Машинное обучение 2023

- Всего планируется 4 лабораторных работы
- Каждая работа оценивается в 15 баллов
- Сумма баллов за лабораторные: 60 баллов
- К каждой лабораторной работе предоставляется отчет в электронном виде
- Защита лабораторных работ происходит в формате персональной беседы по теме выполненной лабораторной работы
- Отчет к лабораторной работе должен быть представлен **ДО** защиты этой работы (заранее скинут на почту преподавателя)
- Баллы за каждую работу формируются после сдачи отчета по ней **И** защиты
- Не защищены все лабораторные / не сданы все отчеты - нет допуска к экзамену
- Во всех лабораторных работах используется один датасет (один набор данных), выбранный в ходе работы с л/р №1. Набор данных, выбранный студентом, должен быть уникален в рамках курса обучения, ссылку на него просьба размещать в таблице “МО, датасеты, осень '23. Курс 3/4” в следующем виде:

Фамилия и инициалы	Номер группы	Ссылка на датасет формата https://www.kaggle.com/datasets/…

  Таблица будет размещена в открытом виде, чтобы у всех была возможность посмотреть, какие датасеты уже были выбраны.

Сроки защиты лабораторных работ без снижения баллов:

л/р №1 - до 08.10;

л/р №2 - до 05.11;

л/р №3 - до 03.12;

л/р №4 - до 31.12.

Баллы снижаются в следующих случаях:

- Работа не защищена вовремя: -5 баллов за каждый дедлайн по последующей лабораторной (например, л/р №1, сданная в сроки для л/р №2, оценивается максимально в 10 баллов, в сроки для л/р №3 - в 5 баллов).
- Работа плохо защищена: снижение баллов (от -1 до -5), опционально - отправка студента переделывать лабораторную

Итоговые оценки (какую оценку можно получить “автоматом”):

- 95%+ от общего числа баллов: 5
- 90%+ от общего числа баллов: 4
- 80%+ от общего числа баллов: 3

Лабораторные работы

Лабораторная работа №1. Подготовка и нормализация данных

Лабораторная работа №1 служит для получения и закрепления навыков предобработки данных для дальнейшего применения методов машинного обучения для решения задач.

Студент самостоятельно выбирает набор данных на сайте Kaggle.com.

Основные требования к выбираемому набору данных:

1. Число столбцов признаков – не менее 10;
2. Число записей – не менее 10000;
3. Набор данных имеет пропуски.

В ходе выполнения лабораторной работы должны быть выполнены следующие этапы:

1. Предварительная обработка данных
    a. Визуализация значимых признаков (диаграммы рассеяния, ящики с усами, гистограммы)
    b. Очистка данных (удаление пропусков, нормализация, удаление дубликатов)
    c. Корреляция данных (матрица корреляций)

В рамках защиты лабораторной работы необходимо продемонстрировать jupyter-notebook с кодом, быть готовым пояснить выполненные действия, продемонстрировать базовое понимание работы используемых в работе методов.

Лабораторная работа №2. Классификация

Лабораторная работа №2 служит для получения и закрепления навыков предобработки данных и применения методов машинного обучения для решения задач классификации.

Набор данных берется из лабораторной работы №1.

В ходе выполнения лабораторной работы должны быть выполнены следующие этапы:

1. Обучение моделей и подбор параметров с помощью Grid Search:
    a. K-ближайших соседей (KNN)
    b. Машина опорных векторов (SVM)
    c. Дерево решений ИЛИ Случайный лес
2. Оценка моделей
    a. Визуализация предсказанных значений
    b. Оценка качества прогноза (precision/recall/f1-score/ROC-AUC)
    c. Визуализация дерева решений
    d. Визуализация Feature Importance для случайного леса и XGBoost

В рамках защиты лабораторной работы необходимо продемонстрировать jupyter-notebook с кодом, быть готовым пояснить выполненные действия, продемонстрировать базовое понимание работы используемых в работе методов.

Лабораторная работа №3. Кластеризация

Лабораторная работа №3 служит для получения и закрепления навыков предобработки данных и применения методов машинного обучения для решения задач кластеризации.

Набор данных берется из лабораторной работы №1.

В ходе выполнения лабораторной работы должны быть выполнены следующие этапы:

1. Обучение моделей и подбор параметров (где применимо):
    a. метод К-средних
    b. DBSCAN
    c. Иерархическая кластеризация
2. Оценка моделей
    a. Экспертная оценка
    b. Сравнение разбиения на классы с помощью кластеризации с реальными.
    c. Визуализация предсказанных значений

В рамках защиты лабораторной работы необходимо продемонстрировать jupyter-notebook с кодом, быть готовым пояснить выполненные действия, продемонстрировать базовое понимание работы используемых в работе методов.

Лабораторная работа №4. Регрессия

Лабораторная работа №4 служит для получения и закрепления навыков предобработки данных и применения методов машинного обучения для решения задач регрессии.

Набор данных берется из лабораторной работы №1.

В ходе выполнения лабораторной работы должны быть выполнены следующие этапы:

1. Предварительная обработка данных
    a. Визуализация значимых признаков (диаграммы рассеяния, ящики с усами, гистограммы)
    b. Очистка данных (удаление пропусков, нормализация, удаление дубликатов)
2. Обучение моделей и подбор параметров (где применимо):
    a. Линейная регрессия
    b. LASSO
    c. Ридж-регрессия
3. Оценка моделей
    a. Вывод метрик
    b. Построение графиков

В рамках защиты лабораторной работы необходимо продемонстрировать jupyter-notebook с кодом, быть готовым пояснить выполненные действия, продемонстрировать базовое понимание работы используемых в работе методов.

Требования к содержанию отчетов

Каждый отчет к лабораторной работе должен содержать:

- Титульный лист
- Оглавление
- Заголовки разделов
- Нумерацию страниц
- Цель работы
- Задачи работы
- Краткую теоретическую информацию по теме работы
- Подробное описание процесса выполнения работы, скриншоты, подтверждающие выполнение шагов.
- Вывод по работе согласно цели и задачам

Для оформления применяется ГОСТ 7.32-2001 (выравнивание по ширине, нумерация страниц, подписи изображений, содержание, оформление Приложения с программным кодом и тд.)

Полезные ссылки (дополняется)

Общая информация:

- https://habr.com/ru/articles/448892/ и другие статьи на хабре по запросу "habr введение в машинное обучение"
- https://www.hse.ru/data/2017/05/14/1171296413/%D0%93%D1%80%D0%B8%D0%B3%D0%BE%D1%80%D0%B8%D0%B9%20%D0%A1%D0%B0%D0%BF%D1%83%D0%BD%D0%BE%D0%B2%20%E2%80%94%20%D0%92%D0%B2%D0%B5%D0%B4%D0%B5%D0%BD%D0%B8%D0%B5%20%D0%B2%20%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5%20%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5.pdf - презентация из ВШЭ про машинное обучение
- https://trends.rbc.ru/trends/industry/60c85c599a7947f5776ad409#card_60c85c599a7947f5776ad409_9 - статья на РБК