetahui seberapa kuat relasi antar fitur, terutama dengan fitur 'diabetes', yang menentukan apakah seorang pasien mengidap diabetes atau tidak.

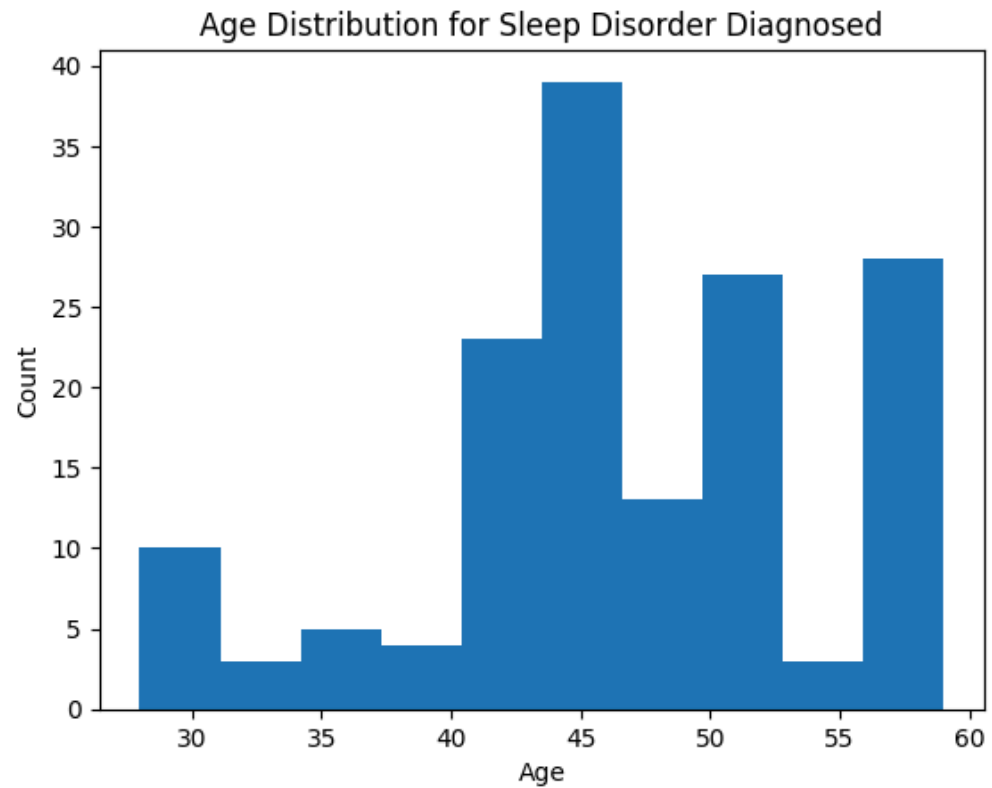
1. Corelasi matrix



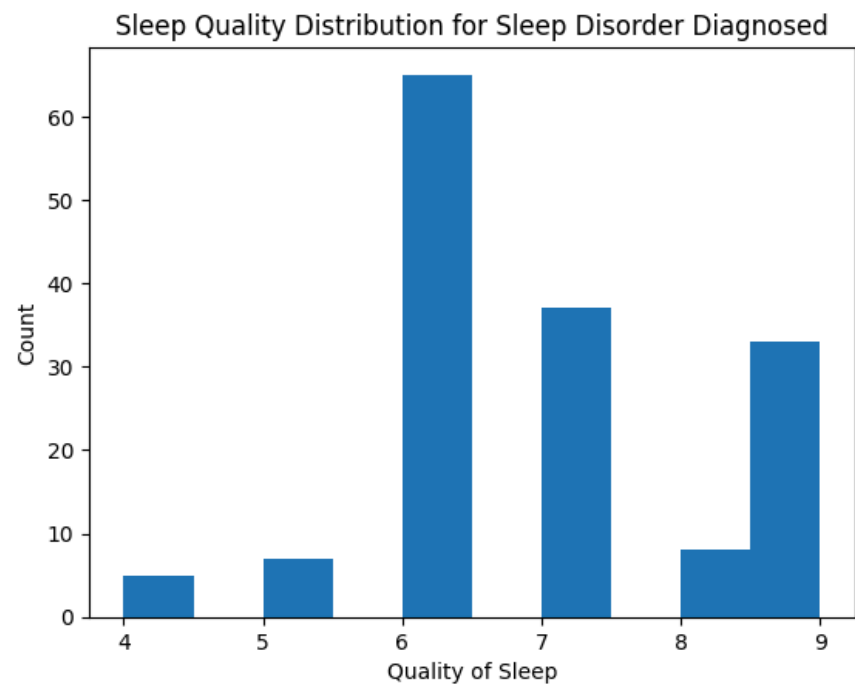
2. Distribusi gender



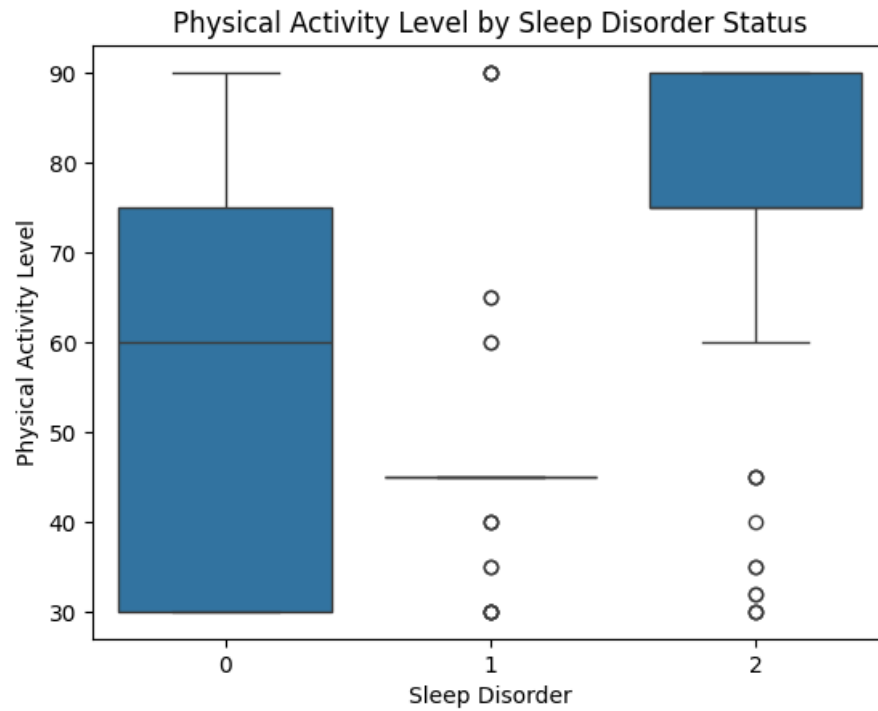
3. Distribusi umur



4. Distribusi kualitas tidur

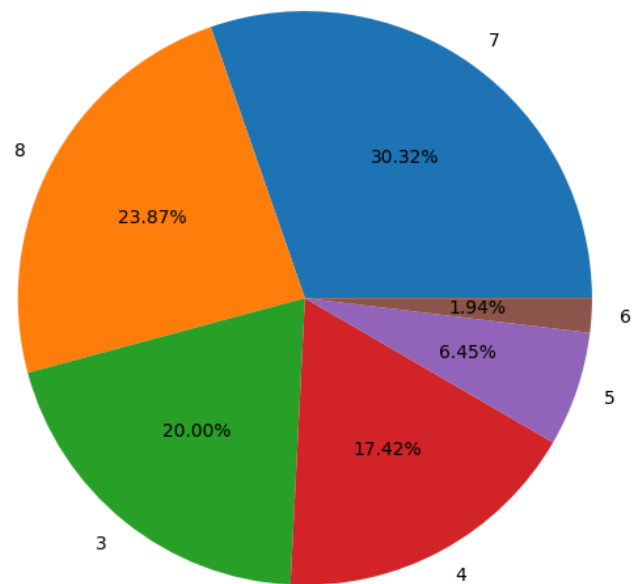


5. Physical activity



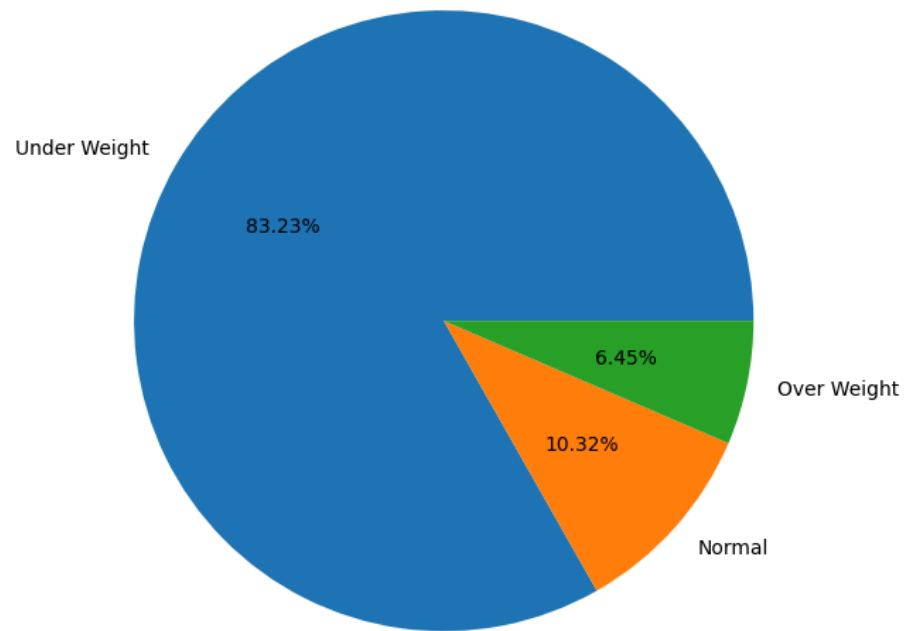
6. Distribusi stress level

Distribution of Stress Level amongst Sleep Disorder Diagnosed

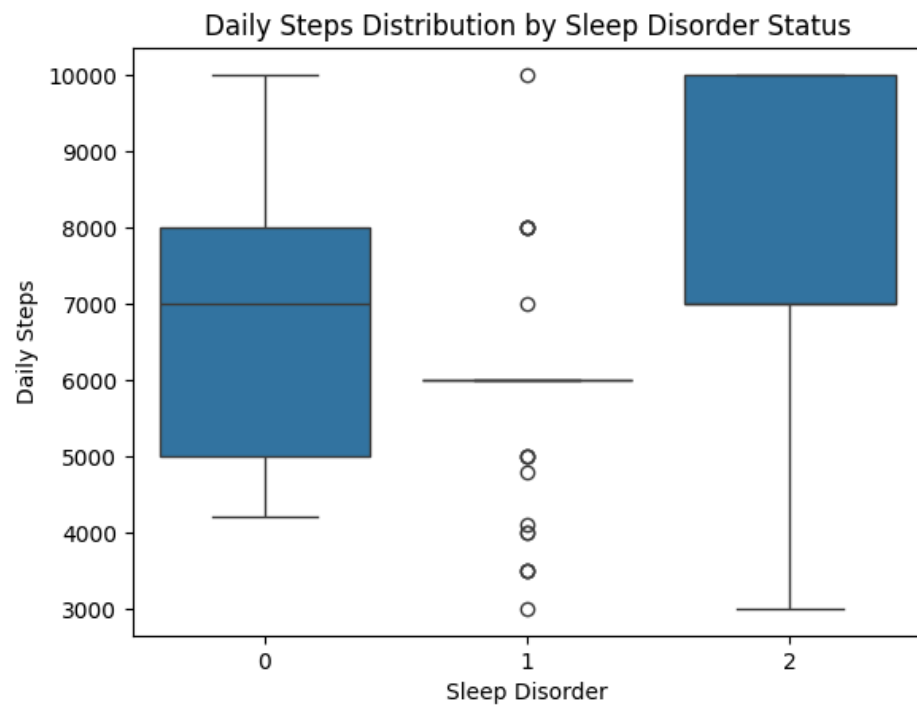


7. Distribusi BMI category

Distribution of BMI Category amongst Sleep Disorder Diagnosed



8. Distribusi Daily Steps



BAB III : HASIL DAN PENEMUAN ANALISIS

Dari model prediksi yang kami sudah gunakan pada dataset kami dapat disimpulkan bahwa dari model RandomForest kami mencapai akurasi 92.21% . Dari sana kita dapat memprediksi dataset dengan pemilihan dari kolom: "Gender" , "Age", "Quality of Sleep", "Physical Activity Level", "Stress Level", "BMI Category", "Daily Steps".

Dari visualisasi yang kami lakukan dapat disimpulkan pada visualisasi berikut:

1. Distribusi Gender

Pada kolom gender dapat disimpulkan bahwa gender dengan berisi Female lebih banyak yang menandakan bahwa wanita banyak mengalami sleep disorder.

2. Distribusi umur

Pada kolom umur dapat disimpulkan bahwa orang yang mengalami gangguan sleep disorder adalah orang yang berumur 40 tahun ke atas.

3. Distribusi kualitas tidur

Pada kolom ini individu menilai sendiri kualitas tidur mereka masing masing dari skala 1 sampai 10. Dari data kami menyimpulkan bahwa byk orang yang menilai 6 dari kualitas tidur mereka.

4. Distribusi aktivitas fisik

Dari kolom tersebut kami menilai bahwa orang yang mengalami gangguan sleep disorder jarang melakukan aktivitas fisik.

5. Distribusi stress

Kami dapat menyimpulkan bahwa orang yang mengalami gangguan disorder mengalami gejala stress. Hal ini dapat disimpulkan dari rating orang-orang pada dataset kami yang memiliki stress level di atas 5 mencapai 50% lebih.

6. Distribusi BMI

Kami menyimpulkan bahwa BMI dari orang-orang yang mengalami gangguan sleep disorder banyak yang memiliki BMI underweight.

Berdasarkan hasil di atas kami menyimpulkan bahwa sebagian besar isi dataset yang kami gunakan menandakan bahwa orang yang berjenis kelamin wanita yang memiliki BMI underweight dan berumur 40 tahun ke atas lebih banyak mengalami gangguan Sleep disorder.

BAB IV : KESIMPULAN

Kesimpulan dari proyek prediksi gangguan tidur ini menunjukkan bahwa dengan menggunakan teknologi machine learning dan analisis data yang canggih, kita dapat secara efektif mengidentifikasi dan memprediksi berbagai jenis gangguan tidur dengan tingkat akurasi yang tinggi. Model prediktif yang telah dikembangkan berhasil memanfaatkan data dari berbagai sumber, termasuk pola tidur, aktivitas harian, dan parameter kesehatan lainnya, untuk memberikan diagnosis awal yang dapat membantu dalam pengambilan keputusan medis. Implementasi dari proyek ini tidak hanya meningkatkan kualitas hidup individu dengan gangguan tidur melalui intervensi yang lebih cepat dan tepat, tetapi juga mengurangi beban pada sistem kesehatan dengan menyediakan alat diagnosis yang lebih efisien dan efektif. Secara keseluruhan, proyek ini menunjukkan potensi besar teknologi dalam bidang kesehatan, khususnya dalam manajemen dan penanganan gangguan tidur.

Referensi

<https://www.halodoc.com/kesehatan/gangguan-tidur>

https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data?select=Sleep_health_and_lifestyle_dataset.csv