Министерство науки и высшего образования Российской Федерации **Федеральное государственное автономное образовательное учреждение  
высшего образования**

**«ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»  
(национальный исследовательский университет)**

Высшая школа электроники и компьютерных наук

Кафедра системного программирования

**ОТЧЁТ ПО ЗАДАНИЮ №3**

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Байесовская классификация

Выполнил

студент группы КЭ-120

Глизница Максим Николаевич

E-mail: letadllo@mail.ru

Челябинск 2021

# Задание

Выполните классификацию набора данных с помощью Байесовской классификации, варьируя соотношение мощностей обучающей и тестовой выборок от 60%:40% до 90%:10% с шагом 5%.

Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм.

# Краткие сведения о наборах данных

Использованный набор данных:

Mushroom Classification (<https://www.kaggle.com/uciml/mushroom-classification>). Содержит информацию о типичных признаках определённых видов грибов, таких как форма шляпки, запах, и других. Также включает информацию о том, является ли данный вид гриба ядовитым. Расшифровка используемых в файле сокращений находится в файле attr.txt. Набор содержит 3916 ядовитых и 4208 неядовитых грибов

# Краткие сведения о средствах реализации

Для реализации методов была использована библиотека scikit-learn, включающая в себя множество алгоритмов для анализа данных.

Репозиторий по дисциплине: <https://github.com/Airpllane/DAAlgorithms>. Каталог для задания: 3. Bayes.

# Визуализация показателей качества

Для визуализации были использованы соотношения мощностей обучающей и тестовой выборки от 60%:40% до 90%:10% с шагом 5%. Был использован классификатор GaussianNB, реализующий алгоритм наивной байесовской классификации с использованием нормального (Гауссовского) распределения вероятностей.

Была выполнена визуализация четырёх метрик качества на выбранных соотношениях. Результаты визуализации приведены на рис. 1.

Из метрик можно увидеть, что, в целом, качество классификации получилось высоким в соответствии со всеми использованными метриками. Наиболее удачными разбиениями оказались 60%:40%, 70%:30% и 85%:15%.

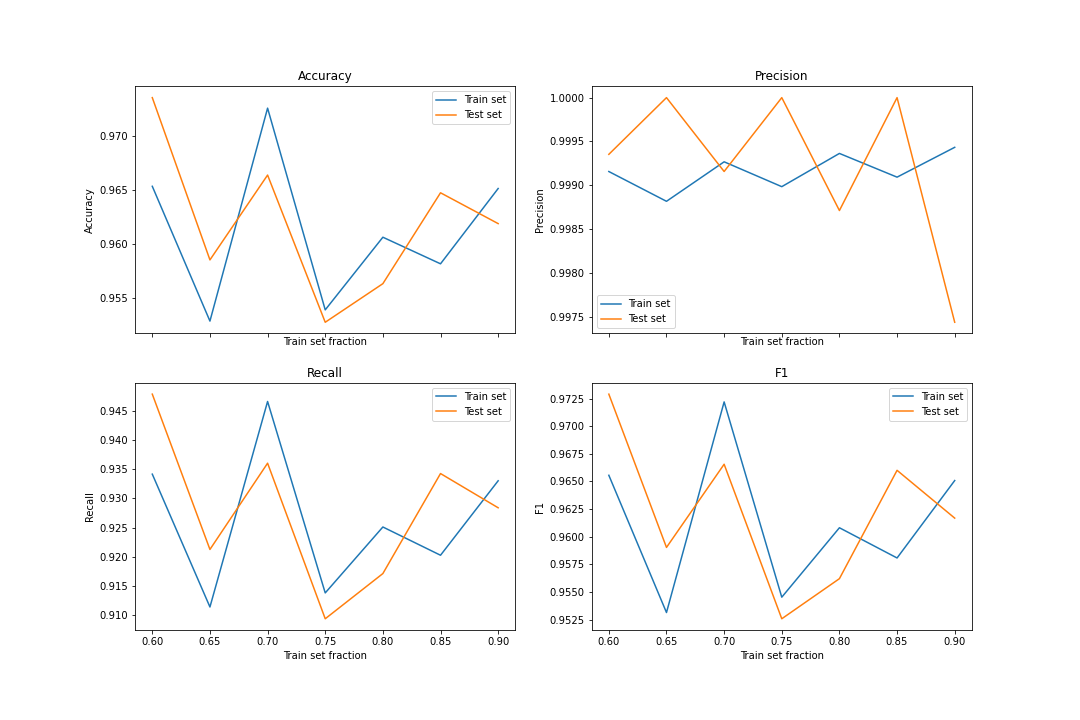


Рис. . Результаты визуализации

Поскольку использованный набор данных содержит данные о ядовитости грибов, классификация должна стремиться уменьшить количество ложноотрицательных результатов (то есть ядовитые грибы не должны быть классифицированы как неядовитые). Таким образом, наиболее важной метрикой является recall (полнота), которая зависит от количества ложноотрицательных результатов. Эта метрика имеет наиболее низкие значения среди четырёх исследованных, что говорит о том, что качество классификации несколько ниже, чем может показаться изначально.