Лабораторная работа №3

Целью данной лабораторной работы является разработка программы, реализующей применение метода линейной регрессии к заданному набору данных.

Набор данных содержит в себе информацию о вариантах португальского вина "Винью Верде". Входные переменные представляют собой 13 столбцов со значениями, полученными на основе физико-химических тестов, а именно:
0 – цвет вина (“red” / ”white”)
1 - фиксированная кислотность
2 - летучая кислотность
3 - лимонная кислота
4 - остаточный сахар
5 - хлориды
6 - свободный диоксид серы
7 - общий диоксид серы
8 - плотность
9 - pH
10 - сульфаты
11 - спирт
Выходная переменная (на основе сенсорных данных):
12 - качество (оценка от 0 до 10, целое число)

Классы упорядочены и не сбалансированы (например, нормальных вин гораздо больше, чем отличных или плохих). В предоставленных данных есть пропуски и неточности. Задания выполняются согласно варианту. Чтобы определить номер варианта, воспользуйтесь следующей формулой:

$$N_{\text{варианта}} = (N_{\text{по списку}}\ mod\ 4) + 1$$

Варианты заданий:

1) Использовать классическую модель LinearRegression
2) Использовать модель LASSO
3) Использовать модель LARS
4) Использовать модель ElasticNet

Задание: Данные необходимо рассматривать как три набора. Данные для красного вина, данные для белого, общие данные вне зависимости от цвета. Необходимо построить модель для каждого из наборов, обучить её и сравнить полученные при помощи модели результаты с известными. Для обучения использовать 70% выборки, для тестирования 30%. Разбивать необходимо случайным образом, а, следовательно, для корректности тестирования

качества модели, эксперимент необходимо провести не менее 10 раз и вычислить среднее значение качества регрессии.

Особенности работы с данными:

1) Данные разнотипные, поэтому необходимо все столбцы привести к одному типу. Все данные должны быть вещественными числами. В данных есть пропуски, а это означает, что при считывании они будут записаны как NaN (либо произойдёт ошибка).
2) Результат работы модели будет тоже вещественным числом. Поэтому для оценки качества работы модели, необходимо использовать не прямое сравнение, а учитывать разницу между настоящим значением и смоделированным.
3) Данные в столбцах имеют разную размерность. Поэтому необходимо их нормализовать. Можно воспользоваться, например, методом preprocessing.normalize().

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \in [0,1]$$

В качестве результата выполненной лабораторной работы должна быть разработанная программа, решающая поставленную задачу и отчёт с содержанием текста программы, краткими комментариями и результатами работы программы.