МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА»

институт информационных технологий и технологического образования
кафедра информационных технологий и электронного обучения

Основная профессиональная образовательная программа
Направление подготовки 09.03.01 Информатика и вычислительная техника
Направленность (профиль) «Технологии разработки программного обеспечения»
форма обучения – очная

Курсовая работа

по дисциплине «Пакеты прикладных программ для статистической обработки и анализа данных»

Статистический анализ качества сна и образа жизни

Обучающегося 3 курса
Воложанина Владислава Олеговича

_____

Руководитель:
д.п.н, профессор
_____Власова Е. З.

«_____» _____ 2024 г.

Санкт-Петербург
2024

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Данная курсовая работа направлена на изучение статистических методов анализа. Целью данной работы является изучение и анализ статистических данных о качестве сна и образе жизни людей, выявление взаимосвязей и зависимостей между качеством сна и различными аспектами образа жизни, а также определение факторов, влияющих на здоровье сна.

Обоснование актуальности курсовой работы:

Актуальность темы обусловлена растущим интересом к изучению здоровья сна в современном обществе, где уровень стресса, неправильное питание и недостаток физической активности влияют на качество жизни. Понимание связей между образом жизни и качеством сна может способствовать разработке эффективных стратегий для улучшения общего состояния здоровья населения.

Объектом исследования являются данные опроса. Предмет исследования – определение связи между качеством сна и образом жизни.

Основная часть курсовой работы состоит из двух глав: теоретический – блок теории, в которой дается обзор методов статистического анализа, таких методов, как: корреляционный анализ, регрессионный анализ, кластеризация, метод машинного обучения. Практический – блок посвящен реализации статистического анализа.

1.Теоретическая часть

1.1 Введение в корреляционный анализ

Корреляционный анализ — это метод по изучению статистической зависимости между случайными величинами с не обязательным наличием строгого функционального характера, при которой динамика одной случайной величины приводит к динамике математического ожидания другой.

Если обозначить через X независимую переменную, а через Y зависимую от нее переменную. Зависимость Y от X называется функциональной, если каждому значению величины X соответствует единственное значение величины Y. Однако гораздо чаще в окружающем нас мире каждому значению переменной X соответствует не одно, а множество значений переменной Y. Причем, заранее сказать, какое именно значение примет величина Y, нельзя. Это объясняется тем, что на результирующую переменную действует не только контролируемый фактор X, а множество других неконтролируемых случайных факторов. Таким образом, каждому значению переменной X соответствует определенное распределение переменной Y. Такая зависимость называется стохастической, или вероятностной.

Корреляционной зависимостью между двумя переменными называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.

Корреляционная зависимость может быть представлена в виде:

$$\varphi(x) = M_x(Y) \tag{1}$$

Или

$$\psi(y) = M_x(X) \tag{2}$$

Уравнения (1) и (2) называются уравнениями регрессии соответственно Y на X и X на Y, а их графики – линиями регрессии. Если

$$\varphi(x) = const, \psi(y) = const, \tag{3}$$

то говорят, что корреляционная связь отсутствует. Переменная X в корреляционном анализе может быть как детерминированной, так и случайной.

Метод корреляционного анализа может применяться в том случае, если имеется большое количество наблюдений о величине результативных и факторных показателей. При этом исследуемые факторы должны быть количественными и отражаться в конкретных источниках. Первое может определяться нормальным законом — в этом случае результатом корреляционного анализа выступают коэффициенты корреляции Пирсона, либо в случае, если признаки не подчиняются этому закону, используется коэффициент ранговой корреляции Спирмена.

Правила отбора факторов корреляционного анализа. При применении данного метода необходимо определиться с факторами, оказывающими влияние на результативные показатели. Их отбирают с учетом того, что между показателями должны присутствовать причинно следственные связи. В случае создания многофакторной корреляционной модели отбирают те из них, которые оказывают существенное влияние на результирующий показатель, при этом взаимозависимые факторы с коэффициентом парной корреляции более 0,85 в корреляционную модель предпочтительно не включать, как и такие, у которых связь с результативным параметром носит не прямолинейный или функциональный характер.

1.2 Коэффициент корреляции

Количественная оценка тесноты взаимосвязи двух случайных величин осуществляется с помощью коэффициента корреляции. Вид коэффициента корреляции и, следовательно, алгоритм его вычисления зависят от шкалы, в которой производятся измерения изучаемых показателей и от формы зависимости. Значение коэффициента корреляции может изменяться в диапазоне от -1 до +1. Абсолютное значение коэффициента корреляции показывает силу взаимосвязи. Чем меньше его абсолютное значение, тем слабее связь. Если он равен нулю, то связь вообще отсутствует. Чем больше значение модуля коэффициента корреляции, тем сильнее связь и тем меньше разброс в значениях *yi* при каждом фиксированном значении *xi*. Знак коэффициента корреляции определяет направленность взаимосвязи: минус – отрицательная, плюс – положительная.

Для оценки степени взаимосвязи величин, измеренных в количественных шкалах, используется коэффициент линейной корреляции, предполагающий, что выборки X и Y распределены по нормальному закону.

Коэффициент корреляции — параметр, который характеризует степень линейной взаимосвязи между двумя выборками, рассчитывается по формуле:

$$r_{xy} = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}} \quad (4)$$

Степень зависимости двух случайных величин X и Y может характеризоваться на основе анализа получаемых результатов (*x1, y1*), … (*xn, yn*). Когда вид распределения неизвестен используют меры связи, не регламентирующие нормальность выборок. Например, коэффициент ранговой корреляции Спирмена:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (5)$$

Где:

d_i^2 - квадраты разности рангов,

n - число наблюдений,

ранг — это порядковый номер значений, расположенных в порядке возрастания или убывания их величин.

1.3 Корреляционная матрица

Корреляционная матрица — это таблица, показывающая коэффициенты корреляции между переменными. Каждая ячейка таблицы показывает корреляцию между двумя переменными. Значение находится в диапазоне от -1 до 1. Если две переменные имеют высокую корреляцию, это означает, что они имеют тенденцию изменяться аналогичным образом.

Высокая положительная корреляция, близкая к 1 подразумевает, что по мере увеличения одной переменной увеличивается и другая.

Высокая отрицательная корреляция, близкая к -1 означает, что по мере увеличения одной переменной другая уменьшается. Корреляция, близкая к 0, указывает на отсутствие линейной зависимости между переменными.

Основная цель корреляционной матрицы заключается в обобщении данных, в качестве входных данных для более продвинутого анализа и в качестве диагностического средства для расширенного анализа. Это помогает понять силу и направление взаимосвязи между переменными.

Часто корреляционные матрицы представлены в виде тепловых карт или матриц с цветовой кодировкой. Эти визуальные инструменты облегчают понимание сложных взаимосвязей между несколькими переменными.

Интерпретация значений: +1: Идеальная положительная корреляция, -1: Абсолютная отрицательная корреляция, 0: Корреляция отсутствует

Значения от 0 до ±1 обозначают степень линейной зависимости между переменными.

1.4 Тепловая карта

Тепловая карта в контексте корреляционной матрицы используется для отображения коэффициентов корреляции между каждой парой рассматриваемых переменных. Она предлагает систему цветового кодирования для быстрого определения силы и типа корреляций, что облегчает интерпретацию сложных наборов данных.

На тепловой карте каждая ячейка представляет корреляцию между двумя переменными. Интенсивность цвета или оттенок в каждой ячейке соответствуют величине коэффициента корреляции: теплые цвета, такой как красный обычно представляют положительную корреляцию, холодные цвета, такой как синий указывают на отрицательную корреляцию, интенсивность цвета соответствует силе корреляции, при этом более темные оттенки обычно указывают на более сильную корреляцию.

Диагональная линия на стандартной тепловой карте корреляционной матрицы диагональ (слева вверху- справа внизу) часто показывает идеальную положительную корреляцию (1, 0), поскольку она представляет корреляцию каждой переменной с самой собой.

Матрица обычно симметрична относительно диагонали, поскольку корреляция между A и B такая же, как корреляция между B и A.

Значения на тепловой карте представляют собой коэффициенты корреляции. Они количественно определяют степень линейной зависимости между парами переменных.

1.5 Регрессия. Метод «наименьших квадратов»

Метод наименьших квадратов — это метод оценки неизвестных параметров в модели линейной регрессии. Метод наименьших квадратов минимизирует сумму квадратов различий между наблюдаемой зависимой переменной в данном наборе данных и теми, которые предсказываются линейной функцией.

Основная цель регрессии - провести линию через точечную диаграмму точек данных, которая минимизирует сумму квадратов вертикальных расстояний точек от линии.

В методе наименьших квадратов линия наилучшего соответствия для набора точек данных определяется путем минимизации суммы квадратов остатков, разностей между наблюдаемыми значениями и значениями, предсказанными моделью.

Математически, если y_i- наблюдаемые значения, а $\hat{y}_i$ - значения, предсказанные линейной моделью, OLS стремится минимизировать $\sum(y_i - \hat{y}_i)^2$.

В простой модели линейной регрессии $y = \beta_0 + \beta_1 x + \epsilon$ оценивает параметры β_0 и β_1, которые наилучшим образом соответствуют данным.

Метод наименьших квадратов также может быть применена к моделям многомерной регрессии, включающим более одной предиктивной переменной.

Коэффициенты регрессионной модели оцениваются с использованием критерия наименьших квадратов. Статистические программные пакеты, такие как R, Python (с библиотеками, такими как pandas и statsmodels) и другие, могут эффективно выполнять оценку.

После оценки параметров модели для определения значимости параметров используются статистические тесты, такие как *t*-тесты и *F*-тесты.

1.6 Кластеризация по K-значениям.

Кластеризация по K-значениям относится к категории алгоритмов в науке о данных и статистике, которые группируют точки данных на основе определенных признаков в заранее определенное количество кластеров. Одним из наиболее широко используемых методов в этой категории является кластеризация по K-значениям.

K-means — это итеративный алгоритм, который разбивает набор данных на K отдельных, неперекрывающихся подгрупп, где каждая точка данных принадлежит только одной группе. Он направлен на минимизацию различий внутри каждого кластера.

Процесс включает в себя выбор K начальных центроидов, по одному для каждого кластера. Затем точки данных присваиваются ближайшему центроиду, и центроиды пересчитываются. Этот процесс повторяется до тех пор, пока центроиды не перестанут существенно изменяться.

Метод Elbow используется для определения оптимального количества кластеров при кластеризации с использованием k-средних значений. Метод отображает объясненное изменение в зависимости от количества кластеров, и пользователь должен выбрать количество кластеров таким образом, чтобы добавление другого кластера не дало намного лучшего моделирования данных.

Чтобы выполнить кластеризацию k-средних в наборе данных для диапазона значений k. Надо для каждого значения k вычислите общую сумму квадратов внутри кластера (wcss), построить кривую wcss в соответствии с количеством кластеров k.

Расположение изгиба на графике обычно считается показателем соответствующего количества кластеров. Эта точка «сгиба» представляет собой точку, в которой добавление большего количества кластеров существенно не улучшает объясненную дисперсию.

Метод силуэта измеряет, насколько хорошо каждый объект был кластеризован. Если объект хорошо сгруппирован в своем кластере, то его силуэт близок к 1. Если объект попал в неправильный кластер, его силуэт близок к -1.

Чтобы выполнить вычисление силуэта, надо для каждого объекта вычислить среднее расстояние до объектов в том же кластере a, затем вычисляется среднее расстояние до объектов в ближайшем другом кластере b. Индекс силуэта для каждого объекта равен:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

Где:

a(i) — среднее расстояние от i-го объекта до других объектов в том же кластере,

b(i) — минимальное среднее расстояние от i-го объекта до объектов в другом кластере, не считая его собственный кластер.

K-medoids или PAM (разделение по медоидам): использует фактические точки данных в качестве центров и более устойчив к шуму и выбросам.

1.7 Модель случайного леса.

Алгоритм случайного леса (Random Forest) — универсальный алгоритм машинного обучения, суть которого состоит в использовании ансамбля решающих деревьев. Само по себе решающее дерево предоставляет крайне невысокое качество классификации, но из-за большого их количества результат значительно улучшается. Также это один из немногих алгоритмов, который можно использовать в абсолютном большинстве задач.

Благодаря своей гибкости Random Forest применяется для решения практически любых проблем в области машинного обучения. Сюда относятся классификации (RandomForestClassifier) и регрессии (RandomForestRegressor), а также более сложные задачи, вроде отбора признаков, поиска выбросов/аномалий и кластеризации.

Основным полем для применения алгоритма случайного дерева являются первые два пункта, решение других задач строится уже на их основе.

По сравнению с другими методами машинного обучения, теоретическая часть алгоритма Random Forest проста. У нее нет большого объема теории, необходима только формула итогового классификатора a(x):

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x) \qquad (7)$$

Где:

N – количество деревьев,

i – счетчик для деревьев,

b – решающее дерево,

x – сгенерированная нами на основе данных выборка.

1.8 Реализация алгоритма Random Forest.

Алгоритм Random Forest для задачи классификации, используя библиотеку *scikit-learn*. Вычисляет площадь под кривой ошибок для тренировочной и тестовой частей модели, чтобы определить ее качество необходимые параметры алгоритма: число деревьев – *n_estimators*, чем больше деревьев, тем лучше качество. Часто при большом увеличении *n_estimators* качество на обучающей выборке может даже доходить до 100%, в то время как качество на тесте выходит на асимптоту, что сигнализирует о переобучении нашей модели.

Также один из самых важных параметров для построения. В библиотеке *sklearn* для задач классификации реализованы критерии *gini* и *entropy*. Они соответствуют классическим критериям расщепления. В свою очередь, для задач регрессии реализованы два критерия (*mse* и *mae*), которые являются функциями ошибок *Mean Square Error* и *Mean Absolute Error* соответственно. Практически во всех задачах используется критерий *mse*.

2. Практическая часть
РАЗРАБОТКА ПРОГРАММЫ

Для создания кода использовался VSCode, а использовался язык программирования python.

2.1 Импорт необходимых библиотек и чтение файла

На рисунке 1 изображены строки, которые импортируют необходимые библиотеки для задач. Библиотеки включают scikit-learn для машинного обучения, pandas для манипулирования данными, seaborn и matplotlib для визуализации и statsmodels для статистического моделирования.

```
from sklearn.cluster import KMeans
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
```

Рис. 1 – импорт библиотек

На рисунке 2 изображена загрузка набора данных, относящихся к здоровью сна и образу жизни, из CSV-файла во фрейм данных pandas.

```
file_path = '/Users/mac/Desktop/Sleep_health_and_lifestyle_dataset.csv'
sleep_data = pd.read_csv(file_path)
```

Рис. 2 – чтение файла

2.2 Реализация корреляционной матрицы

На рисунках 3 и 4 скрипт выбирает определенные интересующие столбцы и вычисляет матрицу корреляции между этими переменными, и визуализирует корреляционную матрицу в виде тепловой карты.

```
columns_of_interest = ['Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Stress Level', 'Heart Rate']
correlation_data = sleep_data[columns_of_interest]
correlation_matrix = correlation_data.corr()
```

Рис. 3 – определение столбцов

```
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Matrix of Sleep Duration, Quality of Sleep, Physical Activity, Stress Level, and Heart Rate")
plt.show()
```

Рис. 4 – вычисление матрицы

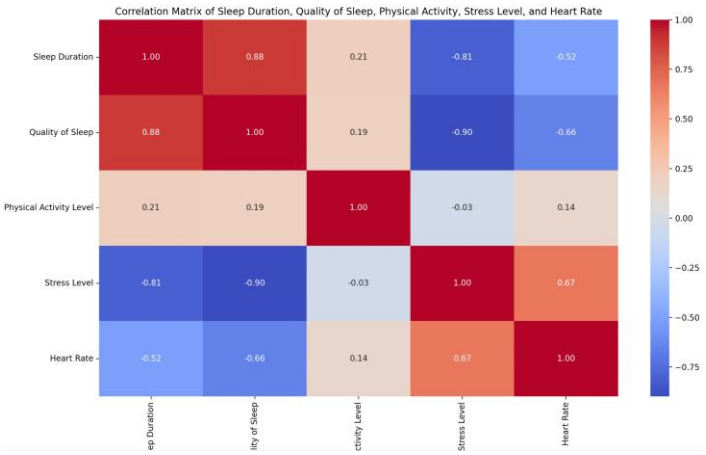

Рис. 5 – тепловая карта

2.3 Реализация линейного регрессионного анализа

Блок кода, который указан на рисунке 6 выполняет линейный регрессионный анализ для моделирования взаимосвязи между "Качеством сна" и другими переменными. Он соответствует модели, делает прогнозы и выводит сводную информацию о модели.

```
X = sleep_data[['Sleep Duration', 'Physical Activity Level', 'Stress Level', 'Heart Rate']]
y = sleep_data['Quality of Sleep']
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
predictions = model.predict(X)
model_summary = model.summary()
print(model_summary)
```

Рис. 6 – выполнение регрессионного анализа

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Quality of Sleep   R-squared:                       0.897
Model:                           OLS   Adj. R-squared:                  0.896
Method:                Least Squares   F-statistic:                     805.5
Date:               Sun, 07 Jan 2024   Prob (F-statistic):          8.02e-181
Time:                       14:20:23   Log-Likelihood:                -171.93
No. Observations:                374   AIC:                             353.9
Df Residuals:                    369   BIC:                             373.5
Df Model:                          4
Covariance Type:           nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                      7.4881      0.567     13.203      0.000       6.373       8.603
Sleep Duration             0.6050      0.045     13.376      0.000       0.516       0.694
Physical Activity Level    0.0065      0.001      6.293      0.000       0.004       0.009
Stress Level              -0.3126      0.022    -13.966      0.000      -0.357      -0.269
Heart Rate                -0.0455      0.007     -6.827      0.000      -0.059      -0.032
==============================================================================
Omnibus:                       14.003   Durbin-Watson:                   0.952
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               14.582
Skew:                          -0.476   Prob(JB):                     0.000682
Kurtosis:                       3.169   Cond. No.                     2.66e+03
==============================================================================
```

Рис. 7 – вывод результатов

2.4 Реализация кластеризации

На рисунке 8 часть кода, которая подготавливает данные для кластеризации по K-среднему значению путем выбора соответствующих столбцов и их нормализации.

```
cluster_columns = ['Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Stress Level', 'Heart Rate']
cluster_data = sleep_data[cluster_columns]
normalized_data = (cluster_data - cluster_data.mean()) / cluster_data.std()
```

Рис. 8 – подготовка данных

Здесь скрипт запускает кластеризацию K-Means для разного количества кластеров (от 1 до 10), чтобы найти оптимальное число, используя метод elbow.

```
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
    kmeans.fit(normalized_data)
    wcss.append(kmeans.inertia_)
```

Рис. 9 – выполнение кластеризации

В этом разделе визуализируется сумма квадратов внутри кластера (WCSS), чтобы помочь определить точку сгиба.

```
plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), wcss)
plt.title('Elbow Method For Optimal k')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

Рис. 10 – визуализация

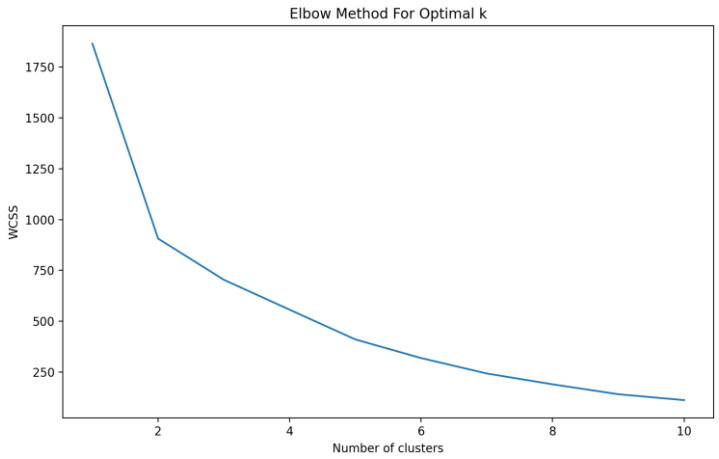

рис. 11 – график суммы квадратов кластера

Выбирается оптимальное количество кластеров, и к данным применяется модель K-средних. Полученные кластеры добавляются в набор данных и печатаются первые несколько строк.

```
optimal_clusters = 4
kmeans = KMeans(n_clusters=optimal_clusters, init='k-means++', max_iter=300, n_init=10, random_state=0)
clusters = kmeans.fit_predict(normalized_data)
sleep_data['Cluster'] = clusters
print(sleep_data.head())
```

Рис. 12 – выбор оптимального количества кластеров

2.5 Реализация алгоритма Random Forest

Эта часть разбивает данные на обучающий и тестовый наборы, подбирает случайный лесной регрессор для обучающего набора и делает прогнозы для тестового набора.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
rf_model = RandomForestRegressor(n_estimators=100, random_state=0)
rf_model.fit(X_train, y_train)
y_pred = rf_model.predict(X_test)
```

Рис 13 – разбив обучающих и тестовых данных

```
   Person ID Gender  Age          Occupation  Sleep Duration  Quality of Sleep  ...  BMI Category  Blood Pressure Heart Rate Daily Steps  Sleep Disorder  Cluster
0          1   Male   27     Software Engineer             6.1                 6  ...    Overweight         126/83         77        4200             NaN        2
1          2   Male   28                Doctor             6.2                 6  ...        Normal         125/80         75       10000             NaN        2
2          3   Male   28                Doctor             6.2                 6  ...        Normal         125/80         75       10000             NaN        2
3          4   Male   28  Sales Representative             5.9                 4  ...         Obese         140/90         85        3000     Sleep Apnea        2
4          5   Male   28  Sales Representative             5.9                 4  ...         Obese         140/90         85        3000     Sleep Apnea        2

[5 rows x 14 columns]
Mean Squared Error: 0.026639870825691962
R-squared: 0.9792624863832664
Predicted Quality of Sleep: [8.         6.         6.75882251 8.         8.         8.
 6.19185745 7.17       6.         6.         9.         8.
 6.         6.         9.         8.         6.         7.07
 9.         9.         8.         7.         6.         6.19185745
 9.         6.         6.19185745 7.         4.78       6.
 9.         8.         6.         6.         8.         7.17
 7.74       8.         6.02       7.         6.         6.
 6.75882251 9.         7.         6.02       9.         9.
 4.         6.09       7.         7.         7.         8.
 7.         6.         8.         7.         8.         8.
 7.         6.67316667 6.         6.19185745 7.         6.
 6.         6.         6.         6.         8.         7.
 7.         9.         9.        ]
```

Рис. 14 – вывод прогноза

Наконец, скрипт оценивает производительность модели путем вычисления среднеквадратичной ошибки и показателей R-квадрата и выводит их на печать.

```
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

Рис. 15 – скрипт оценивания среднеквадратичной ошибки и R-квадрата

```
Mean Squared Error: 0.026639870825691962
R-squared: 0.9792624863832664
```

Рис. 16 – вывод среднеквадратичной ошибки и R-квадрата

ВЫВОДЫ

Вывод по корреляции:

Присутствует корреляция между продолжительностью сна и качеством сна, хотя она может быть не очень сильной. Это говорит о том, что более продолжительный сон может быть связан с лучшим качеством сна.

Уровень физической активности показывает некоторую корреляцию с качеством и продолжительностью сна. Это означает, что более высокая физическая активность может быть связана с лучшим качеством сна или большей продолжительностью сна.

Уровень стресса, по-видимому, коррелирует с качеством и продолжительностью сна. Более высокий уровень стресса может негативно сказаться на качестве и продолжительности сна.

Корреляция между частотой сердечных сокращений и переменными, связанными со сном, присутствует, но не очень сильная. Это указывает на то, что, хотя может существовать некоторая взаимосвязь между частотой сердечных сокращений и режимом сна, на нее влияют множество других факторов.

Вывод по Кластерному анализу:

График показывает изгиб или точку перегиба вокруг 4 кластеров. Это говорит о том, что увеличение числа кластеров сверх 4 не обеспечивает существенного уменьшения суммы квадратов внутри кластера (WCSS).

WCSS является мерой дисперсии внутри каждого кластера; меньшее значение указывает на то, что точки данных в кластере расположены ближе к центру тяжести, что означает лучшую кластеризацию.

Таким образом, график метода Elbow показывает, что сегментирование набора данных о здоровье сна и образе жизни на 4 кластера является оптимальным для отражения внутренней структуры данных.

Вывод по регрессионному анализу:

Регрессионный анализ был проведен для прогнозирования качества сна на основе продолжительности сна, уровня физической активности, уровня стресса и частоты сердечных сокращений.

Значение R-квадрата равно 0,897, что указывает на то, что приблизительно 89,7% вариабельности качества сна объясняется моделью. Это говорит о сильном соответствии модели.

Коэффициенты:

Коэффициент равен 0,605, что означает, что за каждый дополнительный час сна качество сна повышается на 0,605 балла, при условии, что другие переменные постоянны.

Увеличение уровня физической активности на каждую единицу связано с повышением качества сна на 0,0065 балла при прочих равных условиях.

Уровень стресса имеет отрицательный коэффициент -0,3126, указывающий на то, что более высокие уровни стресса связаны со снижением качества сна.

Коэффициент частоты сердечных сокращений равен -0,0455 предполагает, что увеличение частоты сердечных сокращений связано со снижением качества сна.

Все переменные являются статистически значимыми ($p < 0,05$), что указывает на надежную взаимосвязь между этими факторами и качеством сна.

В заключение, регрессионная модель показывает, что продолжительность сна, уровень физической активности и уровень стресса являются значимыми предикторами качества сна, причем продолжительность сна и физическая активность оказывают положительное влияние, а стресс - отрицательное. Частота сердечных сокращений также отрицательно влияет на качество сна. Однако следует проявлять осторожность из-за потенциальных проблем с мультиколлинеарностью.

Модель прогнозирования:

Среднеквадратичная ошибка модели на тестовом наборе составляет приблизительно 0,0266. Это низкое значение указывает на то, что предсказания модели близки к фактическим значениям.

Оценка R2 составляет около 0,9793, что означает, что примерно 97,93% вариабельности качества сна объясняется моделью. Это говорит об очень хорошем соответствии модели.

ЗАКЛЮЧЕНЕ

Для написания курсовой работы была изучена специальная литература – учебные пособия и статьи.

Цель работы состояла в изучение и анализ статистических данных о качестве сна и образе жизни людей, выявление взаимосвязей и зависимостей между качеством сна и различными аспектами образа жизни. В ходе работы были рассмотрены различные методы реализации статистического анализа, были выявлены взаимосвязи.

Практическая часть работы включает реализацию статистических методов анализа данных, а именно реализацию корреляционного анализа, регрессионного анализа, кластеризацию и метод машинного обучения. В конце работы были сделаны выводы по каждому методы анализа данных.

Работа продемонстрировала понимание статистического анализа, анализ и интерпретацию данных. Таким образом цель была достигнута.

ЛИТЕРАТУРА

1. Шашков В.Б. Прикладной регрессионный анализ: многофакторная регрессия. Оренбург, (2003).

2. Шихалёв А.М. Регрессионный анализ. парная линейная регрессия: учебно-методическое пособие. Казань, (2015).

3. Харченко М.А. Корреляционный анализ: учебное пособие для вузов. Издательско-полиграфический центр Воронежского государственного университета, (2008)

4. Нигей Н.В. Корреляционный и регрессионный анализ: методическое пособие для самоподготовки. Благовещенск.

5. Бантикова О. И., Седова Е.Н., Чудинова О.С. Методы кластерного анализа: Оренбург ИПК ГОУ ОГУ, (2011).

ПРИЛОЖЕНИЕ

Рис. 17 QR код на репозиторий с исходным кодом проекта и результатами опроса.