Министерство науки и высшего образования Российской Федерации
ФГБОУ ВО «Волгоградский государственный технический университет»
Факультет электроники и вычислительной техники
Кафедра «Системы автоматизированного проектирования
и поискового конструирования»

Дисциплина Машинное обучение и анализ данных

Утверждаю
Зав. кафедрой САПРиПК
«_____»_____20 _г.

ЗАДАНИЕ

на курсовую работу (проект)

Студент Мухин Дмитрий Андреевич
(фамилия, имя, отчество)

Группа ИВТ-363

1. Тема: Прогнозирование максимального веса в жиме лёжа

Утверждена приказом от «_____» _____20___г. № _____

2. Срок представления работы (проекта) к защите «___»_____20___г.

3. Содержание расчетно-пояснительной записки: _Описание предметной области. Предварительная обработка данных. Стандартизация данных. Понижение размерности данных. Классификация данных. Тестирование модели.

4. Перечень графического материала: __Visual Studio Code

5. Дата выдачи задания «_____» _____20 ___г.

Руководитель работы (проекта)_____А. В. Хайров___
подпись, дата инициалы и фамилия

Задание принял к исполнению_____ _____12.02.2024___Мухин Д.А._____
подпись, дата инициалы и фамилия

Минобрнауки России
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Волгоградский государственный технический университет»

Факультет ___ электроники и вычислительной техники_____
Кафедра _____САПРиПК_____

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к курсовой работе (проекту)

по дисциплине _____Машинное обучение и анализ данных_____

на тему___Прогнозирование максимального веса в жиме лёжа _____

Студент_____Мухин Дмитрий Андреевич _____
(фамилия, имя, отчество)
Группа_____ИВТ-363_____

Руководитель работы (проекта) _____ _____
(подпись и дата подписания) (инициалы и фамилия)

Члены комиссии:

(подпись и дата подписания) (инициалы и фамилия)

(подпись и дата подписания) (инициалы и фамилия)

(подпись и дата подписания) (инициалы и фамилия)

Нормоконтролер _____ _____А.В. Хайров_____
(подпись, дата подписания) (инициалы и фамилия)

Волгоград 2024 г.

Содержание

Глава 1. Введение

1.1 Описание предметной области

Анализируемым объектом в курсовой работе является набор данных, содержащий информацию о спортивных результатах. Набор данных включает следующие сведения:

1. «playerId»: Уникальный id спортсмена;
2. «Name»: Имя спортсмена;
3. «Sex»: Пол спортсмена;
4. «Equipment»: Экипировка спортсмена;
5. «Age»: Возраст спортсмена;
6. «BodyWeightKg»: Вес спортсмена в килограммах;
7. «BestSquatKg»: Лучший результат в приседаниях;
8. «BestDeadliftKg»: Лучший результат в становой тяге.
9. «BestBenchKg»: Лучший результат в жиме лёжа.

1.2 Постановка задачи

На основе выбранного набора данных выполним следующие поставленные задачи:

1. Разведочный анализ данных для определения связей между атрибутами, вбросов и аномалий, распределения частот категориальных переменных;
2. Классификация данных и тренировка модели;
3. Тестирование разработанной модели.

Глава 2. Этапы выполнения работы

2.1. Предварительная обработка данных

Выведем следующую информацию о датасете: первые пять строк, и основная инфомация о столбцах (см рис. 1 и рис. 2).

	playerId	Name	Sex	Equipment	Age	BodyweightKg	BestSquatKg	BestDeadliftKg	BestBenchKg
0	19391.0	Carlos Ceron	M	Raw	23.0	87.30	205.0	235.0	125.0
1	15978.0	Tito Herrera	M	Wraps	23.0	73.48	220.0	260.0	157.5
2	27209.0	Levi Lehman	M	Raw	26.0	112.40	142.5	220.0	145.0
3	27496.0	Stacy Hayford	F	Raw	35.0	59.42	95.0	102.5	60.0
4	20293.0	Brittany Hirt	F	Raw	26.5	61.40	105.0	127.5	60.0

Рисунок 1 – Первые 5 строк DataFrame

```
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   playerId        18904 non-null  float64
 1   Name            18904 non-null  object
 2   Sex             18904 non-null  object
 3   Equipment       18904 non-null  object
 4   Age             18729 non-null  float64
 5   BodyweightKg    18904 non-null  float64
 6   BestSquatKg     18904 non-null  float64
 7   BestDeadliftKg  18904 non-null  float64
 8   BestBenchKg     18904 non-null  float64
```

Рисунок 2 – Тип данных столбцов

Выведем количество дубликатов, удалим их и посмотрим, сколько дубликатов осталось(см. рис. 3). Поиск производится по колонке Name и PlayerId.

```
Кол-во дубликатов в датасете:  3
Кол-во дубликатов в датасете:  0
```

Рисунок 3 - работа с дубликатами

Уберем ненужные признаки (колонки “playerId”, “Name”) (см. рис. 4). Преобразуем категориальные данные с помощью LabelEncoder. Колонку “Пол” приводим к виду: женщина - 0, мужчина - 1. Колонку “Экипировка”: однослойная экипировка - 0, нет экипировки - 1, многослойная экипировка - 2, кистевые бинты - 3. удалим дубликаты (см. рис. 5-6).

```
Названия колонок датасета:  ['Sex', 'Equipment', 'Age', 'BodyweightKg', 'BestSquatKg', 'BestDeadliftKg', 'BestBenchKg']
```

Рисунок 4 - Названия колонок после удаления ненужных признаков

	Sex	Equipment	Age	BodyweightKg	BestSquatKg	BestDeadliftKg	BestBenchKg
0	1	1	23.0	87.30	205.0	235.0	125.0
1	1	3	23.0	73.48	220.0	260.0	157.5
2	1	1	26.0	112.40	142.5	220.0	145.0
3	0	1	35.0	59.42	95.0	102.5	60.0
4	0	1	26.5	61.40	105.0	127.5	60.0

Рисунок 5 - 5 строк DataFrame после преобразования категориальных переменных

```
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Sex             18901 non-null  int32  
 1   Equipment       18901 non-null  int32  
 2   Age             18726 non-null  float64
 3   BodyweightKg    18901 non-null  float64
 4   BestSquatKg     18901 non-null  float64
 5   BestDeadliftKg  18901 non-null  float64
 6   BestBenchKg     18901 non-null  float64
```

Рисунок 6 - типы данных столбцов DataFrame после преобразования

2.2. Стандартизация данных

Стандартизируем данные, применив метод MinMaxScaler (см. рис.7)

	Sex	Equipment	Age	...	BestSquatKg	BestDeadliftKg	BestBenchKg
0	1.0	0.50	0.210526	...	0.644578	0.555969	0.278673
1	1.0	1.00	0.210526	...	0.662651	0.620050	0.356817
2	1.0	0.50	0.250000	...	0.569277	0.517520	0.326761
3	0.0	0.50	0.368421	...	0.512048	0.216338	0.122385
4	0.0	0.50	0.256579	...	0.524096	0.280419	0.122385
5	1.0	1.00	0.210526	...	0.710843	0.684131	0.362828
6	1.0	1.00	0.210526	...	0.710843	0.684131	0.398894
7	1.0	0.50	0.203947	...	0.593373	0.466255	0.236595
8	1.0	0.50	0.223684	...	0.638554	0.536744	0.410916
9	0.0	0.75	0.164474	...	0.536145	0.280419	0.128396

Рисунок 7 – DataFrame со стандартизированными данных

Создадим новый датасет для обучения, убрав колонку BestBenchKg (см. рис. 8)

	Sex	Equipment	Age	BodyweightKg	BestSquatKg	BestDeadliftKg
0	1.0	0.50	0.210526	0.349803	0.644578	0.555969
1	1.0	1.00	0.210526	0.270773	0.662651	0.620050
2	1.0	0.50	0.250000	0.493338	0.569277	0.517520
3	0.0	0.50	0.368421	0.190370	0.512048	0.216338
4	0.0	0.50	0.256579	0.201693	0.524096	0.280419
5	1.0	1.00	0.210526	0.499628	0.710843	0.684131
6	1.0	1.00	0.210526	0.499628	0.710843	0.684131
7	1.0	0.50	0.203947	0.368388	0.593373	0.466255
8	1.0	0.50	0.223684	0.422428	0.638554	0.536744
9	0.0	0.75	0.164474	0.143364	0.536145	0.280419

Рисунок 8 – DataFrame после преобразования

2.3. Разведочный анализ данных

Выведем «ящик с усами» для колонок Equipment и Age после стандартизации данных(см. рис. 9-10)

«Ящик с усами»— это график, использующийся в описательной статистике, который компактно изображает одномерное распределение вероятностей.

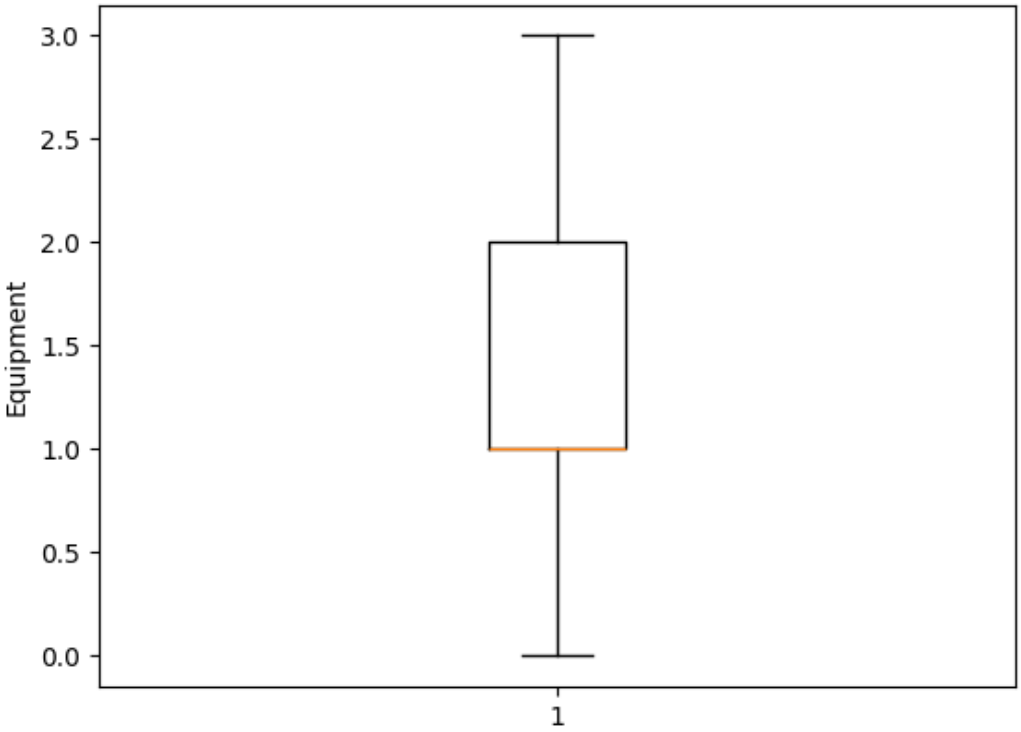


Рисунок 9 - график столбца Equipment

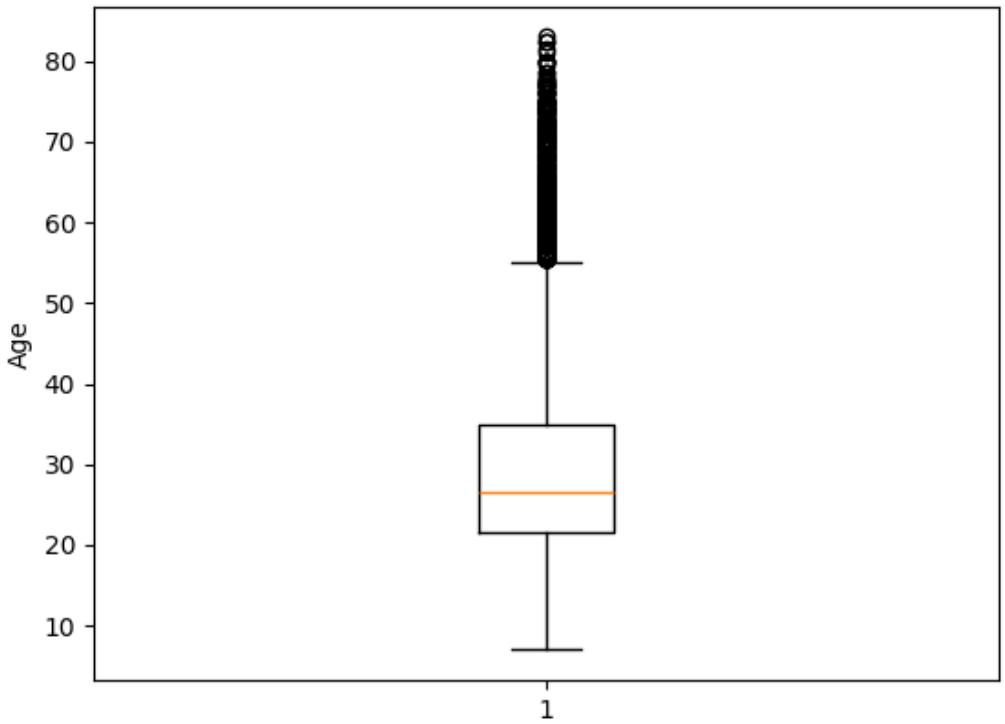


Рисунок 10 - график столбца Age

Выведем тепловую карту для нашего датасета. С помощью него определим

силы и направление корреляции между парами объектов в датасете, что поможет определить наиболее важные функции для модели, а также повысить точность модели (см. рис. 11). После чего выведем график, на котором будет видно, насколько часто используют тот или иной вид экипировки (см. рис. 12)

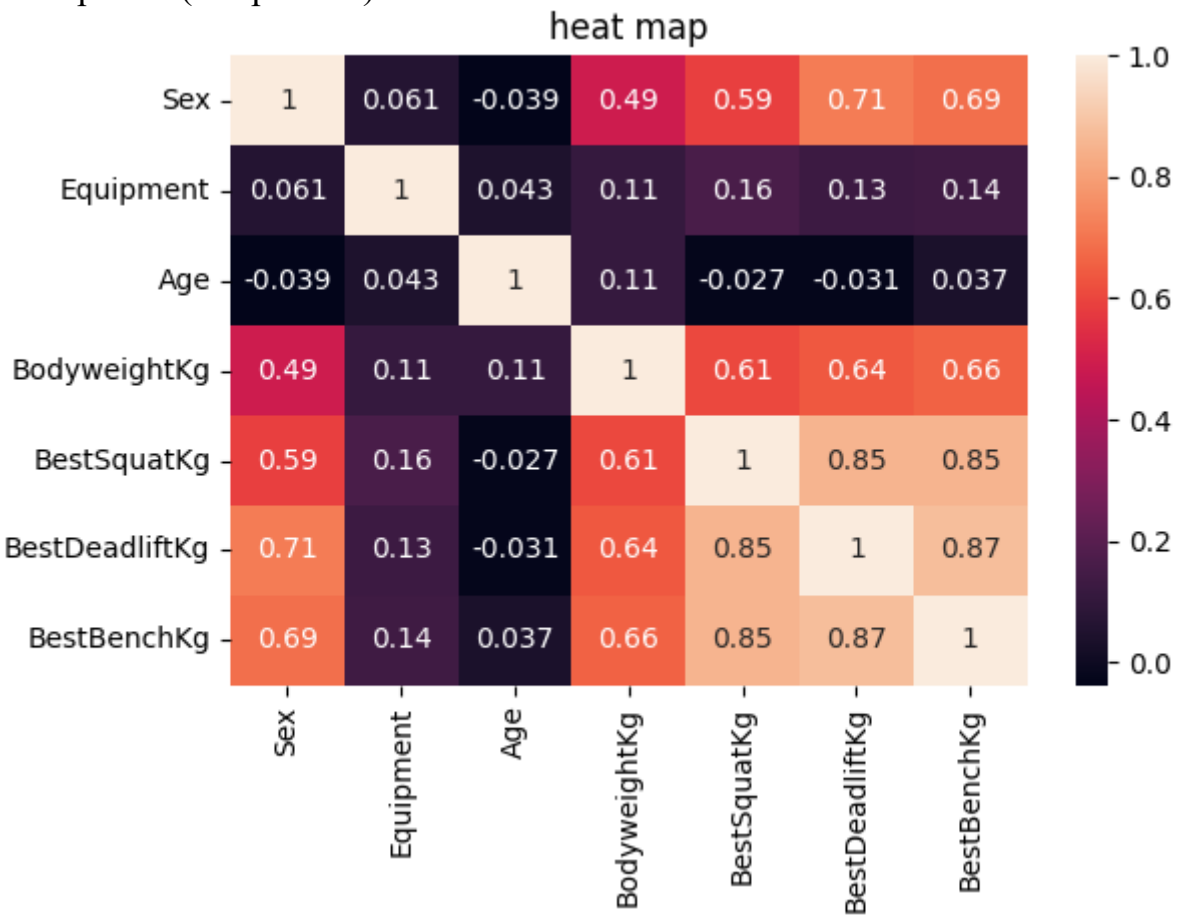


Рисунок 11 - Тепловая карта

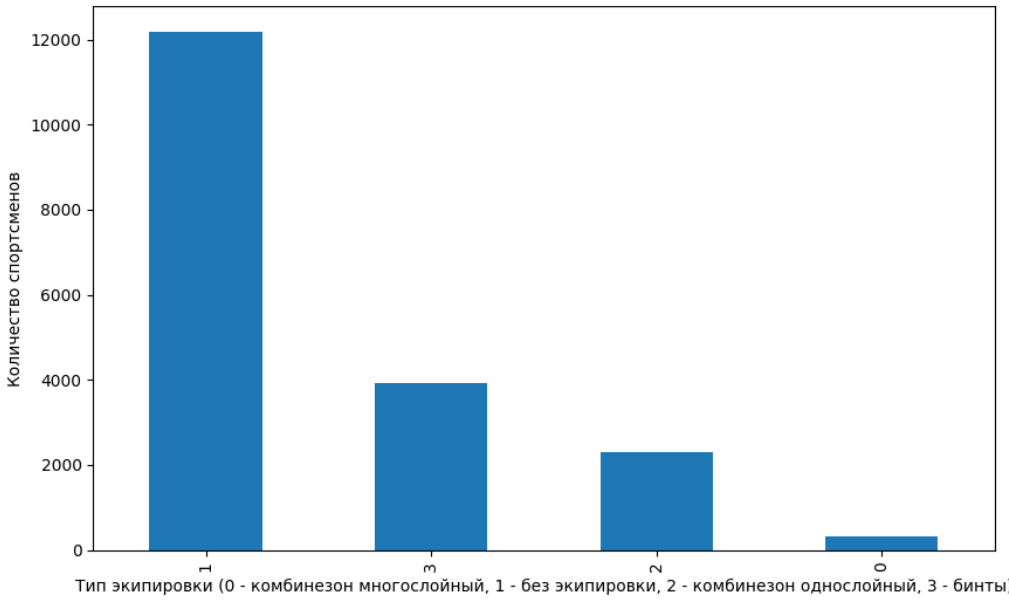


Рисунок 12 - Распределение частот используемых видов экипировки

2.4. Классификация данных

Проведём анализ наивным Байесовским методом и методом k-ближних

соседей. (см. рис. 13-14)

Наивный байесовский метод (Naive Bayes) - это простой вероятностный классификатор, основанный на применении теоремы Байеса с "наивным" предположением о независимости между признаками. Этот метод широко используется в задачах классификации текстов и анализе данных, особенно когда имеется большое количество признаков.

Метод k-ближайших соседей (k-Nearest Neighbors, k-NN) - это простой и интуитивно понятный алгоритм классификации и регрессии. Он основывается на принципе "похоже на похожее": объекты классифицируются на основе классов их k ближайших соседей в пространстве признаков. В качестве метрики сходства часто используется евклидово расстояние.

```
Метод k-ближних соседей
Верных наблюдений:  804
Неверных наблюдений 133
Наивный байесовский метод
Верных наблюдений:  807
Неверных наблюдений 130
```

Рисунок 13 - Количество верных и неверных наблюдений для методов

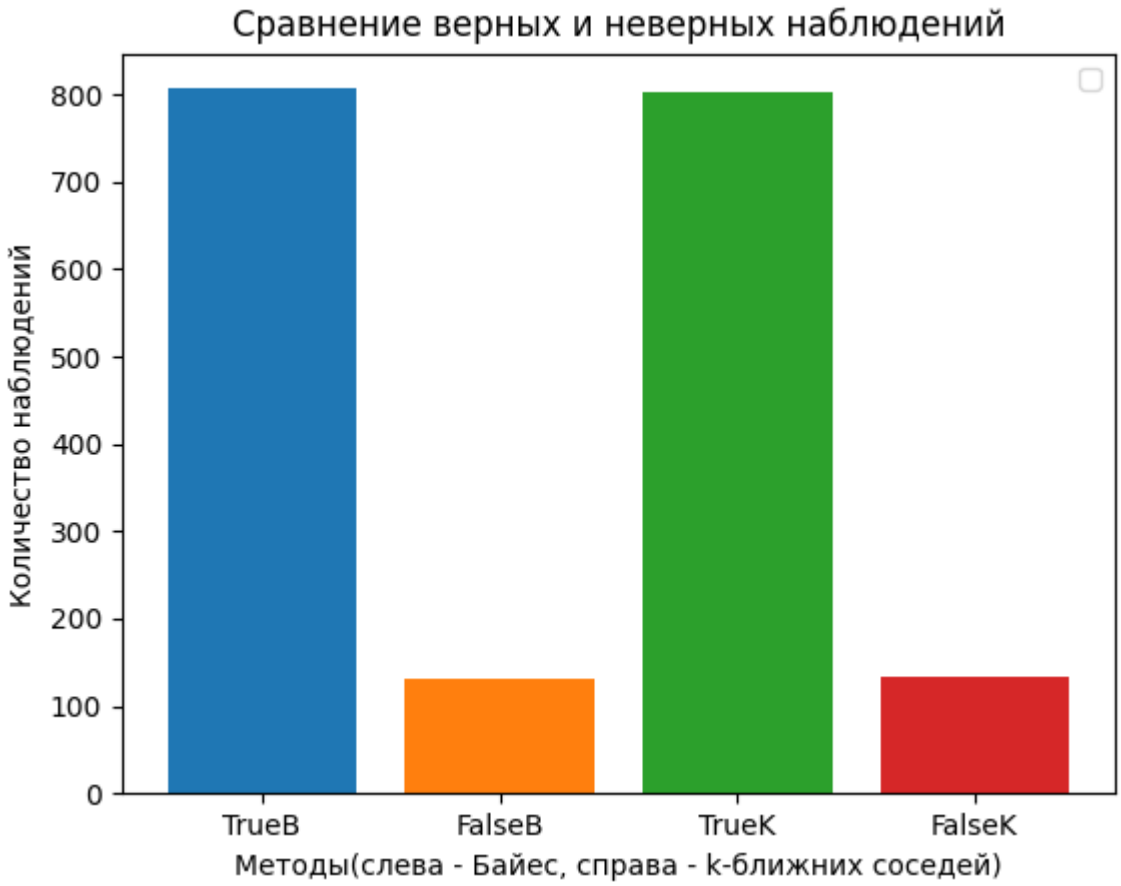
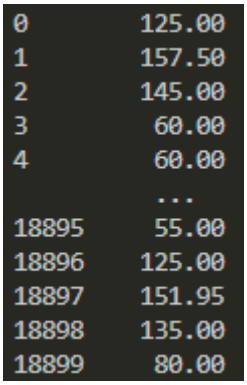


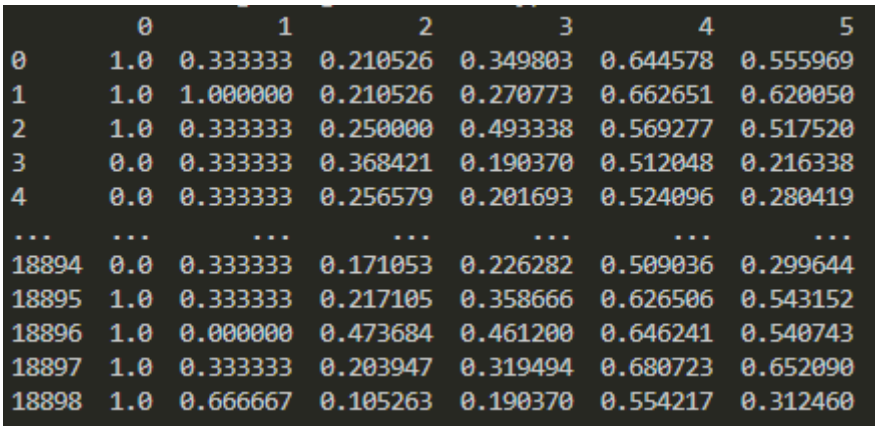
Рисунок 14 - График неверных наблюдений

Выделяем данные и метки для обучения модели. Столбец “BestBenchKg” берём как ответ, остальные как данные для получения ответов (см. Рис.15-16)



0	125.00
1	157.50
2	145.00
3	60.00
4	60.00
	...
18895	55.00
18896	125.00
18897	151.95
18898	135.00
18899	80.00

Рисунок 15 - данные для ответов



	0	1	2	3	4	5
0	1.0	0.333333	0.210526	0.349803	0.644578	0.555969
1	1.0	1.000000	0.210526	0.270773	0.662651	0.620050
2	1.0	0.333333	0.250000	0.493338	0.569277	0.517520
3	0.0	0.333333	0.368421	0.190370	0.512048	0.216338
4	0.0	0.333333	0.256579	0.201693	0.524096	0.280419
...	...	...	...	...	...	...
18894	0.0	0.333333	0.171053	0.226282	0.509036	0.299644
18895	1.0	0.333333	0.217105	0.358666	0.626506	0.543152
18896	1.0	0.000000	0.473684	0.461200	0.646241	0.540743
18897	1.0	0.333333	0.203947	0.319494	0.680723	0.652090
18898	1.0	0.666667	0.105263	0.190370	0.554217	0.312460

Рисунок 16 - данные для получения ответов

Далее обучаем модель, используя пропорции выборки 95/5. Параметры RandomForestRegressor(): n_estimators = 100, random_state = 50

Random Forest Regressor - это модель машинного обучения, которая используется для решения задач регрессии. Она основана на алгоритме Random Forest, который в свою очередь является ансамблевым методом обучения. В этой модели ансамбль из множества деревьев решений обучается на разных подмножествах данных и выдает прогноз в виде среднего значения (для регрессии) по прогнозам каждого отдельного дерева.

Каждое дерево в лесу решений обучается на случайной подвыборке данных, а также в каждом узле дерева случайно выбирается подмножество признаков для разделения. Это позволяет уменьшить переобучение и повысить обобщающую способность модели.

Когда требуется сделать прогноз для новых данных, каждое дерево в лесу выдает свой собственный прогноз, а затем эти прогнозы усредняются (в случае регрессии), чтобы получить окончательный результат.

2.5. Тестирование модели

Реализуем механизм ввода значений параметров спортсмена для входных данных разработанной модели, необходимо ввести пол (int, 0 - женщина, 1 -

мужчина), тип экипировки (int, 0 - однослойная экипировка, 1 - нет, 2 - многослойная экипировка, 3 - кистевые бинты), возраст, вес тела, макс. результат в приседаниях и становой тяге,

Были введены следующие данные

1) Пол (женщина - 0, мужчина - 1): 1
   Экипировка (0 - комбинезон многослойный, 1 - без экипировки, 2 - комбинезон однослойный , 3 - бинты): 1
   Возраст: 23
   Вес (кг): 87.3 Вес на снаряде для приседаний (кг): 205
   Вес на снаряде для становой тяги (кг): 235(см. рис. 17)
2) Пол (женщина - 0, мужчина - 1): 1
   Экипировку (0 - комбинезон многослойный, 1 - без экипировки, 2 - комбинезон однослойный , 3 - бинты): 3
   Возраст: 23
   Вес (кг): 73.48
   Вес на снаряде для приседаний (кг): 220
   Вес на снаряде для становой тяги (кг): 260(см. рис. 18)
3) Пол (женщина - 0, мужчина - 1): 0
   Экипировку (0 - комбинезон многослойный, 1 - без экипировки, 2 - комбинезон однослойный , 3 - бинты): 2
   Возраст: 19
   Вес (кг): 51
   Вес на снаряде для приседаний (кг): 115
   Вес на снаряде для становой тяги (кг): 127.5(см. рис. 19)
4) Пол (женщина - 0, мужчина - 1): 1
   Экипировку (0 - комбинезон многослойный, 1 - без экипировки, 2 - комбинезон однослойный , 3 - бинты): 0
   Возраст: 28
   Вес (кг): 75
   Вес на снаряде для приседаний (кг): 322
   Вес на снаряде для становой тяги (кг): 250(см. рис. 20)

```
Параметры спортсмена:
Пол (женщина - 0, мужчина - 1): 1
Экипировку (0 - комбинезон многослойный, 1 - без экипировки, 2 - комбинезон однослойный , 3 - бинты): 1
Возраст: 23
Вес (кг): 87.3
Вес на снаряде для приседаний (кг): 205
Вес на снаряде для становой тяги (кг): 235
Прогнозируемый вес снаряда: ~ 130
Ожидаемый результат: ~125
```

Рисунок 17 - Спортсмен-мужчина без экипировки

```
Параметры спортсмена:
Пол (женщина - 0, мужчина - 1): 1
Экипировку (0 - комбинезон многослойный, 1 - без экипировки, 2 - комбинезон однослойный , 3 - бинты): 3
Возраст: 23
Вес (кг): 73.48
Вес на снаряде для приседаний (кг): 220
Вес на снаряде для становой тяги (кг): 260
Прогнозируемый вес снаряда: ~ 151
Ожидаемый результат: ~157.5
```

Рисунок 18 - Спортсмен-мужчина с экипировкой (бинты)

```
Параметры спортсмена:
Пол (женщина - 0, мужчина - 1): 0
Экипировку (0 - комбинезон многослойный, 1 - без экипировки, 2 - комбинезон однослойный , 3 - бинты): 2
Возраст: 19
Вес (кг): 51
Вес на снаряде для приседаний (кг): 115
Вес на снаряде для становой тяги (кг): 127.5
Прогнозируемый вес снаряда: ~ 61
Ожидаемый результат: ~62.5
```

Рисунок 19 - Спортсмен-женщина с однослойным комбинезоном

```
Параметры спортсмена:
Пол (женщина - 0, мужчина - 1): 1
Экипировку (0 - комбинезон многослойный, 1 - без экипировки, 2 - комбинезон однослойный , 3 - бинты): 0
Возраст: 28
Вес (кг): 75
Вес на снаряде для приседаний (кг): 322
Вес на снаряде для становой тяги (кг): 250
Прогнозируемый вес снаряда: ~ 174
Ожидаемый результат: ~185
```

Рисунок 20 - Спортсмен-мужчина с многослойным комбинезоном

Тестирование модели показало примерно ожидаемые результаты. Результаты могут немного отличаться, т.к. наши данные не могут учитывать некоторые индивидуальные параметры тела спортсмена. Так, влияние таких параметров, как максимальные веса в приседаниях и становой тяге, а также пол и вес тела оказывают наибольшее влияние на вес снаряда в жиме лёжа.

Заключение

В результате выполнения курсовой работы был проанализирован и обработан набор данных по теме «Прогнозирование максимального веса в жиме лёжа», а также реализована и обучена на пред обработанных данных классификационная модель, определяющая максимальный вес в жиме лёжа с помощью входных параметров спортсмена.

Список источников

1. Методические указания к лабораторным работам №1-8 по дисциплине «Машинное обучение и анализ данных», ВолгГТУ, 2024 г.
2. Kaggle - https://www.kaggle.com/datasets/kukuroo3/powerlifting-benchpress-weight-predict
3. Визуализация данных в python - https://habr.com/ru/articles/502958/
4. Документация библиотеки scikit в python - https://scikit-learn.ru/1-9-naive-bayes/?ysclid=lvtm3vjbdu138569632