Вы должны выбрать датасет для решения задачи классификации или регрессии. Рекомендую брать понравившийся датасет с kaggle. Например,

https://www.kaggle.com/purumalgi/music-genre-classification
https://www.kaggle.com/sagnik1511/car-insurance-data
https://www.kaggle.com/fedesoriano/heart-failure-prediction/version/1
https://www.kaggle.com/teertha/personal-loan-modeling
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
https://www.kaggle.com/rashikrahmanpritom/177k-english-song-data-from-20082017
https://www.kaggle.com/shivan118/hranalysis?select=train.csv
https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction?select=fake_job_postings.csv
https://www.kaggle.com/naveengowda16/logistic-regression-heart-disease-prediction
https://www.kaggle.com/kaushiksuresh147/customer-segmentation?select=Train.csv
https://www.kaggle.com/jsphyg/weather-dataset-rattle-package
https://www.kaggle.com/sobhanmoosavi/us-accidents/version/10
https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction
https://www.kaggle.com/crowdflower/twitter-user-gender-classification
https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey
https://www.kaggle.com/mssmartypants/water-quality
https://www.kaggle.com/code/gauravduttakiit/covid-19-sentiment-analysis-on-train-data/
https://www.kaggle.com/datasets/raghadalharbi/all-products-available-on-sephora-website

Рекомендуется брать датасет с несколькими тысячами-десятками тысяч объектов и несколькими десятками признаков.

Задачу надо решать в Jupyter notebook и выложить на гит.

1. Загрузите данные
2. Опишите задачу словами. В том числе напишите, что значит каждый признак
3. Разбейте данные на обучающую и тестовую выборки
4. Визуализируйте данные из обучающей выборки. В частности, имеет смысл построить диаграммы рассеивания для количественных признаков. Построить гистограммы распределений и т.п. Вычислить основные характеристики (среднее, разброс, корреляционную матрицу и т.д.). Интерпретируйте результаты
5. Обработать пропущенные значения (или убедиться, что их нет)
6. Исключить нерелевантные признаки (объяснить, как вы их нашли)
7. Если необходимо, то обработать коррелированные признаки
8. Обработать категориальные признаки
9. Провести масштабирование (или объяснить, почему в вашем случае она не нужна)
10. Вам может понадобиться другая предобработка. Например, если в вашем датасете есть текстовые признаки с уникальными значениями (например, аннотации товаров, отзывы пользователей, другие тексты), как в двух последних датасетах из перечисленных, то вам понадобится этап извлечения признаков, т.е. простые методы NLP, как, например, bag-of-words. Воспользуйтесь библиотеками re, nltk
11. После шагов 5–10 разумно вернутся к шагу 4 (а может, возвращаться к нему после каждого из этапов 5–10.

12. Попробуйте как минимум 3 метода классификации (регрессии). Объясните ваш выбор. Найдите значения метрик на обучающей и тестовой выборке. Сделайте вывод.
13. На одном из методов (объясните выбор) найдите оптимальное значение параметров. Постройте график зависимости ошибок (на обучающей выборке и валидационной/CV) от значения гиперпарамтра. Для найденного оптимального значения параметра (параметров) снова обучите модель. Сделайте вывод.
14. Довольны ли вы результатами? В частности, если классы не сбалансированы, то результат может оказаться неприемлемым. В этом случае можете применить методы балансировки из библиотеки imbalanced-learn.
15. Сделать общие выводы

Вы можете пользоваться готовыми примерами решения подобных задач. Например, здесь есть основные шаги, правда мало комментариев:
https://github.com/NikolaiZolotykh/MachineLearningCourse/blob/master/VADII_06_%D0%9E%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D1%81%D1%82%D0%BE%D0%B8%D0%BC%D0%BE%D1%81%D1%82%D0%B8_%D0%BA%D0%B2%D0%B0%D1%80%D1%82%D0%B8%D1%80%D1%8B.ipynb
Обратите внимание, что это задача восстановления регрессии - большинство же из вас выберет классификацию