Отчет по лабораторной работе № 1 по курсу «Машинное обучение»

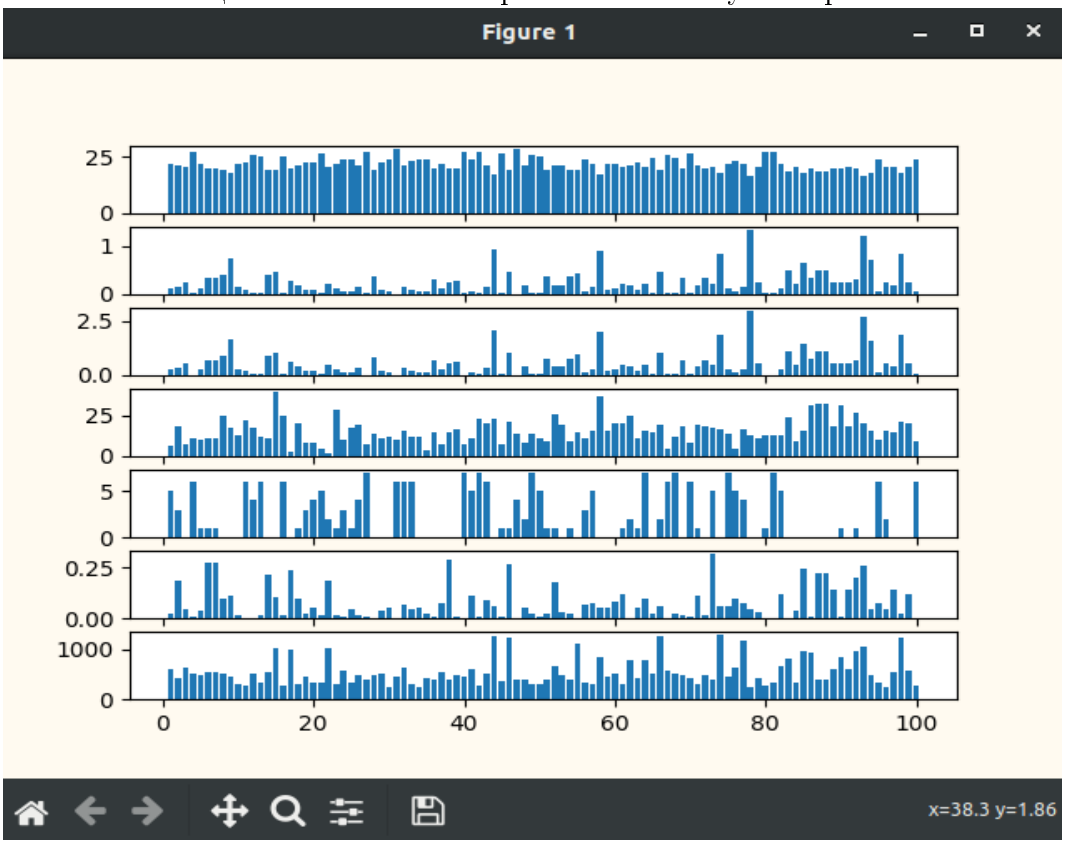
Студент группы М8О-308б-18 МАИ Игитова Александра

1. Постановка задачи

Найти себе набор данных, для следующей лабораторной работы, и проанализировать его. Выявить проблемы набора данных, устранить их. Визуализировать зависимости, показать распределения некоторых признаков. Реализовать алгоритмы K ближайших соседей с использованием весов и Наивный Байесовский классификатор и сравнить с реализацией библиотеки sklearn.

2. Реализация

В качестве набора данных была взята статистика о метеоритах собранная NASA с сервиса Kaggle. Метериты разделяются на 2 класса: опасные и не оспасные. При анализе данных были убраны Кеплеровы элементы орбиты как невлияющие на опасность. С помощью библиотеки matplotlib были визуализированы оставшиеся признаки.



По визуализации можно увидеть что выбросы отсутствуют. Далее парсим наши данные, разделяя их на фичи и принадлежность к классу. Затем проверяем точность алгоритмов k ближайших соседей и наивный Байесовский классификатор реализованных в библиотеке sklearn.

```
trol53@trol53-Inspiron-15-3567:~/AI/MLlab1$ python3 test.py
KNN: 66.16853932584269
Naive bayes: 66.62531017369727
```

Алгоритмы на наших данных выдали 66 и 65 процентов точности. Далее мы реализуем свои алгоритмы. Идея алгоритма к ближайших соседей в следующем: рассматривается n-мерное пространство, где n - количество аттрибутов а сами аттрибуты являются координатами, далее при поступлении тестовых данных мы находим декартово расстояние между новым объектом и всем данными в тренировочной выборке, затем выбирается к ближайших к нему точек и новому объекту присваивается тот класс, к которому принадлежит большинство его ближайших соседей. Так же можно суммировать количество соседей с весом, в нашем случае весом является расстояние, т.е. суммируются $1/dist$. Наша реализация выдала результат около 75%, что выше результата реализации sklearn. Теперь о реализации наивного Байесовского классификатора. В основе лежит теорема Байеса, а если точнее идея о том что все аттрибуты влияют на принадлежность к классу независимо друг от друга. Мы находим среднее значение и среднее отклонение по каждому аттрибуту. Далее мы группируем полученные данные по классам. Затем для новых объектов рассчитывается Гауссовская функция плотности для всех атрибутов. Затем объединяя их и сравнивая с тренировочными данными получаем вероятность принадлежности к классу. Наша реализация выдает точность около 94% что сильно выше реализации sklearn.

```
trol53@trol53-Inspiron-15-3567:~/AI/AI/ML1$ python3 bayes.py
Accuracy: 94.94949494949495%
trol53@trol53-Inspiron-15-3567:~/AI/AI/ML1$ python3 knn.py
Accuracy: 75.14124293785311%
```

3. Выводы

В данной лабораторной работе были рассмотрены самые простые алгоритмы классификации. Также было полезным изучить работу с данными, в частности такие аспекты как поиск, парсинг и чистка данных. Около половины времени ушло именно на это. Также полезным было изучение основ такой библиотеки как sklearn, которая является неотъемлемой частью при работе с данными.