Sleep Health and lifestyle analysis

Group members: Tai Le, Haoming Chen

Introduction

Sleeping is an essential activity for maintaining good health and well-being throughout human life

- restore our body through muscle growth, tissue repair, protein synthesis, and hormone growth
- eliminate unwanted chemicals and restructure the synapses inside our brain
- Make you feel energized and ready for the day

Introduction

a lack of quality sleep can have a significant impact on one's physical and mental functioning

- Short term negative effect on
  - emotions, fatigue
  - ability to learn, focus, react, make decisions, and solve problem
- long term negative health consequences including chronic medical conditions like
  - diabetes
  - obesity
  - heart disease

Introduction

Our goal is for this project..

- apply data mining techniques to a large public health dataset to find patterns between sleep quality, personal attributes and health habits.
- develop a model that can predict signs of bad quality sleep so that proactive measures can be taken to prevent the problem

Preprocessing

Dataset: The 2022 Behavioral Risk Factor Surveillance System (BRFSS) dataset

- a total of 326 columns features
- a total of 445132 valid data entries.

Preprocessing

Feature selection:

- Target label: ‘sleep_time’
- Feature columns: 28 in total
    - states, sex, age, education, income
    - IBM, exercise, physical health, mental health, stress, life satisfaction
    - cigarettes, tobacco, e-cigarettes, marijuana, drinking
    - heart attack, coronary heart disease, stroke, asthma, kidney disease, arthritis, diabetes, depressive disorder

Preprocessing

In order to measure sleep quality, we applied a transformation to our target label ‘sleep_time’

Healthy_Sleep = 1, if 6 < ‘sleep_time’ < 10

Healthy_Sleep = 0, if ‘sleep_time’ < 7 and ‘sleep_time’ is > 9

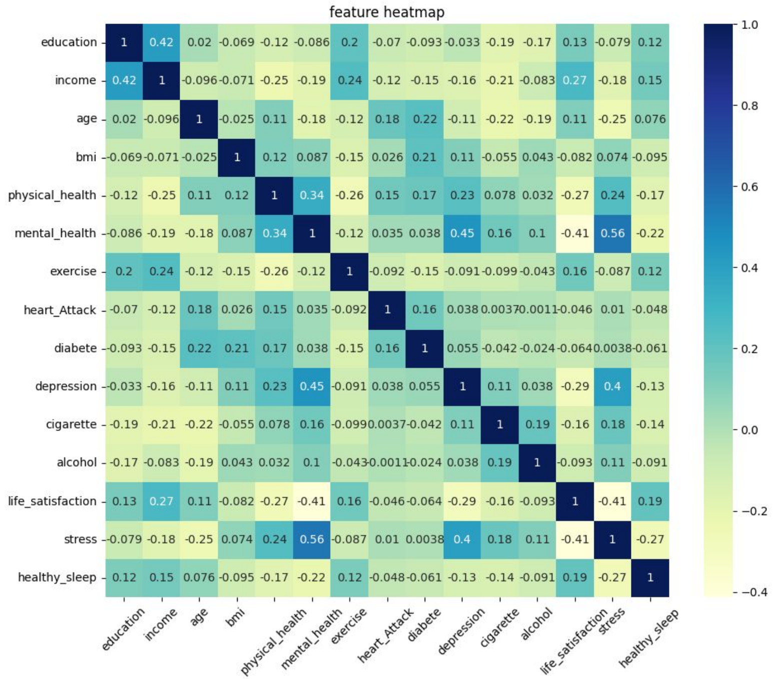
Preprocessing

Feature transformation:

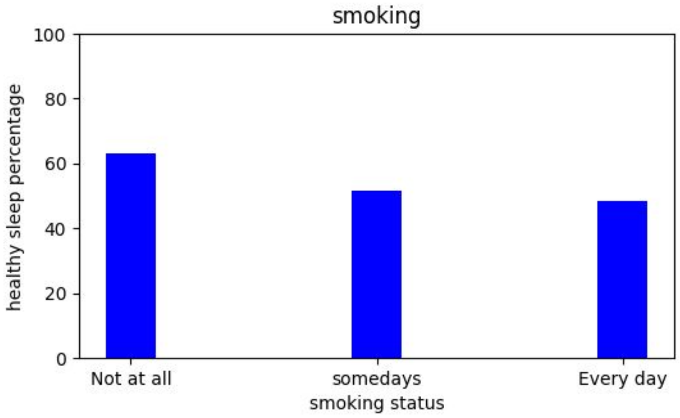
- Smoking:

Categorical	Numerical
Everyday	2
Someday	1
Not at all	0

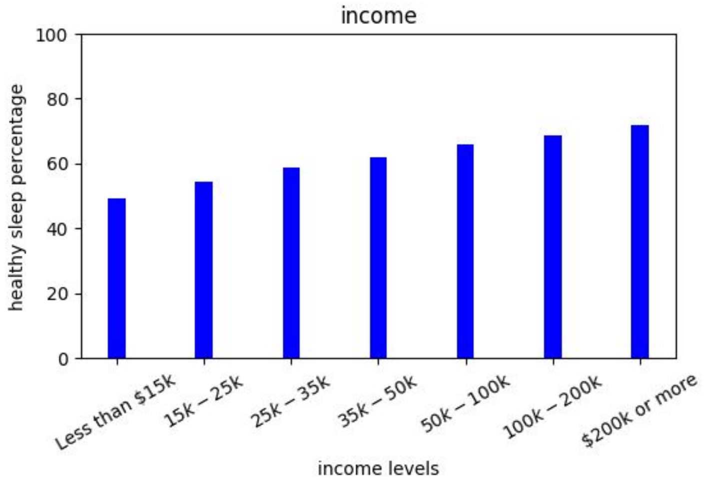
Analysis


feature heatmap
education 1 0.42 0.02 -0.069 -0.12 -0.086 0.2 -0.07 -0.093 -0.033 -0.19 -0.17 0.13 -0.079 0.12
income 0.42 1 -0.096 -0.071 -0.25 -0.19 0.24 -0.12 -0.15 -0.16 -0.21 -0.083 0.27 -0.18 0.15
age 0.02 -0.096 1 -0.025 0.11 -0.18 -0.12 0.18 0.22 -0.11 -0.22 -0.19 0.11 -0.25 0.076
bmi -0.069 -0.071 -0.025 1 0.12 0.087 -0.15 0.026 0.21 0.11 -0.055 0.043 -0.082 0.074 -0.095
physical_health -0.12 -0.25 0.11 0.12 1 0.34 -0.26 0.15 0.17 0.23 0.078 0.032 -0.27 0.24 -0.17
mental_health -0.086 -0.19 -0.18 0.087 0.34 1 -0.12 0.035 0.038 0.45 0.16 0.1 -0.41 0.56 -0.22
exercise 0.2 0.24 -0.12 -0.15 -0.26 -0.12 1 -0.092 -0.15 -0.091 -0.099 -0.043 0.16 -0.087 0.12
heart_Attack -0.07 -0.12 0.18 0.026 0.15 0.035 -0.092 1 0.16 0.038 0.0037 -0.0011 -0.046 0.01 -0.048
diabete -0.093 -0.15 0.22 0.21 0.17 0.038 -0.15 0.16 1 0.055 -0.042 -0.024 -0.064 0.0038 -0.061
depression -0.033 -0.16 -0.11 0.11 0.23 0.45 -0.091 0.038 0.055 1 0.11 0.038 -0.29 0.4 -0.13
cigarette -0.19 -0.21 -0.22 -0.055 0.078 0.16 -0.099 0.0037 -0.042 0.11 1 0.19 -0.16 0.18 -0.14
alcohol -0.17 -0.083 -0.19 0.043 0.032 0.1 -0.043 -0.0011 -0.024 0.038 0.19 1 -0.093 0.11 -0.091
life_satisfaction 0.13 0.27 0.11 -0.082 -0.27 -0.41 0.16 -0.046 -0.064 -0.29 -0.16 -0.093 1 -0.41 0.19
stress -0.079 -0.18 -0.25 0.074 0.24 0.56 -0.087 0.01 0.0038 0.4 0.18 0.11 -0.41 1 -0.27
healthy_sleep 0.12 0.15 0.076 -0.095 -0.17 -0.22 0.12 -0.048 -0.061 -0.13 -0.14 -0.091 0.19 -0.27 1
education income age bmi physical_health mental_health exercise heart_Attack diabete depression cigarette alcohol life_satisfaction stress healthy_sleep
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Analysis



smoking
100
80
60
40
20
0
healthy sleep percentage
Not at all
somedays
Every day
smoking status

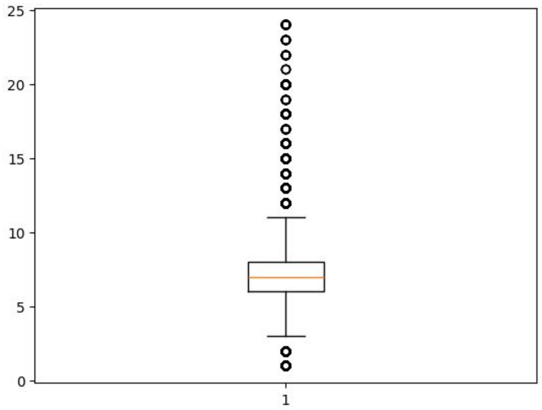


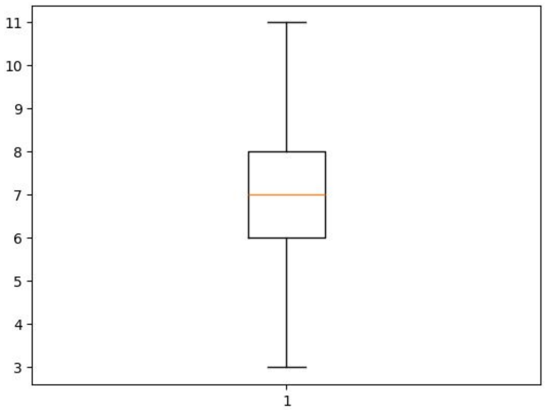
income
100
80
60
40
20
0
healthy sleep percentage
Less than $15k
15k – 25k
25k – 35k
35k – 50k
50k – 100k
100k – 200k
$200k or more
income levels

Modeling

Modeling

Removing outliers based on target label ‘sleep_time’:





Modeling

Create new ylabel ‘bad_sleep’ from original target label ‘sleep_time’

bad_sleep = 0, if 6 < ‘sleep_time’ < 10

bad_sleep = 1, if ‘sleep_time’ < 6 and ‘sleep_time’ is > 9

	bad_sleep
0	291829
1	152840

Modeling

Compute imputed values (scikit-learn library):

- One Hot Encoder
  - Transform state feature into a 50 independent binary data columns
- Simple Imputer
  - Fill in all missing values with most-frequent values within each column

```
RangeIndex: 444669 entries, 0 to 444668
Data columns (total 73 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   sexvar      444669 non-null  float64
 1   educa       444669 non-null  float64
 2   income3     444669 non-null  float64
 3   x.age80     444669 non-null  float64
 4   x.age.g     444669 non-null  float64
 5   x.bmi5      444669 non-null  float64
 6   physhlth    444669 non-null  float64
 7   menthlth    444669 non-null  float64
 8   exerany2    444669 non-null  float64
 9   cvdinfr4    444669 non-null  float64
 10  cvdcrhd4    444669 non-null  float64
 11  cvdstrk3    444669 non-null  float64
 12  diabete4    444669 non-null  float64
 13  addepev3    444669 non-null  float64
 14  smokday2    444669 non-null  float64
 15  avedrnk3    444669 non-null  float64
 16  lsatisfy    444669 non-null  float64
 17  sdhstre1    444669 non-null  float64
 18  bad_sleep   444669 non-null  float64
 19  x.state_1   444669 non-null  float64
 20  x.state_2   444669 non-null  float64
 21  x.state_4   444669 non-null  float64
```

Modeling

Data imbalance:

- upsampling the minority class in the target label
    - Before upsampling:

0	291829
1	152840

    - After upsampling:

0	291829
1	291829

Modeling

Logistic Regression:

logistic_regressor = LogisticRegression(C = 0.1)

Training time: 10 sec

	Train	Test
Accuracy	0.61602	0.61209
Precision	0.63223	0.62843
Recall	0.55468	0.54910
F1	0.59092	0.58609

Modeling

Soft Vector Machine (SVM):

dataset_svm = dataset.sample(n=15000)

svm_model = SVC(kernel = 'rbf', C = 0.1)

Training time: 30 sec

	Train	Test
Accuracy	0.62025	0.60133
Precision	0.63852	0.62973
Recall	0.57816	0.56510
F1	0.60685	0.59567

Modeling

Decision tree:

decision_tree = DecisionTreeClassifier(min_samples_leaf=500)

Training time: 10 sec

	Train	Test
Accuracy	0.62889	0.62398
Precision	0.63661	0.63174
Recall	0.60052	0.59510
F1	0.61804	0.61287

Modeling

Random Forest Tree:

random_forest_decision_tree = RandomForestClassifier(n_estimators=100, max_depth=10)

	Train	Test
Accuracy	0.63100	0.62511
Precision	0.64146	0.63481
Recall	0.59395	0.58969
F1	0.61679	0.61142

Training time: 1 min

Modeling

Gradient Boosting Tree:

gradient_boosting_classifier = GradientBoostingClassifier(n_estimators=500, learning_rate=0.2)

	Train	Test
Accuracy	0.63670	0.63038
Precision	0.64454	0.63761
Recall	0.60950	0.60469
F1	0.62653	0.62072

Training time: 3 min

Modeling

Kth-Nearest Neighbor (KNN):

knn_classifier = KNeighborsClassifier(n_neighbors = 7)

	Train	Test
Accuracy	0.76686	0.65519
Precision	0.74697	0.64205
Recall	0.80720	0.70018
F1	0.77592	0.66985

Inference time: 6 min

Modeling

Kth-Nearest Neighbor (KNN) variation:

knn_classifier = KNeighborsClassifier(n_neighbors = 7, weights='distance')

Inference time: 6 min

	Train	Test
Accuracy	0.99840	0.76413
Precision	0.99827	0.70414
Recall	0.99851	0.91131
F1	0.99839	0.79444

Modeling

KNN cross validation:

Mean: 0.76271

standard deviation: 0.00134

	run#1	run#2	run#3	run#4	run#5
score	0.76379	0.76021	0.76398	0.76258	0.76297

Modeling

In conclusion:

- KNN has the best performance but suffers from long inference time
- SVM fits the dataset poorly and training time increases exponentially as the dataset increases
- Decision tree provide the best balance between good performance, short training time and low inference cost

Modeling

Challenges and future improvements:

- Data quality, noise in the dataset
- Feature selection and engineering

Community contribution

- Monitor individual well-being, providing insights to user’s sleep pattern
- Early detection of sleep issue
- Raising awareness about user's sleep health
- Allowing users to take proactive measures to improve their sleep quality
- Access to a larger database for more realistic result with different users and improve the model accuracy

Q&A