1. **БД для больших данных**
   a) В соответствие с вариантом, выбрать одну из БД: MongoDB, PostgreSQL;
   b) Установить на ПК и по предложенной модели данных создать БД;
   c) Заполнить БД несколькими тестовыми записями.

2. **Работа с большими данными**
   a) Поместить выданный по варианту набор данных в БД (из лаб. №1);
   b) Написать 3 запроса к БД, в соответствие с бизнес-кейсом.

3. **Анализ и интерпретация данных**

a. В соответствие с набором данных из лаб. №2 рассчитать статистические характеристики признаков данных:

1. медиана;
2. мода;
3. среднее;
4. перцентили;
5. стандартное отклонение;
6. минимальное и максимальное значения;
7. число пропущенных и уникальных значений.

b) Визуализировать признаки, например, посмотреть распределение или сезонность, того или иного ряда;

c) Провести вышеуказанные процедуры для очищенных (лаб №3) и неочищенных данных;

d) Построить практически значимые выводы.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Sklearn, Matplotlib

4. **Очистка и трансформация данных**

В соответствие с набором данных из лаб. №2 провести очистку данных:

1. удаление выбросов;
2. заполнение пропущенных значений;
3. исправление некорректных значений;
4. кодирование категориальных признаков.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Sklearn.

5. Задача классификации и её метрики качества

a) В соответствие с вариантом выбрать набор данных для задачи классификации;

b) Обучить на выбранном наборе данных несколько различных моделей:

1. логистическая регрессия;
2. наивный байесовский классификатор;
3. метод К-ближайших соседей;
4. метод опорных векторов;
5. дерево решений;

c) С помощью обученных моделей сделать прогноз для тестовых данных;

d) Оценить прогноз каждой из обученных моделей по следующим метрикам:

1. точность;
2. полнота;
3. f1-мера;

e) Сделать выводы по проделанным экспериментам;

f) Предоставить полученные в ходе обучения параметры моделей:

1. веса;
2. граф дерева решений;

g) Определить наиболее подходящую метрику для оценки качества и обосновать свой выбор.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Sklearn

6. **Задача кластеризации и её метрики качества**

a) В соответствие с вариантом выбрать набор данных для задачи кластеризации;

b) Обучить несколько различных моделей:

1. метод К-средних;
2. DBSCAN;
3. OPTICS;

c) Оценить каждую из обученных моделей по следующим метрикам:

1. Silhouette coefficient;
2. Dunn Index;
3. Davies Bouldin Index;

d) Сделать выводы по проделанным экспериментам;

e) Интерпретировать результаты метрик качества, визуализировать кластеры.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Sklearn, Matplotlib

7. **Задача восстановления регрессии и её метрики качества**

a) В соответствие с вариантом выбрать набор данных для задачи восстановления регрессии;

b) Обучить несколько различных моделей:

1. линейная регрессия;
2. ridge-регрессия (регуляризация Тихонова);
3. lasso-регрессия;

c) С помощью полученных моделей сделать прогноз для тестовых данных;

d) Оценить прогноз каждой из обученных моделей по следующим метрикам:

1. средняя абсолютная ошибка;
2. средняя квадратическая ошибка;
3. коэффициент детерминации;

e) Сделать выводы по проделанным экспериментам;

f) Интерпретировать результаты метрик качества;

g) Предоставить параметры обученных моделей.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Sklearn

8. **Композиции алгоритмов**

a) В соответствие с вариантом выбрать набор данных;

b) Построить композиции базовых алгоритмов в зависимости от типа задачи (классификация или регрессия) с использованием трех подходов:

1. boosting;
2. bagging;
3. голосование;

c) С помощью обученных моделей сделать прогноз для тестовых данных;

d) Оценить прогнозы по метрикам из лаб. №6 или лаб. №7;

e) Сравнить приведенные методы построения композиций алгоритмов и сделать выводы. Уметь объяснить базовое устройство каждого из них.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Sklearn

9. **Рекомендательные системы**

a) В соответствии с вариантом выбрать набор данных для построения рекомендательной системы;

b) Построить рекомендательную систему, используя один из подходов:

1. подход, на основе анализа контента;
2. коллаборативная фильтрация;

c) Составить рекомендации для 10 случайных пользователей из набора данных;

d) Сформулировать практически значимые выводы по результатам.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Sklearn

10. Нейронные сети: перцептрон с несколькими слоями

a) В соответствие с вариантом выбрать набор данных;

b) Построить и обучить перцептрон с несколькими слоями;

c) Путем оценки модели по соответствующим типу задачи (классификация или регрессия) метрикам подобрать наилучшие гиперпараметры модели:

1. число эпох;
2. число слоёв;
3. learning rate;
4. batch-size;
5. число нейронов в слое;

d) Отразить результаты обучения в таблице;

e) Привести график снижения ошибки в ходе обучения.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Keras, Matplotlib

11. Свёрточные нейронные сети

a) В соответствие с вариантом выбрать набор данных;

b) Построить и обучить свёрточную нейронную сеть;

c) Путем оценки построенной модели по метрикам классификации подобрать наилучшие гиперпараметры:

1. число эпох;
2. размер свёрточного фильтра;
3. learning rate;
4. batch-size;
5. размер свёрточного слоя;

d) Отразить результаты обучения в таблице;

e) Вывести график снижения ошибки в ходе обучения.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Keras, Matplotlib

12. Рекуррентные нейронные сети

a) В соответствие с вариантом выбрать набор данных;

b) построить и обучить рекуррентную нейронную сеть.

c) Путем оценки модели по метрикам классификации подобрать наилучшие гиперпараметры:

1. число эпох
2. размер рекуррентного слоя
3. learning rate
4. batch-size
5. число рекуррентных слоёв.

d) Отразить результаты обучения в таблице

e) вывести график снижения ошибки в ходе обучения

f) Внутри группы оценить применимость сверточных и рекуррентных нейронных сетей для задач классификации текста и изображений.

Для реализации использовать язык программирования Python 3.x и библиотеки Pandas, Numpy, Keras, Matplotlib

Наборы данных для лабораторных 2, 3, 4 (работа с данными)

№	Датасет	Запросы
1	Данные по штрафам за парковку в Нью-Йорке https://www.kaggle.com/new-york-city/nyc-parking-tickets	1. Число штрафов сгруппированные по штатам; 2. Наиболее частый тип кузова, получающий штраф; 3. Среднее число штрафов в день по штату.
2	Данные о заболеваемости COVID-19 https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset	1. Число смертей за 1-й месяц наблюдения по странам; 2. Наиболее “заражаемые” штаты США; 3. Число заражений по дням за последние 30 дней наблюдения.
3	Данные о скачиваниях и рейтингах приложений в Google play https://www.kaggle.com/lava18/google-play-store-apps	1. Список категорий приложений с наиболее высоким рейтингом (средним); 2. Максимальное число отзывов для платных и бесплатных приложений; 3. Наиболее популярный жанр для приложений дороже 5 долларов.
4	Данные о статистике суицидов по странам с 1985 по 2016 годы https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016	1. Самые частотные страны по суициду из ТОП10 низкого ВВП; 2. Среднее число суицидов для мужчин и женщин в год вашего рождения; 3. Самые частотные возрастные категории по суициду.
5	Данные по БУ авто с Craiglist https://www.kaggle.com/austinreese/craigslist-carstrucks-data	1. Средняя цена авто по марке производителя 2. Список наиболее дешевых марок для 6-цилиндровых авто 3. Число авто дешевле 5000$ по годам выпуска

Наборы данных для лабораторных 5, 6, 8, 10 (классификация и кластеризация)

№	Датасет	Описание
1	Walmart Recruiting: Trip Type Classification https://www.kaggle.com/c/walmart-recruiting-trip-type-classification/data	Данные о посещениях магазина Walmart покупателями. Целевой признак: тип посещения магазина.
2	IEEE-CIS Fraud Detection https://www.kaggle.com/c/ieee-fraud-detection/data	Данные об онлайн-транзакциях. Целевой признак: является ли транзакция мошеннической
3	Home Credit Default Risk https://www.kaggle.com/c/home-credit-default-risk/data	Данные о заемщиках банка. Целевой признак: способен ли заёмщик выплатить кредит

Наборы данных для лабораторных 7, 8, 10 (регрессия)

№	Датасет	Описание
1	House Prices: Advanced Regression Techniques https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data	Данные о продаваемых домах и их характеристиках. Целевой признак: цена дома
2	Restaurant Revenue Prediction https://www.kaggle.com/c/restaurant-revenue-prediction/data	Данные о ресторанах, их местоположении, типе и т.д. Целевой признак: выручка ресторана за год
3	Sberbank Russian Housing Market https://www.kaggle.com/c/sberbank-russian-housing-market/data	Данные о купленной недвижимости в России и макроэкономике России. Целевой признак: стоимость недвижимости

Наборы данных для лабораторной 9 (рекомендательные системы)

№	Датасет	Описание
1	Expedia Hotel Recommendations https://www.kaggle.com/c/expedia-hotel-recommendations/data	Данные о пользователях сайта Expedia. Необходимо “рекомендовать” тип отеля, наиболее подходящий для пользователя.
2	Airbnb New User Bookings https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings	Данные о пользователях сайта Airbnb. Необходимо “рекомендовать” страну, в которой пользователь сделает свою первую бронь.

Наборы данных для лабораторных 11, 12

№	Датасет	Описание
1	MNIST database https://www.kaggle.com/c/digit-recognizer/data	Классификация цифр на изображениях
2	IMDB movie Reviews https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews	Классификация эмоциональной окраски текстового отзыва фильма