Санкт-Петербургский политехнический университет
Высшая школа прикладной математики и вычислительной физики,
ФизМех

Направление подготовки
“01.03.02_01 Математическое моделирование и искусственный интеллект”

Тема: “Прогнозирование расхода калорий”

Дисциплина: “Основы машинного обучения"

Выполнил студент: Чуев В. Ю. (гр. 5030102/00101)

Преподаватель: Кацман В. И.

Санкт-Петербург
2024

Оглавление

1. Постановка задачи

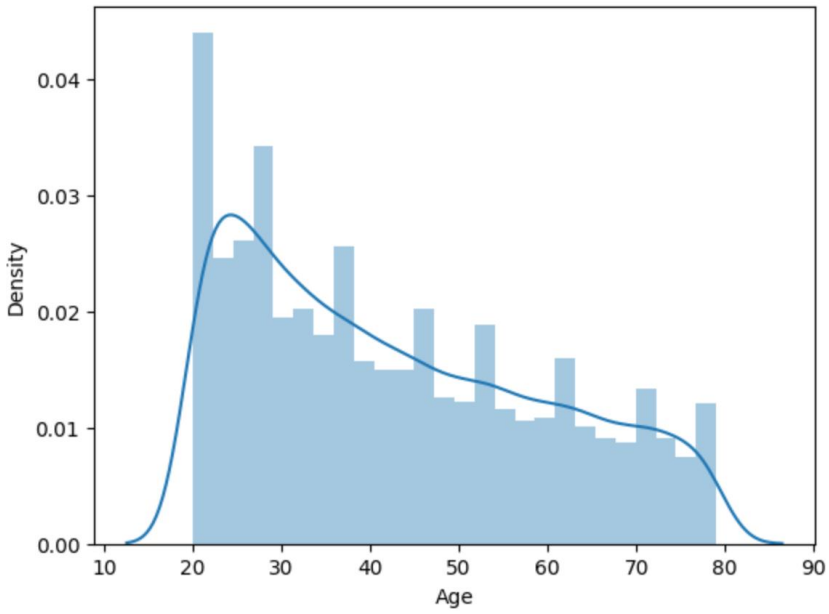
Обучить модель, которая будет получать на вход параметры человека(пол, возраст, рост, вес, пульс, температура) и время тренировки и выдавать ожидаемый расход калорий.

2. Описание датасета

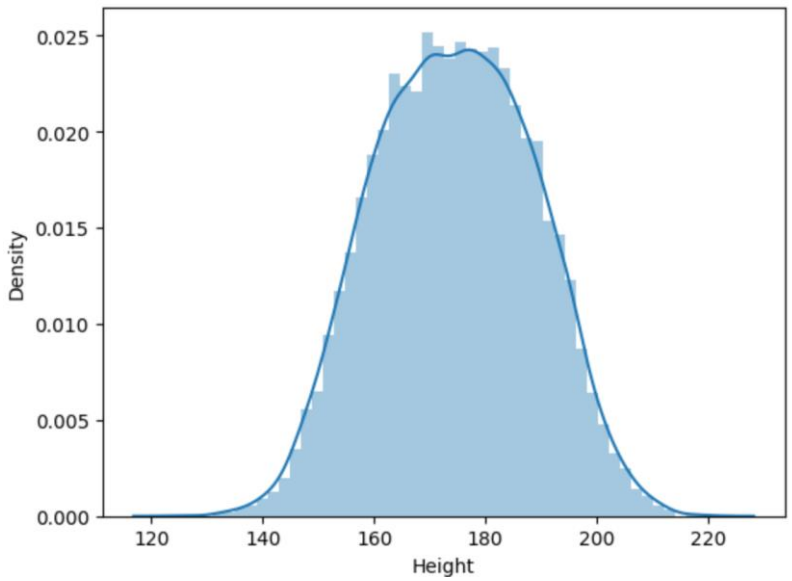
Датасет состоит из 15000 строк следующего вида:

	User_ID	Gender	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
0	14733363	male	68	190.0	94.0	29.0	105.0	40.8	231.0
1	14861698	female	20	166.0	60.0	14.0	94.0	40.3	66.0
2	11179863	male	69	179.0	79.0	5.0	88.0	38.7	26.0
3	16180408	female	34	179.0	71.0	13.0	100.0	40.5	71.0
4	17771927	female	27	154.0	58.0	10.0	81.0	39.8	35.0

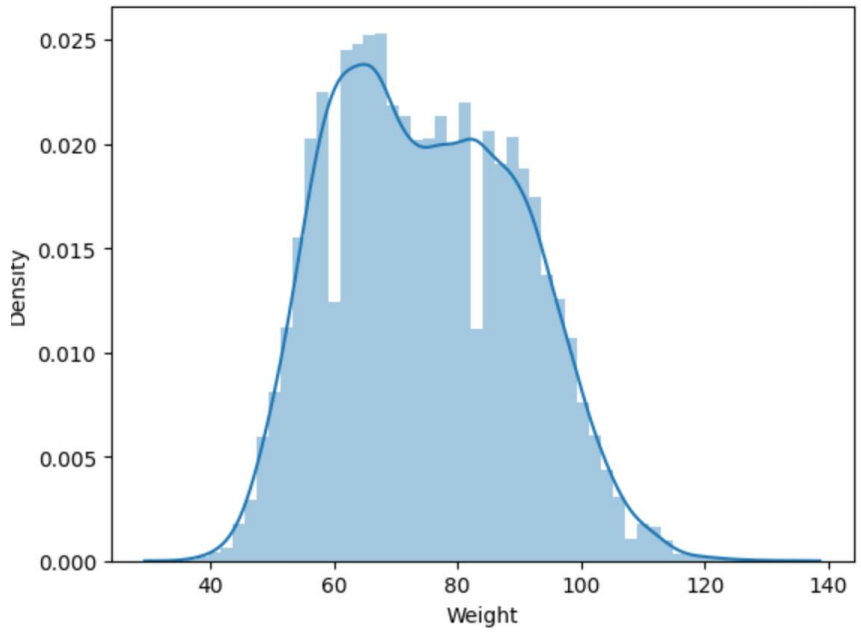
Распределение по возрасту:



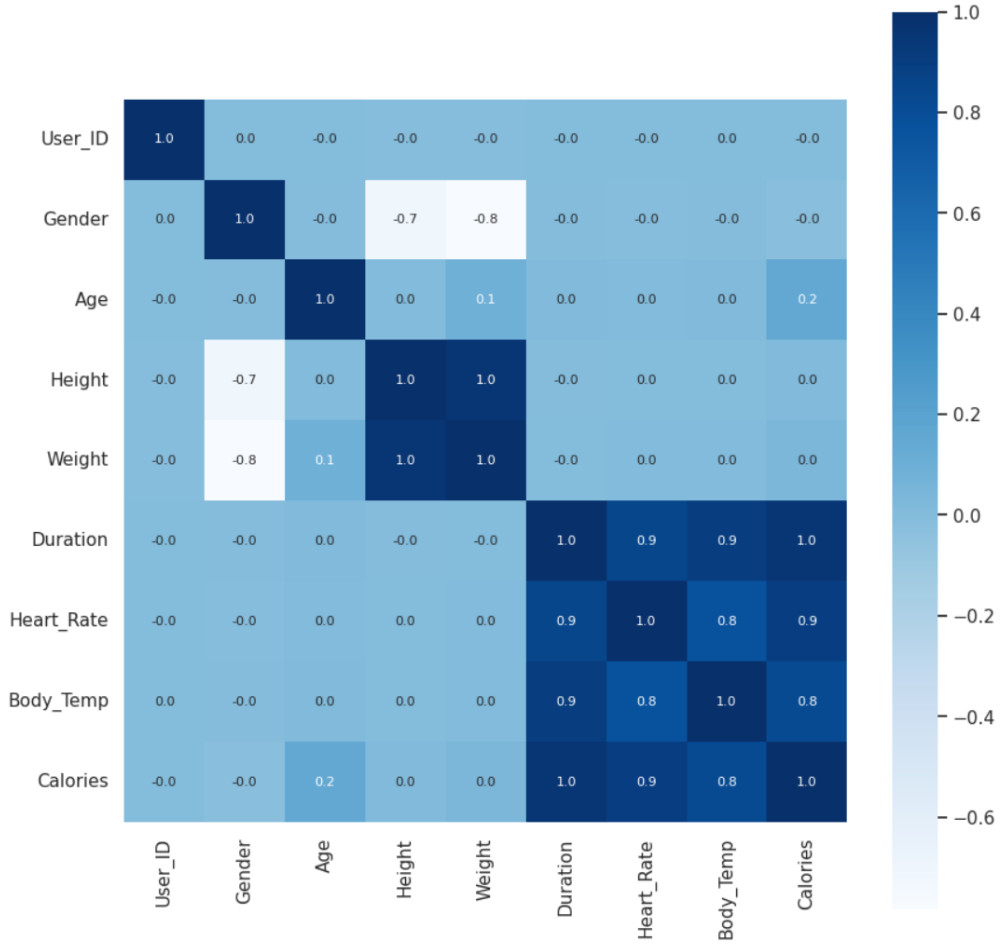
Распределение по росту:



Распределение по весу:



Корреляция параметров:



3. Метод линейной регрессии

Датасет был разбит на обучающую выборку, состоящую из 12000 строк и тестовую выборку, состоящую из 3000 строк и построена модель линейной регрессии.

В качестве критерия качества предсказания количества калорий будем использовать среднее значение модуля разности предсказанного значения со значением из тестовой выборки(ккал), а также среднее значение модуля разности относительно значения из тестовой выборки (%)

На тестовой выборке модель показала следующий результат: средняя погрешность равна 8.39 ккал и 27.79 %. Наибольшая относительная погрешность возникает на данных, где время тренировки мало, а наибольшая абсолютная на данных, где время тренировки велико. Из этого можно сделать вывод, что такая большая погрешность возникает из-за нелинейной зависимости затраченных калорий от времени тренировки.

4. Метод опорных векторов

Построим модель методом опорных векторов, используя такое же разбиение и критерии качества. Полученные результаты: средняя погрешность равна 10.62 ккал и 26.76 %. Наибольшая погрешность возникает на тех же данных, что и в прошлом методе. Сделаем следующие итерации для улучшения этих результатов:

1) Масштабируем данные, чтобы среднее было равно 0, а дисперсия равна 1.

2) Подбираем коэффициент регуляризации C=200.

3) Подбираем коэффициент gamma=0.05 Для RBF ядра.

Изменение погрешности после проделанных итераций:

Погрешность SVR	Начальная	Масштабирование	Подбор C	Подбор gamma
MeanAbs(kcal)	10.62	2.36	0.30	0.28
MeanAbsPercent(%)	26.76	8.44	1.04	0.84

5. Выводы

Задача прогнозирование расхода калорий была решена с методов линейной регрессии и опорных векторов. Изначально оба метода показали невысокое качество предсказания с погрешностью 8.39 ккал/27.79% и 10.62 ккал/26.76% соответственно из-за нелинейной зависимости затраченных калорий от времени тренировки. После итераций по улучшению модели(масштабирование данных, подбор C, подбор gamma), погрешность метода опорных векторов равна 0.28 ккал/0.84% , что сопоставимо погрешности измерения расхода калорий и является хорошим результатом.