# БД для больших данных

1. В соответствие с вариантом, выбрать одну из БД: MongoDB, PostgreSQL;
2. Установить на ПК и по предложенной модели данных создать БД;
3. Заполнить БД несколькими тестовыми записями.

# Работа с большими данными

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

# Анализ и интерпретация данных

* 1. 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.
8. Визуализировать признаки, например, посмотреть распределение или сезонность, того или иного ряда;
9. Провести вышеуказанные процедуры для очищенных (лаб №3) и неочищенных данных;
10. Построить практически значимые выводы.

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

# 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.

# Задача классификации и её метрики качества

1. В соответствие с вариантом выбрать набор данных для задачи классификации;
2. Обучить на выбранном наборе данных несколько различных моделей:
3. логистическая регрессия;
4. наивный байесовский классификатор;
5. метод К-ближайших соседей;
6. метод опорных векторов;
7. дерево решений;
8. С помощью обученных моделей сделать прогноз для тестовых данных;
9. Оценить прогноз каждой из обученных моделей по следующим метрикам:
10. точность;
11. полнота;
12. f1-мера;
13. Сделать выводы по проделанным экспериментам;
14. Предоставить полученные в ходе обучения параметры моделей:
15. веса;
16. граф дерева решений;
17. Определить наиболее подходящую метрику для оценки качества и обосновать свой выбор.

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

# Задача кластеризации и её метрики качества

1. В соответствие с вариантом выбрать набор данных для задачи кластеризации;
2. Обучить несколько различных моделей:
3. метод К-средних;
4. DBSCAN;
5. OPTICS;
6. Оценить каждую из обученных моделей по следующим метрикам:
7. Silhouette coefficient;
8. Dunn Index;
9. Davies Bouldin Index;
10. Сделать выводы по проделанным экспериментам;
11. Интерпретировать результаты метрик качества, визуализировать кластеры.

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

# Задача восстановления регрессии и её метрики качества

1. В соответствие с вариантом выбрать набор данных для задачи восстановления регрессии;
2. Обучить несколько различных моделей:
3. линейная регрессия;
4. ridge-регрессия (регуляризация Тихонова);
5. lasso-регрессия;
6. С помощью полученных моделей сделать прогноз для тестовых данных;
7. Оценить прогноз каждой из обученных моделей по следующим метрикам:
8. средняя абсолютная ошибка;
9. средняя квадратическая ошибка;
10. коэффициент детерминации;
11. Сделать выводы по проделанным экспериментам;
12. Интерпретировать результаты метрик качества;
13. Предоставить параметры обученных моделей.

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

# Композиции алгоритмов

1. В соответствие с вариантом выбрать набор данных;
2. Построить композиции базовых алгоритмов в зависимости от типа задачи (классификация или регрессия) с использованием трех подходов:
3. boosting;
4. bagging;
5. голосование;
6. С помощью обученных моделей сделать прогноз для тестовых данных;
7. Оценить прогнозы по метрикам из лаб. №6 или лаб. №7;
8. 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

# Рекомендательные системы

1. В соответствии с вариантом выбрать набор данных для построения рекомендательной системы;
2. Построить рекомендательную систему, используя один из подходов:
3. подход, на основе анализа контента;
4. коллаборативная фильтрация;
5. Составить рекомендации для 10 случайных пользователей из набора данных;
6. Сформулировать практически значимые выводы по результатам.

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

# Нейронные сети: перцептрон с несколькими слоями

1. В соответствие с вариантом выбрать набор данных;
2. Построить и обучить перцептрон с несколькими слоями;
3. Путем оценки модели по соответствующим типу задачи (классификация или регрессия) метрикам подобрать наилучшие гиперпараметры модели:
4. число эпох;
5. число слоёв;
6. learning rate;
7. batch-size;
8. число нейронов в слое;
9. Отразить результаты обучения в таблице;
10. Привести график снижения ошибки в ходе обучения.

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

# Свёрточные нейронные сети

1. В соответствие с вариантом выбрать набор данных;
2. Построить и обучить свёрточную нейронную сеть;
3. Путем оценки построенной модели по метрикам классификации подобрать наилучшие гиперпараметры:
4. число эпох;
5. размер свёрточного фильтра;
6. learning rate;
7. batch-size;
8. размер свёрточного слоя;
9. Отразить результаты обучения в таблице;
10. Вывести график снижения ошибки в ходе обучения.

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

# Рекуррентные нейронные сети

1. В соответствие с вариантом выбрать набор данных;
2. построить и обучить рекуррентную нейронную сеть.
3. Путем оценки модели по метрикам классификации подобрать наилучшие гиперпараметры:
4. число эпох
5. размер рекуррентного слоя
6. learning rate
7. batch-size
8. число рекуррентных слоёв.
9. Отразить результаты обучения в таблице
10. вывести график снижения ошибки в ходе обучения
11. 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  [h ttps://www.kaggle.com/new-york-city/n](https://www.kaggle.com/new-york-city/nyc-parking-tickets) [y c-parking-tickets](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  [h ttps://www.kaggle.com/sudalairajkumar](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_deaths.csv)  [/ novel-corona-virus-2019-dataset](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_deaths.csv) | 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  [h ttps://www.kaggle.com/lava18/google-](https://www.kaggle.com/lava18/google-play-store-apps) [p lay-store-apps](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  [h ttps://www.kaggle.com/russellyates88/](https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016) [s uicide-rates-overview-1985-to-2016](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  [h ttps://www.kaggle.com/austinreese/crai](https://www.kaggle.com/austinreese/craigslist-carstrucks-data) [g slist-carstrucks-data](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  h [ttps://www.kaggle.com/c/walmart-recruiting-](https://www.kaggle.com/c/walmart-recruiting-trip-type-classification/data) t [rip-type-classification/data](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  h [ttps://www.kaggle.com/c/ieee-fraud-detectio](https://www.kaggle.com/c/ieee-fraud-detection/data) n [/data](https://www.kaggle.com/c/ieee-fraud-detection/data) | Данные об онлайн-транзакциях. Целевой признак: является ли транзакция мошеннической |
| 3 | Home Credit Default Risk  h [ttps://www.kaggle.com/c/home-credit-defaul](https://www.kaggle.com/c/home-credit-default-risk/data) t [-risk/data](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  h [ttps://www.kaggle.com/c/house-prices](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data)  - [advanced-regression-techniques/data](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  h [ttps://www.kaggle.com/c/restaurant-re](https://www.kaggle.com/c/restaurant-revenue-prediction/data) v [enue-prediction/data](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  h [ttps://www.kaggle.com/c/sberbank-rus](https://www.kaggle.com/c/sberbank-russian-housing-market/data) s [ian-housing-market/data](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  h [ttps://www.kaggle.com/c/expedia-hot](https://www.kaggle.com/c/expedia-hotel-recommendations/data) e [l-recommendations/data](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  h [ttps://www.kaggle.com/c/airbnb-recr](https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings) u [iting-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  [h ttps://www.kaggle.com/c/digit-recognizer/data](https://www.kaggle.com/c/digit-recognizer/data) | Классификация цифр на изображениях |
| 2 | IMDB movie Reviews  [h ttps://www.kaggle.com/lakshmi25npathi/imdb-dataset](https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews)  [- of-50k-movie-reviews](https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews) | Классификация эмоциональной окраски текстового отзыва фильма |