***Лабораторная работа №1***

**ПОСТРОЕНИЕ И АНАЛИЗ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ**

**ВСЕ ЗНАЧЕНИЯ КОТОРЫЕ ВЫ ПОЛУЧАЕТЕ В КОНСОЛИ ДОЛЖЫ БЫТЬ ПЕРЕНЕСЕНЫ В ОТЧЁТ, А НЕ ВСТАВЛЕН СКРИН!!!!**

**Описание датасетов**

**1.**Набор данных ***«Рост-Вес-Возраст-Позиция»***. Размер исследуемого датасета – 115 элементов. Выборка сформирована из коллекции [***SOCR Data***](http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data) ***- 1035 Records of Heights and Weights of Major League Baseball Players***, в которой содержится 1035 записей о росте и весе игроков американской бейсбольной лиги. Для каждого элемента выборки приводятся значения четырех переменных: рост (дюймы), вес (фунты) и возраст (годы), местоположение на поле:

1) *Height*;

2) *Weight*;

3) *Age*.

4) *Position*: *Starting\_Pitcher* – «подающий» (бросает бейсбольный мяч в сторону кетчера) – **«1»**; *Catcher* – «принимающий» («ловящий») – **«2»**; First\_Baseman – «первый бейсмен» занимает место около первой базы – **«3»**, *Outfielder* – «аутфилдер», один из трёх игроков, занимающих оборонительную позицию во внешнем поле – **«4»**; *Relief\_Pitcher* – подающий, выходящий на временную замену – **«5»**.

**2.** Набор данных ***«Iris Data Set»* –**  ***«Ирисы Фишера».*** Выборка взята из коллекции ***UC Irvine Machine Learning Repository*** и состоит из 150 цветков [ирисов](https://ru.wikipedia.org/wiki/%D0%98%D1%80%D0%B8%D1%81_(%D1%80%D0%B0%D1%81%D1%82%D0%B5%D0%BD%D0%B8%D0%B5)), распределенных по трем классам. Каждый класс содержит одинаковое число элементов (50 экземпляров): [Ирис щетинистый](https://ru.wikipedia.org/wiki/%D0%98%D1%80%D0%B8%D1%81_%D1%89%D0%B5%D1%82%D0%B8%D0%BD%D0%B8%D1%81%D1%82%D1%8B%D0%B9) (*Iris setosa*), [Ирис виргинский](https://ru.wikipedia.org/w/index.php?title=%D0%98%D1%80%D0%B8%D1%81_%D0%B2%D0%B8%D1%80%D0%B3%D0%B8%D0%BD%D1%81%D0%BA%D0%B8%D0%B9&action=edit&redlink=1) (*Iris virginica*) и [Ирис разноцветный](https://ru.wikipedia.org/w/index.php?title=%D0%98%D1%80%D0%B8%D1%81_%D1%80%D0%B0%D0%B7%D0%BD%D0%BE%D1%86%D0%B2%D0%B5%D1%82%D0%BD%D1%8B%D0%B9&action=edit&redlink=1) (*Iris versicolor*).

Для каждого элемента выборки приводятся метка класса и значения 4-х признаков (в сантиметрах):

1) *Sepal Length* (длина чашелистика);

2) *Sepal Width* (ширина чашелистика);

3) *Petal Length* (длина лепестка);

4) *Petal Width* (ширина лепестка).

**Описание настроек**

1. Устанавливаем библиотекy:

pip install pingouin

Основные библиотеки в проекте (они будут установлены при установке pingoin)

* 1. scikit-learn (sklearn) –для построения моделей и много другого
  2. matplotlib – для построения графиков
  3. pandas – убоство работы с данными
  4. seaborn - для построения графиков
  5. scipy – тут есть всё для статистики и построения моделей.
  6. pingouin – собрана основные статистические методы

1. Датасет «***Ирисы Фишера***» есть в библиотеки scikit-learn (так же дополнительно переведём в формат библиотеки pandas)

import pandas as pd

from sklearn.datasets import load\_iris

iris = load\_iris()

iris\_pd=pd.DataFrame(data=np.c\_[iris['data'], iris['target']], columns=iris['feature\_names'] + ['target'])

**Задание на лабораторную работу**

1. Для каждого датасета провести разведочный анализ данных:
   * построить диаграмму Тьюки, оценить диапазон изменения данных (для этого используем matplotlib). В отчёт включить диаграмму и написать к ней выводы.
   * проанализировать корреляционные зависимости между исследуемыми переменными. Необходимо построить тепловую карту (для этого используем pandas + seaborn). В отчёт включить тепловую карту и написать к ней выводы.
   * рассчитать частные коэффициенты корреляции, сравнить их со значениями парных коэффициентов корреляции. Необходимо построить тепловую карту (для этого используем pandas + seaborn + pingouin (у датафрейма пандаса будет доступен метод pcorr)). В отчёт включить тепловую карту и написать к ней выводы.
   * проверить предположение о распределении признаков по нормальному закону критерием Колмогорова-Смирнова. Необходимо рассчитать значения статистик (для этого используем scipy). В отчёт включить таблицу значений и написать к ней выводы.
   * сделать выводы о степени однородности данных, силе зависимости между переменными, виде функции распределения, наиболее информативных переменных (Вставляем в конец абзаца).
2. Определить входные и выходные переменные. Построить парную регрессию и множественную регрессию (Используем sklearn). Сравнить результаты парной и множественной регрессии. В отчёт включить график с предсказанными значения (на одном графике), график предсказанных значений и остатков (на одном графике), и значения множественного коэффициента корреляции, коэффициента детерминации, скорректированного коэффициента детерминации и стандартной ошибки. Выбрать наиболее адекватную модель (в отчёте написать обоснование).
3. Проверить гипотезу о нормальном распределении остатков. Рассчитать статистику Дурбина-Уотсона. В отчёт включить рассчитанные значения и написать, как определили исходя из полученного значения статистики.
4. Проанализировать наличие выбросов (*смотрим по «*диаграмме Тьюки*»*). Сделать вывод о наличии (отсутствии) выбросов (с учетом результатов, полученных в предыдущих пунктах). В отчёт продублировать диаграмму и написать выводы.
5. Провести предсказание зависимой переменной по вычисленному уравнению множественной регрессии (*Predict Dependent Variable*), оценить точность. Для этого делим выборку на тестовую и обучающую. После чего рассчитываем СКО, коэффициент детерминации и строим график ошибок. В отчёт включить график, значение СКО, коэффициента детерминации и написать выводы о качестве модели.
6. Сформулировать предложения по улучшению регрессионной модели.
7. По указанию преподавателя провести дополнительные исследования для одного из датасета, исключив выбросы, обнаруженные в исходной выборке. Оценить, насколько изменяется регрессионная зависимость после удаления выбросов.
8. Провести корреляционный анализ и сделать выводы о целесообразности построения регрессии для датасетов:

- *Barotrop.sta*;

- *Chemical\_process.sta*;

- *Factor.sta*.