

ИДЗ

В рамках выполнения ИДЗ необходимо на предложенном датасете решить связанную с ним задачу. При выполнении ИДЗ надо самостоятельно провести подготовку данных и выбрать подходящую модель МО. Выполнять ИДЗ можно в бригадах до 3 студентов (студенты могут быть из разных групп).

При выборе задания необходимо минимизировать кол-во повторяющихся вариантов. Поэтому, если будет слишком много повторений, то бригаде может быть изменен вариант на усмотрение преподавателей.

Требования

Требования к разрабатываемой модели

- Модель должна быть разработана на языке Python. Разрешается использовать существующие библиотеки (например, SKLearn или mlxtend)
- В исходном коде должны быть поясняющие комментарии
- Для апробации качества модели необходимо разбивать датасет на обучающее и тестовое подмножество. Разбиение необходимо повторить несколько раз, тем самым показав, что выбранная модель дает одинаковый результат вне зависимости от обучающего подмножества
- *Плюсом будет сравнение нескольких моделей МО*

Требования к отчету

- В отчете должно быть описание датасета, а также описание решаемой задачи
- В отчете должен быть проведен начальный анализ данных. Проведение статистического анализа, анализа на необходимость нормировки данных, обоснование применения определенного вида нормировки, и.т.д.
- В отчете должно быть обоснование выбора модели
- В конце отчета должен быть приведен анализ результирующей модели, а также перечислены возникшие проблемы (и как они были решены) и проблемы, которые решить не удалось. Плюсом будет предложение по улучшению модели.
- Если работа выполнена бригадой, то в отчете должно быть указано, кто в бригаде за что отвечал (написание отчета не является зоной ответственности)
- В приложении должен быть исходный код

Варианты

После названия датасета в скобках указана сложность задачи. Сложность задачи влияет на строгость оценивания ИДЗ. Для более сложных задач допускается получение не самого лучшего результата.

Также для каждого датасета указано количество задач. Разные бригады могут решать разные задачи на одном датасете.

Zoo (1) - 1 задача

Датасет булевых значений. Включает в себя информацию о 101 животном из зоопарка, которые относятся к 7 классам.

- Задача заключается в классификации животного по его описанию.

Wine Quality (1) - 1 задача

Числовой датасет. Включает себя информацию о химических характеристик вина и его оценку по шкале от 0 до 10. Состоит из 4898 наблюдений.

- Задача заключается в предсказании качества вина на основе его химических характеристик. (Задача регрессии)

Groceries dataset (1) - 1 задача

Датасет содержит id покупателя, дату покупки, и наименование купленного товара. Датасет содержит 38.8 тысяч записей.

- Задача заключается в построении ассоциативных правил и анализе полученных правил

Lonosphere (2) - 1 задача

Датасет содержит информацию о сигнале и том, отразился ли он от ионосферы. Датасет состоит из 351 наблюдения.

- Задача кластеризации. Составить кластеры сигналов и провести анализ полученных кластеров и того, как они связаны с тем, что сигнал отразился или нет.

Chocolate Bar Ratings (2) - 1 задача

Датасет содержит информацию о шоколаде. Датасет содержит 1795 наблюдений. Один из признаков отвечает за рейтинг шоколада.

- Задача заключается в выделение значимых признаков и построении регрессионной модели для предсказания рейтинга.

Mobile Price Classification (2) - 1 задача

Датасет содержит информацию о характеристиках телефона и диапазоне цены (от 0 до 3). Предварительно датасет разбит на обучающую и тестовую выборку. Обучающая выборка содержит 2000 наблюдений, а тестовая 1000 наблюдений.

- Задача заключается в предсказании ценового диапазона телефона по его характеристикам (Задача классификации)

Student Grade Prediction (2) - 3 задачи

Датасет содержит информацию о студентах и их оценках за 3 семестра. Датасет состоит из 395 наблюдений.

- Задача регрессии. Предсказать оценки студента на основе информации о нем
- Задача кластеризации. Провести кластеризацию студентов и анализ полученных кластеров. При проведении кластеризации информацию об оценках использовать не надо, но использовать ее при оценке кластеров.
- Задача ассоциативного анализа. На основе ассоциативного анализа найти зависимости между признаками и провести анализ полученных правил. Информацию об оценках использовать не надо.

Gas Turbine CO and NOx Emission (2) - 1 задача

Датасет содержит измерения 11 показателей турбины и выбросов газа CO и NOx. Датасет состоит из 36733 наблюдений

- Задача регрессии. Предсказание значений выбросов на основе показателей турбины

The Blog Authorship Corpus (3) - 2 задачи

Датасет текста. Включает в себя 681288 записей в блоге от 19320 людей, а также информацию о людях.

Для данного датасета

- Задача заключается в определении пола и возраста человека по тексту (Задача классификации)
- Задача ассоциативного анализа. Построение ассоциативных правил для текста с выделением того, какие слова в каких случаях и кем используются. Требуется тщательной подготовки исходных данных.