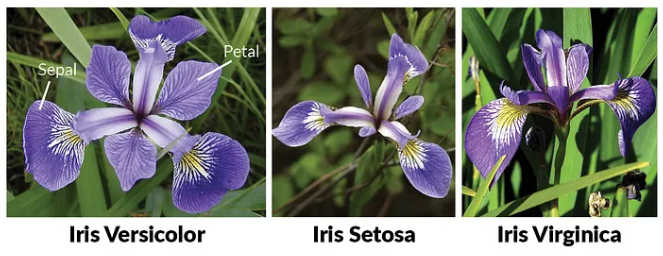
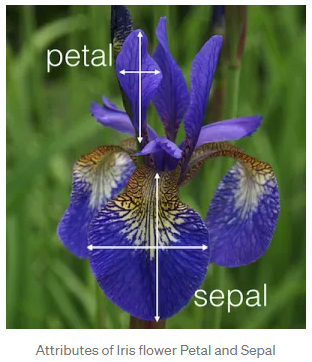
Лабораторная работа #3 часть 1

**Применение логистической регрессии для решения задачи бинарной классификации**

Лабораторная работа будет выполнятся на учебном наборе данных Iris который поставляется в составе Scikit-Learn. Несколько слов о датасете. Датасет описывает 3 сорта цветков ириса: setosa, versicolor, и virginica путем измерения их лепестков. См. картинки ниже:



В датасете содержатся 4 параметра - 'sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)' и собственно целевая переменная, определяющая сорт. По каждому сорту содержится 50 записей, всего – 150 строк с данными (<https://scikit-learn.org/stable/datasets/toy_dataset.html#iris-dataset>

)

Загрузка датасета:

from sklearn.datasets

import load\_irisiris = load\_iris()

Далее будет удобнее перейти к обьекту датафрейм Pandas, например так

import pandas as pd

df = pd.DataFrame(iris.data, columns = iris.feature\_names)

df['target'] = iris.target

Выведем пример из датасета. Обратите внимание, что в получившимся датафрейме имена сортов уже закодированы числами. Мы можем посмотреть имена сортов используя команду

print(iris.target\_names)

А свойство iris.target которое мы использовали при создании датафрейма, как легко убедится содержит номера, где 0 = setosa, 1 = versicolor, 2 = virginica :

print(iris.target)

**Задания к работе:**

Немного поисследуем датасет:

1. Используя Matplotlib отрисовать в цвете для всех 3 сортов зависимости: 'sepal length - sepal width’ и 'petal length - petal width'

Вы заметите что сорт setosa заметно отделен от двух других.

1. Использовать библиотеку seaborn и метод pairplot вывести результат либо для всего датасета либо для обучающей выборки. Запомнить на будущее 😊
2. Подготовим из имеющегося набора данных 2 датасета: в первом оставить setosa и versicolor, во втором – versicolor и virginica

Переходим к машинному обучению:

1. Каждый датасет разбить на обучающую и тестовые выборки (понадобится для следующих частей лабораторной работы с метриками классификации)
2. Использовать для обучения модель:

from sklearn.linear\_model import LogisticRegression

с параметрами по умолчанию, например:

clf = LogisticRegression(random\_state=0)

1. Обучить модель (fit)
2. Сделать предсказание (predict)
3. Вывести значение точности модели (score)

Разделы 4 – 8 проделать для 2 датасетов полученных в п. 3.

1. Давайте теперь сгенерируем датасет случайным образом и проведем для его бинарную классификацию.

Сгенерируем выборку для классификации самостоятельно, используя make\_classification из библиотеки scikit-learn.

X, y = make\_classification(n\_samples=1000, n\_features=2, n\_redundant=0, n\_informative=2,random\_state=1, n\_clusters\_per\_class=1)

Отрисовать полученный датасет используя Matplotlib и провести бинарную классификацию по пунктам 5 - 8