МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение

высшего образования

«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра «Измерительно-вычислительные комплексы»

«Методы искусственного интеллекта»

Отчёт по лабораторной работе №5

Вариант - 9

Выполнил:

студент группы ИСТбд-42

Исхаков Тимур

Ульяновск

2022

Исследование инструментов классификации библиотеки Scikit-learn.

1) Для исследования в лабораторной работе были выбраны следующие классификаторы:

- метод опорных векторов

- дерево решений

- логистическая регрессия

И набор данных для классификации: <https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link>, который содержит информацию о пациентах с раком легких.

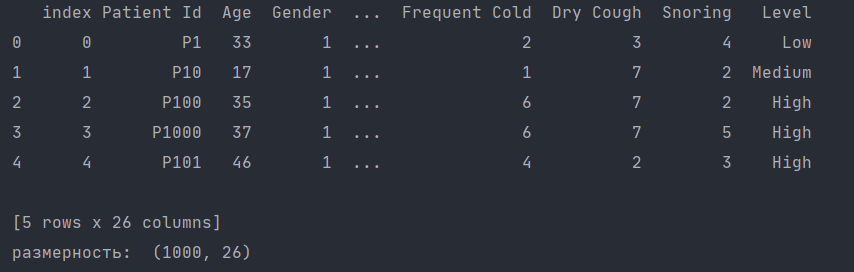
2) Обучение происходит по всем признакам.

3) Используемый датасет содержит следующий набор признаков:

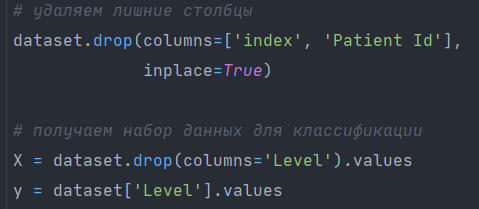
* Age - возраст пациента
* Gender – пол пациента
* Air Pollution - подверженность пациента загрязнению воздуха
* Alcohol use - употребление алкоголя пациентом
* Dust Allergy - аллергия на пыль у пациента
* OccuPational Hazards – опасности для пациента из-за особенностей профессии
* Genetic Risk - генетический риск пациента
* chronic Lung Disease – уровень хронического заболевания легких у пациента
* Balanced Diet - уровень сбалансированного питания пациента
* Obesity – уровень ожирения пациента
* Smoking – уровень курения пациента
* Passive Smoker - пассивное курения пациента
* Chest Pain - боль в груди у пациента
* Coughing of Blood - кашель кровью пациента
* Fatigue - усталость пациента
* Weight Loss - потеря веса пациента
* Shortness of Breath - одышка пациента
* Wheezing - свистящее дыхания пациента
* Swallowing Difficulty - затруднение глотания у пациента.
* Clubbing of Finger Nails - утолщение ногтей на пальцах рук пациента

Для классификаторов определен целевой столбец – Level (Вероятность заболевания – High – высокий уровень, Medium – средний уровень, Low – низкий уровень).

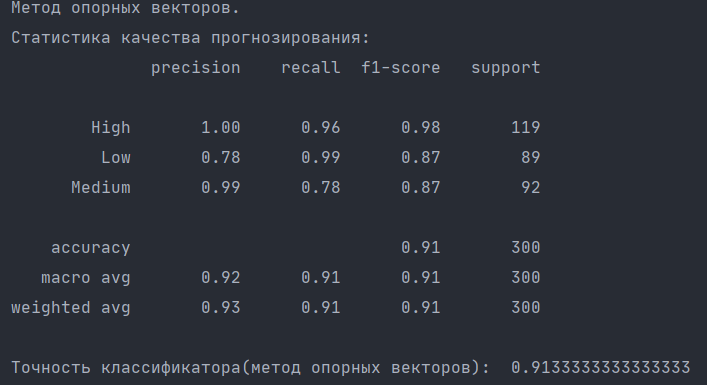
4) Просмотрим первые строки набора данных и получим информацию о размерности датасета



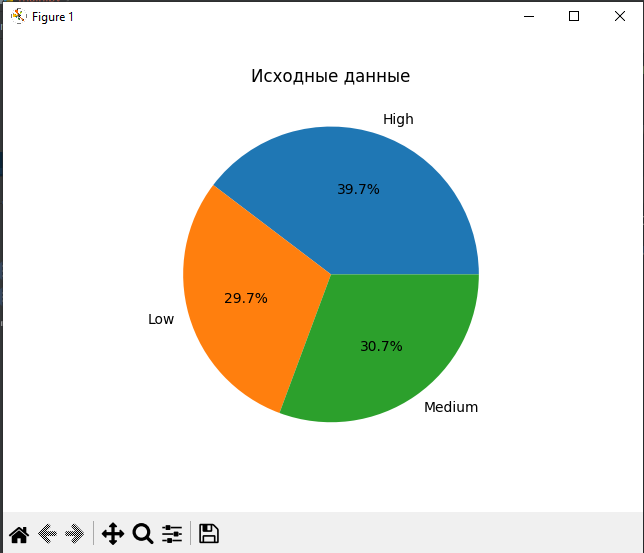
5) Проведем предобработку данных, удалим лишние столбцы, разделим данные на тренировочную и тестовую выборку.

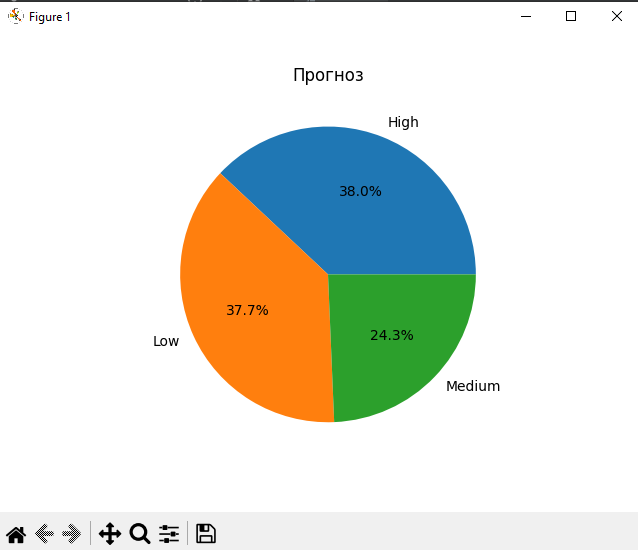


5) Используем классификатор по методу опорных векторов. Оценим точность его работы:

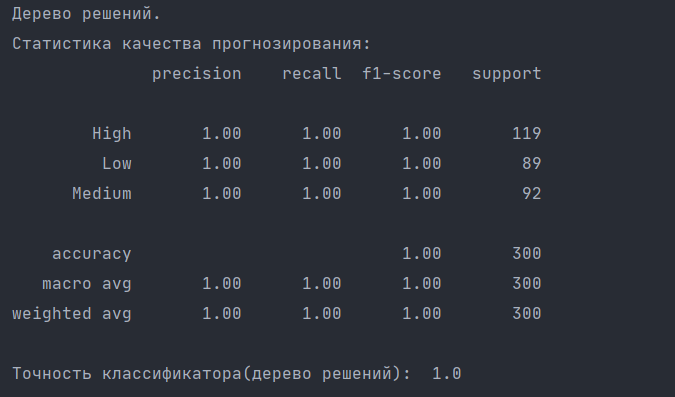


Представим результаты анализа в графической форме:

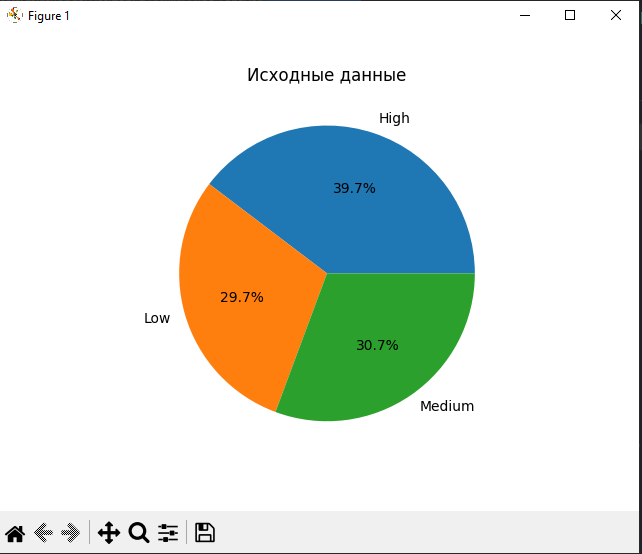


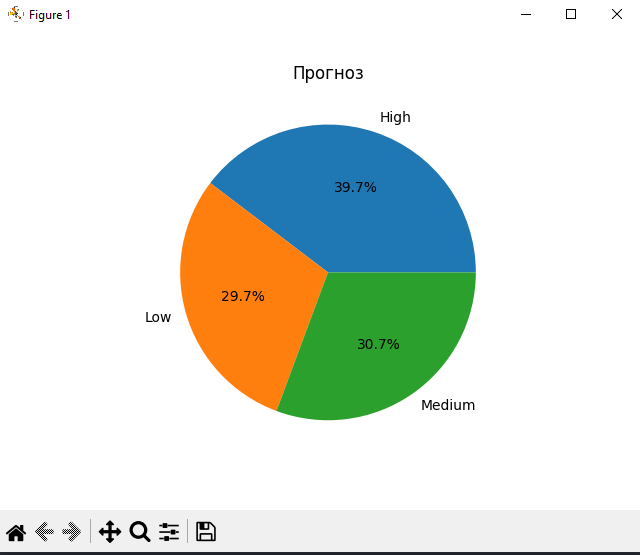


6) Используем классификатор дерева решений:

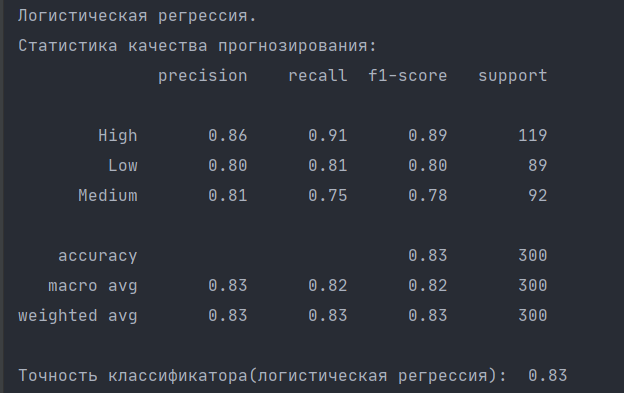


Результаты анализа в графической форме:

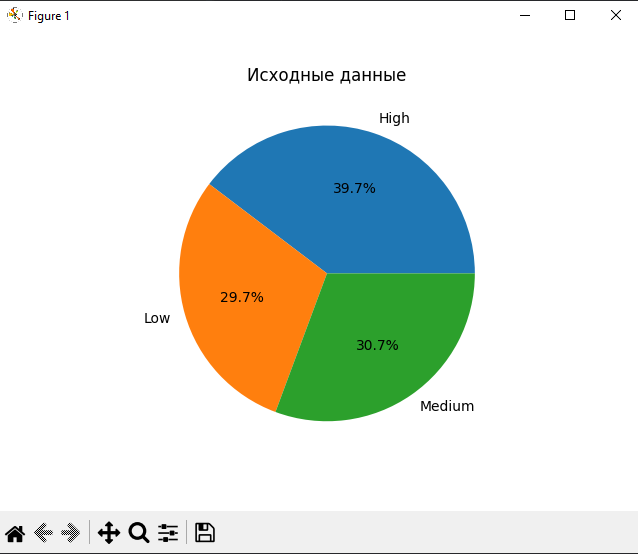


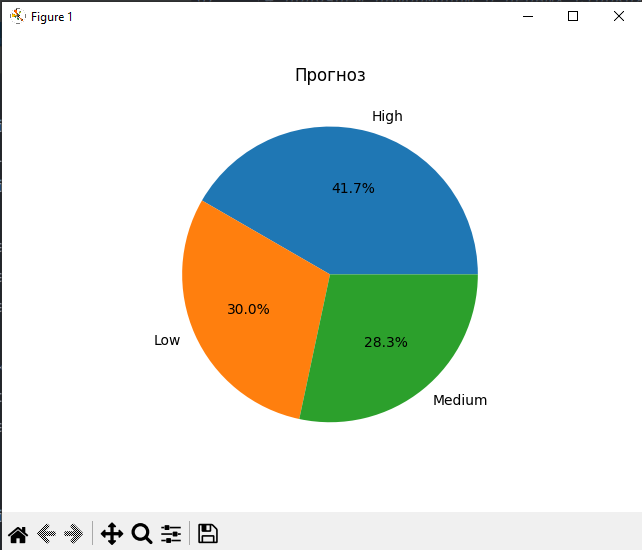


7) Используем классификатор логистической регрессии:



Результаты анализа в графической форме:





Таким образом, в результате выполнения лабораторной работы были исследовали инструменты классификации библиотеки Sklearn, исходя из полученных табличных и графических результатов, можно сказать, что наиболее точным классификатором является классификатор дерева решений, который безошибочно справляется со своей задачей. Наихудшим из рассмотренных классификаторов проявил себя классификатор логистической регрессии.