Московский авиационный институт
(национальный исследовательский университет)

Институт №8 «Информационные технологии и прикладная математика»
Кафедра 806 «Вычислительная математика и программирование»
Дисциплина «Искусственный Интеллект»

Лабораторная работа №0

Тема: Data Mining и исследование данных

Студент: Глушатов И.С.
Группа: М8О-307Б-19
Преподаватель: Ахмед Самир. Х.
Дата:
Оценка:

Москва, 2022

Цель работы: научиться работать с данными, прослеживать зависимости, избавляться от лишних и добавлять нужные признаки. Визуализировать данные. Подготавливать их к дальнейшей работе с алгоритмами обучения.

Задание:

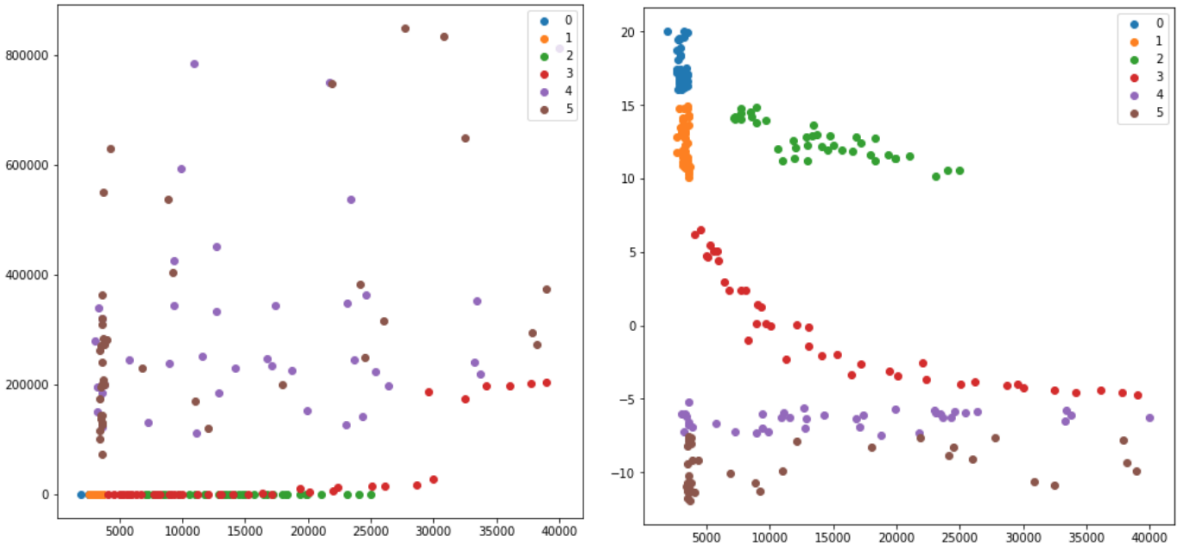
В данной лабораторной работе вы выступаете в роли предприимчивого начинающего стартапера в области машинного обучения. Вы заинтересовались этим направлением и хотите предложить миру что-то новое и при этом неплохо заработать. От вас требуется определить задачу, которую вы хотите решить и найти под нее соответствующие данные. Так как вы не очень богаты, вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще. Вы готовитесь представить отчет ваши партнерам и спонсорам, от которых зависит дальнейшая ваша судьба.

Ход работы

В качестве задания я выбрал – классификацию звезд на основе множества данных, а именно: температура звезды, светимость, радиус, звездная величина. Также представлены сведения о цвете звезды и её спектральном классе.

	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute magnitude(Mv)	Star type	Star color	Spectral Class
0	3068	0.002400	0.1700	16.12	0	Red	M
1	3042	0.000500	0.1542	16.60	0	Red	M
2	2600	0.000300	0.1020	18.70	0	Red	M
3	2800	0.000200	0.1600	16.65	0	Red	M
4	1939	0.000138	0.1030	20.06	0	Red	M

Так как спектральный класс и цвет звезды – не непрерывные величины, я решил исключить эти данные из рассмотрения, сосредоточившись на непрерывных величинах.



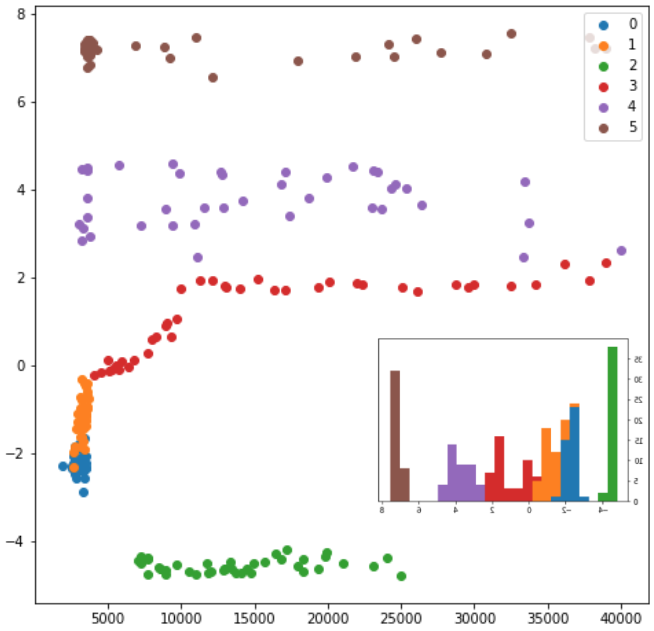
На приведенных рисунках изображены графики зависимости:

1) Слева – Светимость от Температуры
2) Справа – Звёздная величина от Температуры.

Если не рассматривать для каждого класса отдельно, то мы видим, что как таковой корреляции между данными параметрами нет. Однако на данных изображениях можно выделить особенности для каждого класса:
4 и 5 классы имеют очень большую светимость относительно других классов, при этом звёздная величина (ЗВ) находится в диапазоне от -14 до -7 для 5-го класса и от -7 до -5 для 4-го класса.
ЗВ 3-го класса лежит в пределах от -4 до 8, причем никакого класса больше в этом диапазоне нет.
ЗВ класса 0 находится в пределах от 15 до 20, а классы 1 и 2 от 10 до 15. Фактически по ЗВ мы можем однозначно определить 4 из 6 класса. А если брать в расчет температуру, то можно построить R – дерево, которое сможет очень хорошо

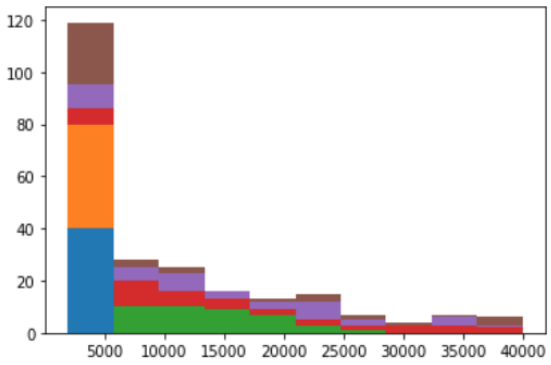
классифицировать звезды по этим двум параметрам. Однако есть зависимость еще лучше.

На рисунке приведена зависимость логарифма (натурального) радиуса звезды от температуры.
Видно, что опять же нет никакой линейной зависимости между параметрами, однако можно заметить, что по логарифму радиуса можно практически наверняка определить класс звезды.

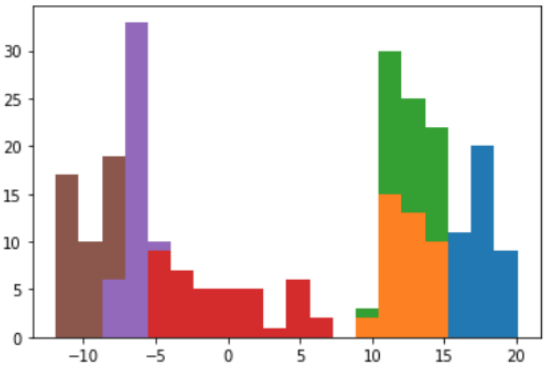


Тот же самый вывод можно сделать, глядя на гистограмму распределения классов по логарифму радиуса, что находится внутри рисунка.

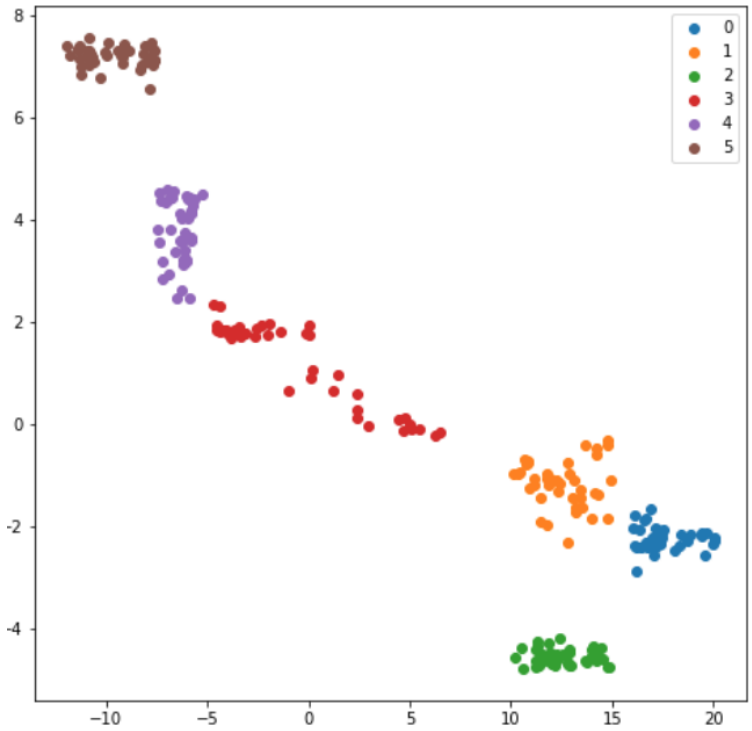
Вторая гистограмма показывает распределение звезд по температуре. Видно, что температурный признак мало говорит нам о классе объекта.



Третья гистограмма показывает распределение звезд по звездной величине. Оно достаточно хорошо разделяет классы.



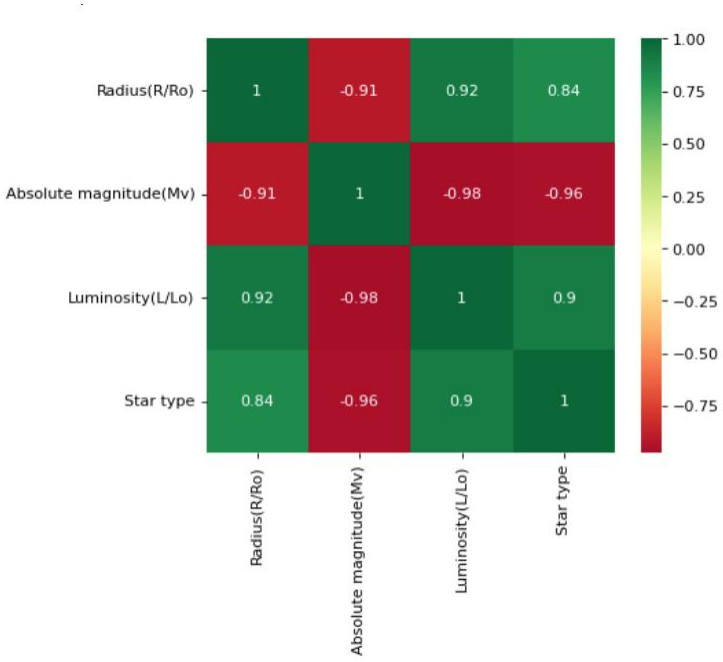
На основе уже полученных данных мы можем предпринять следующие: температуру, как признак можно в принципе выбросить и не учитывать. Следует ввести новый признак – логарифм радиуса. Учитывая хорошие показатели звездной величины и логарифма радиуса, можно построить график зависимости между этими величинами:



Получился очень даже неплохой результат. Каждую группу можно четко выделить, а между самими величинами прослеживает корреляция и мы видим, что с увеличением звездной величины логарифм радиуса, а следовательно, и сам радиус уменьшаются. Т. е. мы имеем обратную пропорциональность.

Похожий результат мы можем получить и с другими данными. Если также ввести новый признак – логарифм светимости, то в совокупности мы получим очень хорошо коррелируемые признаки, что нам может продемонстрировать матрица корреляции снизу.

Однако 3В и логарифм радиуса остаются наилучшими признаками для классификации объектов.



Выводы

В ходе лабораторной работы я определил задачу классификации звезд на основе признаков температуры, радиуса, светимости и звездной величины. Нашел данные для анализа. Всего звезды разделены на 6 классов. Я считаю, что данных достаточно, и что по ним можно построить хорошие классификаторы, так как в итоге получилось найти такие зависимости, благодаря которым можно чётко увидеть разделения по группам. В работе я визуализировал признаки, провел оценку их состоятельности, как признака, необходимого для классификации. В итоге: температура звезды была исключена, так как не носила информативный характер при классификации; также были исключены признаки цвета звезды и её спектрального класса, так как являлись перечисляемыми типами; признаки радиуса звезды и светимости были заменены на их логарифмы – такой ход позволил выделить зависимость между признаками и безопасно повысить коэффициент коррелируемости. Следует заметить, что по приведенной таблице (судя по графику), можно сделать хорошее R – дерево, способное классифицировать звезды. Github файл.