МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра «Информационно-вычислительные комплексы»

«Методы искусственного интеллекта»

Отчет по лабораторной работе №5

Выполнила:

студентка группы ИСТбд-42

Арсланова Эмилия

Проверил:

доцент кафедры ИВК, к.т.н.

Шишкин В.В.

Для выполнения лабораторной работы был выбран датасет «Breast Cancer Diagnosis» с сайта https://www.kaggle.com/datasets.

Этот набор данных содержит данные о диагностике рака молочной железы. Данные состоят из идентификатора пациента, диагноза (злокачественное или доброкачественное образование) и 30 вычисленных характеристик, извлеченных из оцифрованного изображения тонкоигольной аспирации (ТАБ) образования молочной железы. Особенности включают радиус, текстуру, периметр, площадь, гладкость, компактность, вогнутость и вогнутость точек, а также симметрию и фрактальную размерность.

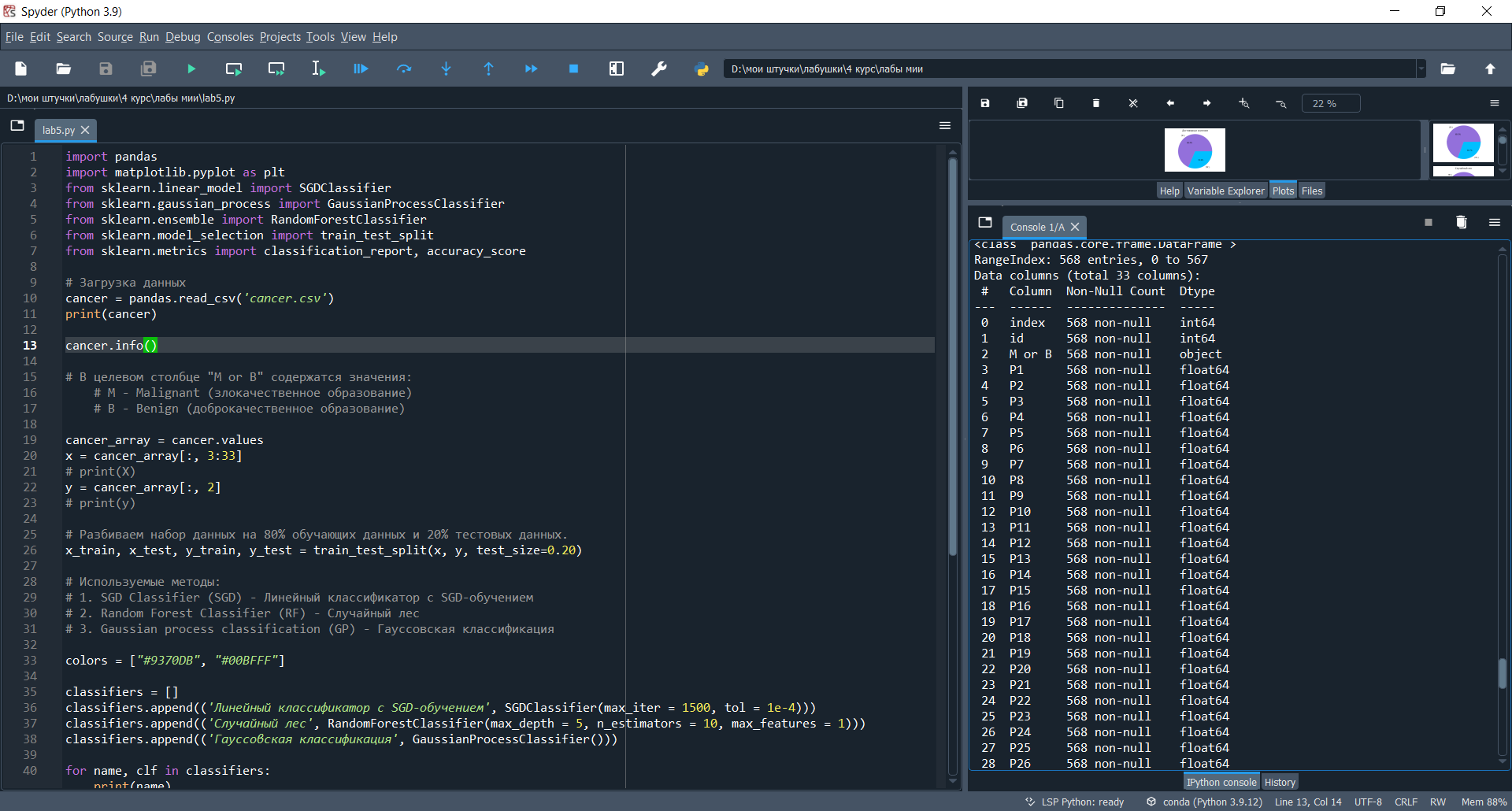
Создано известными исследователями в области общей хирургии и компьютерных наук из Университета Висконсин-Мэдисон под руководством доктора Уильяма Вольберга при участии профессора Ника Стрита и Ольви Л. Мангасаряна. Этот набор данных был использован в некоторых новаторских исследованиях для прогнозирования течения рака молочной железы с использованием методов линейного программирования.

В качестве целевого был выбран столбец, содержащий информацию о диагнозе (M - Malignant (злокачественное образование) B – Benign доброкачественное образование)). В качестве набора признаков выбраны столбцы, содержащие характеристики, описывающие каждый из представленных случаев заболевания.

Для исследования также выбраны следующие классификаторы:

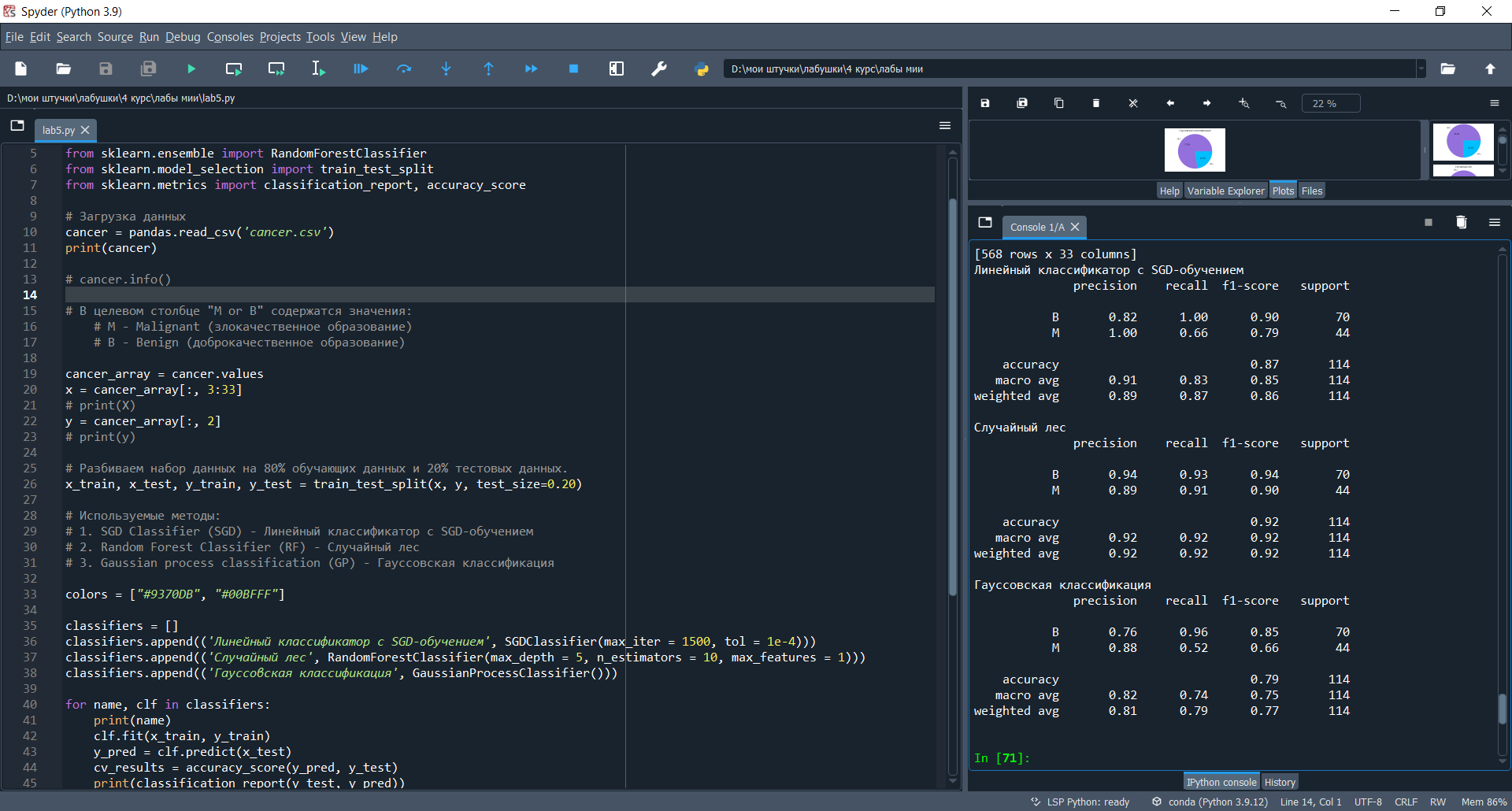
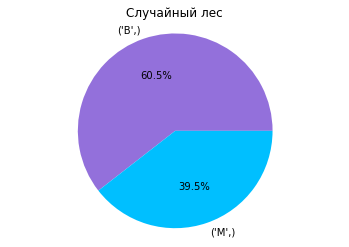
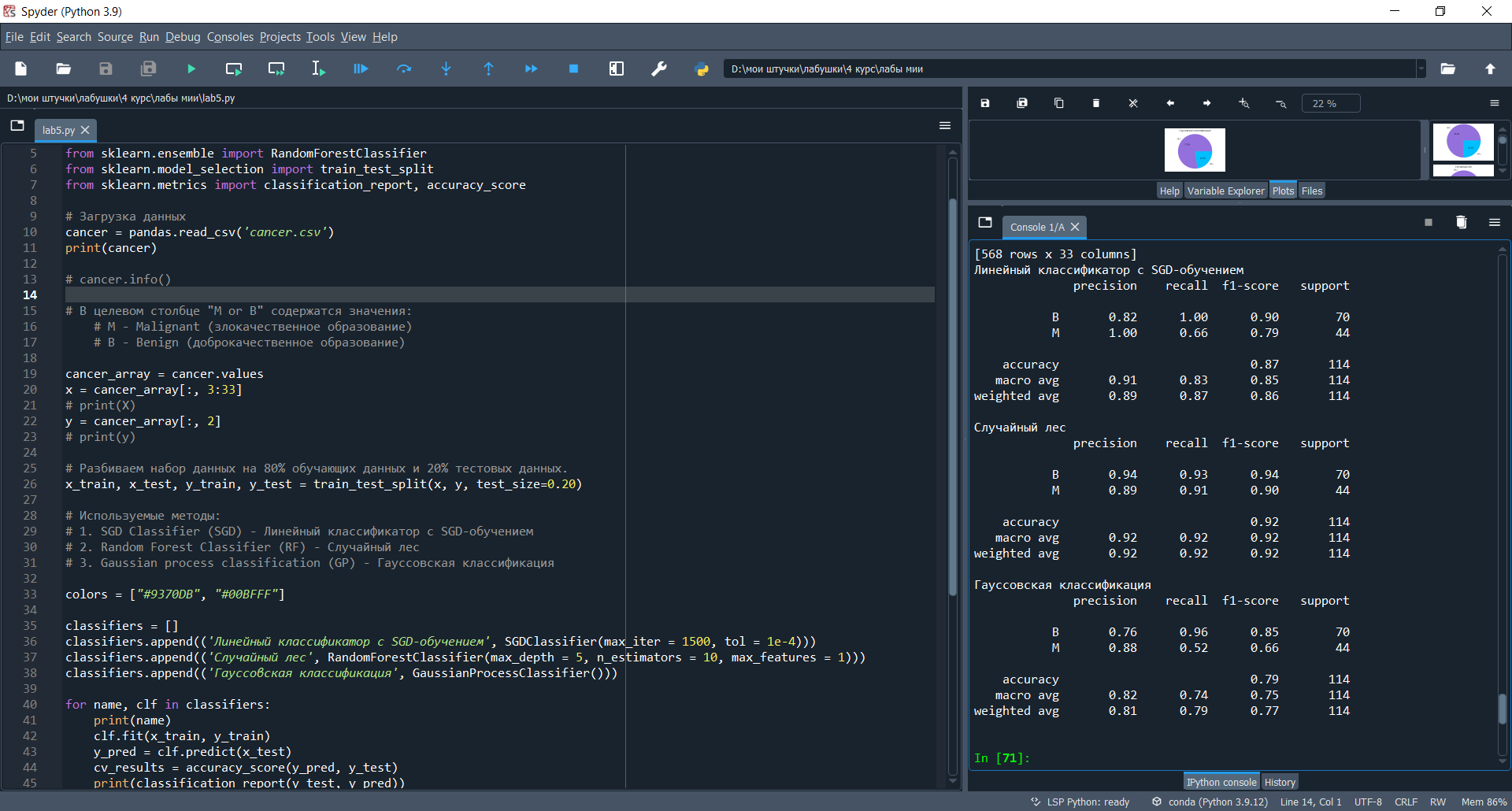
1. SGD Classifier - Линейный классификатор с SGD-обучением
2. Random Forest Classifier - Случайный лес
3. Gaussian process classification - Гауссовская классификация

При подготовке датасета было выявлено, что он не нуждается в очистке:



В результате обучения и оценки моделей получены следующие результаты:





Таким образом, самым точным оказался классификатор «Случайный лес», его точность составляет 92%. Далее – линейный классификатор с SGD-обучением, его точность – 87%. Самый неточный – Гауссовская классификация, точность – 79%.

