Отчет по лабораторной работе №0
по курсу «Искусственный интеллект»

Выполнил студент группы М8О-304б-16 Величутин Андрей

Тема: Получение и предобработка данных

Задача:
Требуется сформировать/получить два набора данных соответствующие следующим критериям:
1. Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна
2. Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классификации). Если такого признака нет, спроектировать

По каждому датасету построить распределения признаков (в случае корпуса документов – построить распределение слов) и объяснить имеющуюся картину. Вычислить статистические характеристики признаков. Обнаружить и решить возможные проблемы с данными. Если решить данную проблему невозможно, объяснить почему.

Оборудование студента:
Intel(R) Core(TM) i5-5200 CPU @ 2.20GHz 2.19GHz
ОЗУ 6,00 ГБ

Программное обеспечение:
Windows 10, Python 3.7.4(С библиотеками Pandas, Numpy, Seaborn и Scikit-Learn), Jupyter notebook 4.4.0

Ход работы:

Для начала рассмотрим первый датасет: https://www.kaggle.com/c/ghouls-goblins-and-ghosts-boo

В данном датасете имеются следующие столбцы:

- id — идентификатор существа
- bone_length — средняя длина кости, нормализована
- rotting_flesh — процент гнилой плоти
- hair_length — средняя длина волос, нормализована
- has_soul — процент души существа
- color — цвет существа, категориальный признак
- type — тип существа, предсказываемое значение

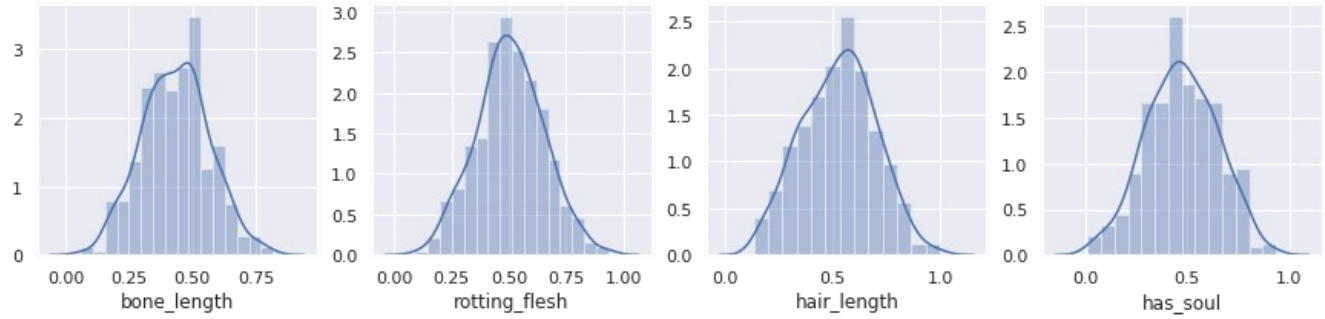
id	bone_length	rotting_flesh	hair_length	has_soul	color	type
0	0.354512	0.350839	0.465761	0.781142	clear	Ghoul
1	0.575560	0.425868	0.531401	0.439899	green	Goblin
2	0.467875	0.354330	0.811616	0.791225	black	Ghoul
4	0.776652	0.508723	0.636766	0.884464	black	Ghoul
5	0.566117	0.875862	0.418594	0.636438	green	Ghost
7	0.405680	0.253277	0.441420	0.280324	green	Goblin
8	0.399331	0.568952	0.618391	0.467901	white	Goblin
11	0.516224	0.536429	0.612776	0.468048	clear	Ghoul
12	0.314295	0.671280	0.417267	0.227548	blue	Ghost
19	0.280942	0.701457	0.179633	0.141183	white	Ghost

Библиотека pandas позволяет одним методом вывести основные характеристики числовых и категориальных признаков:

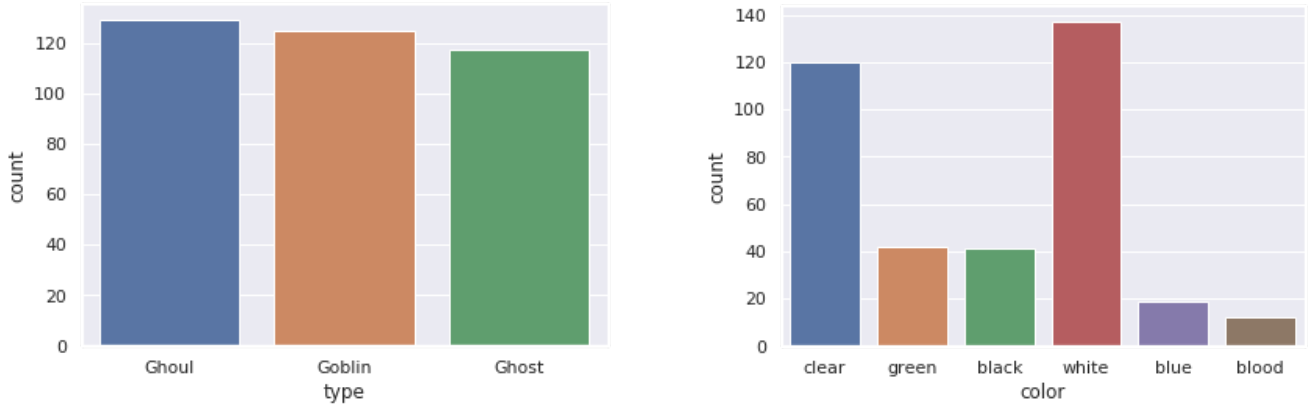
	bone_length	rotting_flesh	hair_length	has_soul
count	371.000000	371.000000	371.000000	371.000000
mean	0.434160	0.506848	0.529114	0.471392
std	0.132833	0.146358	0.169902	0.176129
min	0.061032	0.095687	0.134600	0.009402
25%	0.340006	0.414812	0.407428	0.348002
50%	0.434891	0.501552	0.538642	0.466372
75%	0.517223	0.603977	0.647244	0.600610
max	0.817001	0.932466	1.000000	0.935721

	color	type
count	371	371
unique	6	3
top	white	Ghoul
freq	137	129

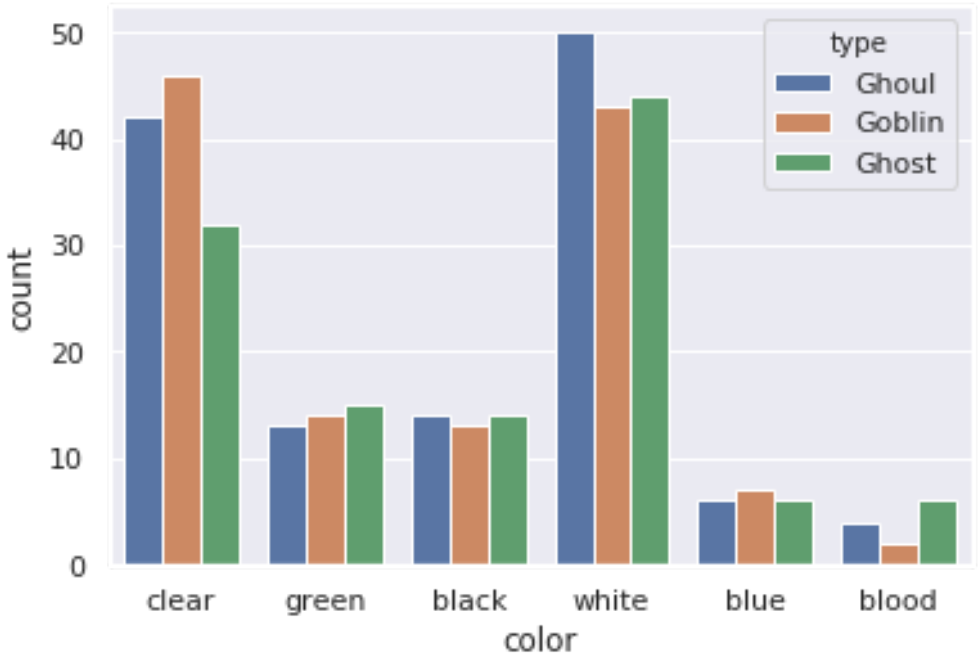
С помощью библиотеки seaborn можно построить распределение числовых признаков на графике (seaborn.distplot):

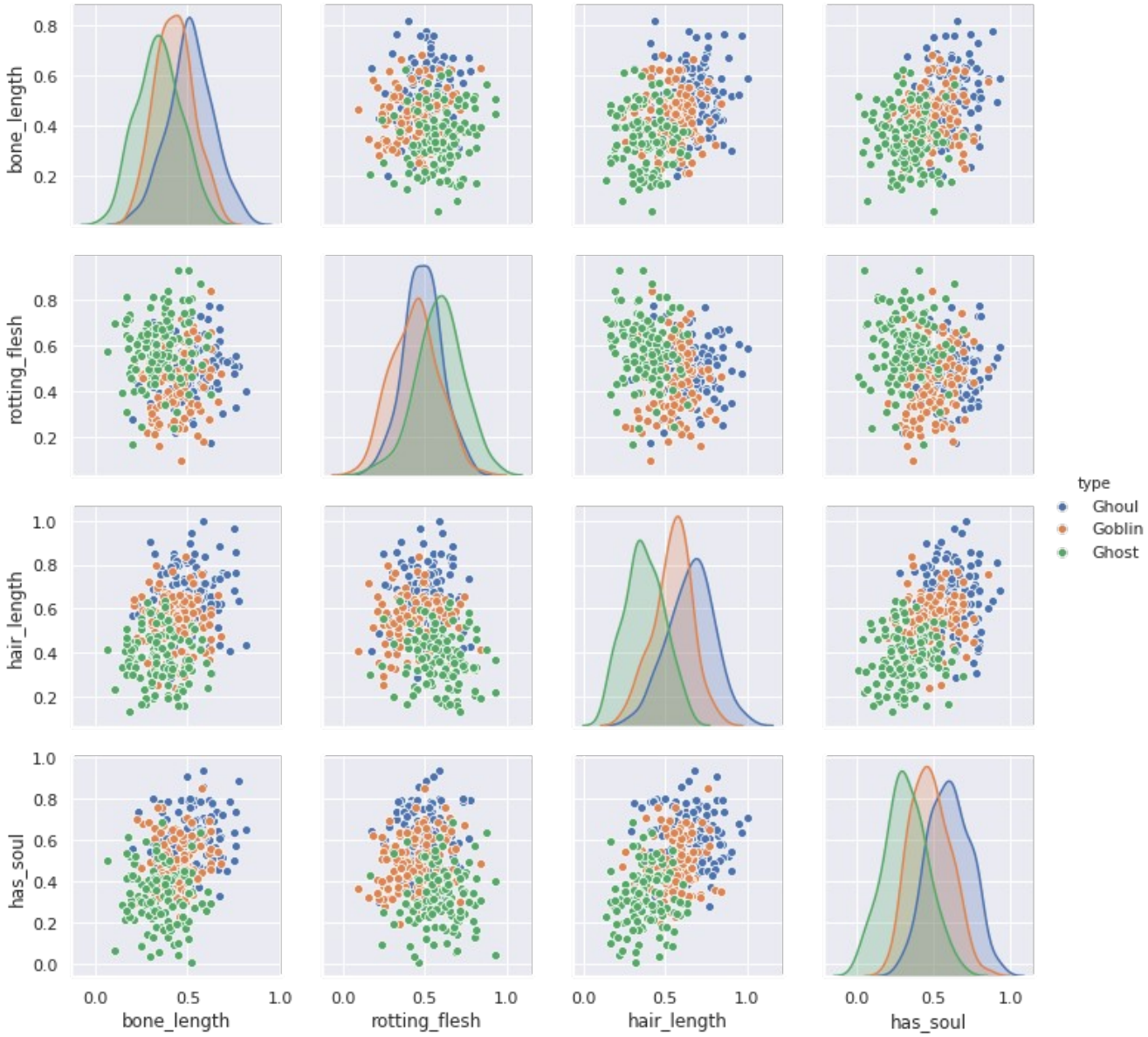


И распределение категориальных признаков (seaborn.countplot):



Можно посмотреть как распределена целевая пременная в зависимости от значений признаков:





Intel(R) Core(TM) i5-5200 CPU @ 2.20GHz 2.19GHz
ОЗУ 6,00 ГБ

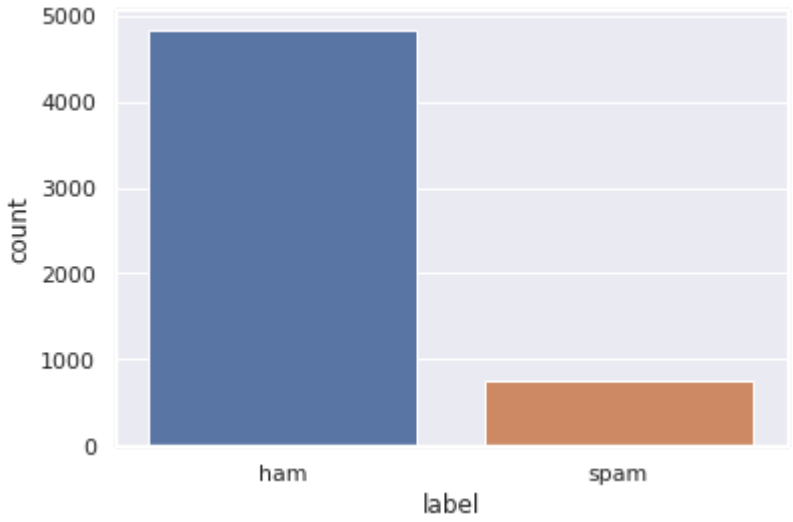
На графиках отчетливо видно три кластера, которые соответствуют трем различным классам. Скорее всего даже линейные классификаторы дадут хорошую точность.

Рассмотри второй датасет: https://www.kaggle.com/uciml/sms-spam-collection-dataset

Он представляет из себя набор текстовых документов (А точнее sms). По тексту сообщений неообходимо будет относить документ либо к классу «спам», либо к классу «не спам»

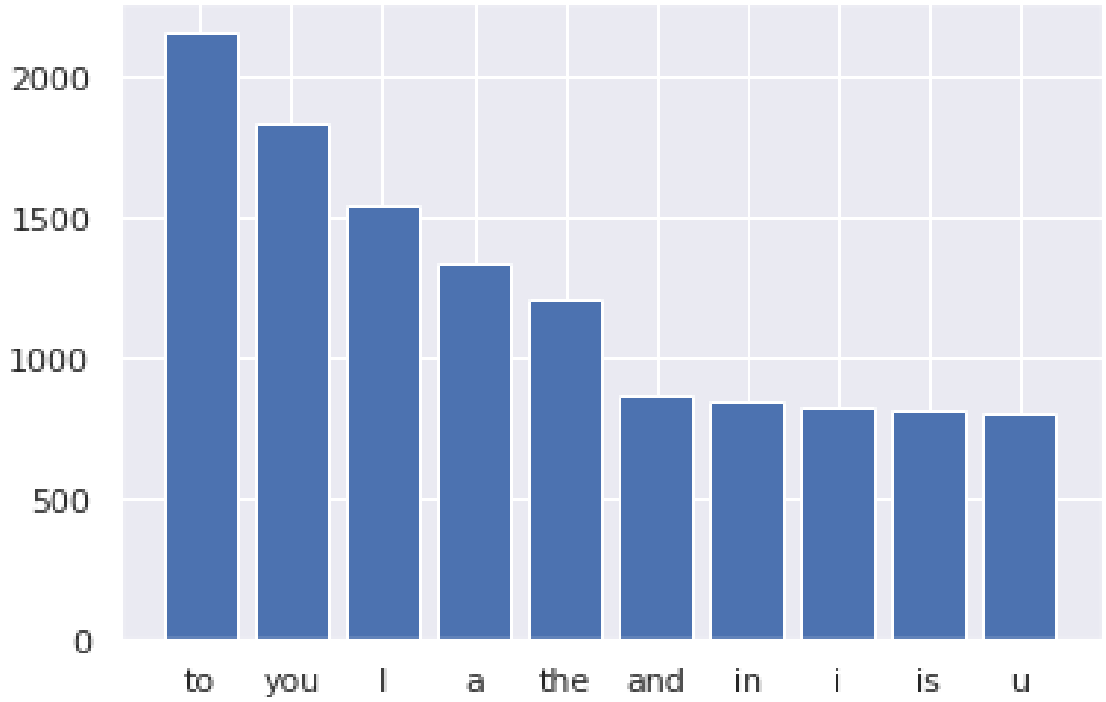
	label	message
0	ham	"Go until jurong point, crazy.. Available only...
1	ham	Ok lar... Joking wif u oni...\n
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	"Nah I don't think he goes to usf, he lives ar...
5	spam	"FreeMsg Hey there darling it's been 3 week's ...
6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...

Можно посмотреть на соотношение классов:



На глаз явный дисбаланс классов, это может быть проблемой. Также есть небольшая проблема с сообщениями. Какие то из них обернуты в ковычки, имеются лишние переносы на конце строки и др. Однако, это не должно мешать извлечению слов. К тому же формат csv не очень пригоден для хранения текстовых данных, так как разделители могут быть частью этих данных. Именно из-за этого пришлось вручную прописывать чтение данных, а не использовать готовую функцию pandas.read_csv(...)

Теперь можно рассмотреть распределение самых частоупотребляемых слов в сообщениях:



Как видно, самые часто употребляемые слова относятся в основном к артиклям, предлогам и местоимениям.

Весь код, который был написан для выполнения задания находится по ссылке: https://github.com/BeJIuK/ML/blob/master/Lab0/Lab0.ipynb

Выводы:

Данная лабораторная работа показалась мне достаточно интересной. Я познакомился с библиотекой pandas и seaborn. Визуализация данных очень часто оказывается полезной в том плане, что она позволяет отобразить возможные проблемы в датасете (Или наоборот — какие-то подсказки, которые позволят лучше классифицировать данные)