Домашнее задание
по теме «Алгоритм Random Forest»

Формулировка задания

Выполните подготовку данных для решения задач классификации и регрессии.

Проведите классификацию методом Случайного леса. Обратите внимание на подбор гиперпараметров алгоритма. Качество оценить минимум по 3 критериям качества для классификации: confusion_matrix, accuracy, precision, recall, f1_score, roc_auc.

Проведите регрессию методом Случайного леса. Обратите внимание на подбор гиперпараметров алгоритма. Качество оценить минимум по 3 критериям качества для регрессии: MAE, MSE, RMSE, MAPE, RMSLE, R^2 и др.

Для классификации, регрессии и оценки качества использовать библиотеку scikit-learn.

Результирующий код должен быть читаемым, с единой системой отступов и адекватными названиями переменных.

Описание плана работы

Задача 1. Классификация

1) Загрузите данные из дополнительных материалов или по ссылке: https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals?select=train_dataset.csv из заданий “Классификация SVM” и “Классификация Decision Tree”.

Если на наборе данных задача решается долго, то провести исследование на части данных. Использовать случайное сэмплирование или другие методы.

2) EDA(Exploratory Data Analysis) и подготовку данных использовать из прошлых заданий. Нормализовывать данные не нужно.

3) Обучите алгоритм RandomForestClassifier (метод случайного леса из библиотеки scikit-learn). Посчитайте качество классификации и напишите ответы на следующие вопросы:

a) Какие значения гиперпараметров алгоритма подойдут для задачи?

b) Насколько ваш алгоритм верно предсказывает целевую переменную?
c) Какие критерии качества классификации получились для задачи?

Задача 2. Регрессия

1) Загрузите данные как в задаче “Регрессия SVM” и “Регрессия Decision Tree” из дополнительных материалов или по ссылке: https://www.kaggle.com/competitions/playground-series-s4e12

Если на наборе данных задача решается долго, то провести исследование на части данных. Использовать случайное сэмплирование или другие методы.

2) EDA (Exploratory Data Analysis) используем прошлых домашних заданий по регресии. Нормализовывать данные не нужно.

4) Обучить модель регрессии RandomForestRegressor.

5) Оценить качество алгоритма. Сравнить с прошлыми решениями. Получилась ли модель лучшего качества?

6) Подобрать гиперпараметры RandomForestRegressor через GridSearchCV или другой метод подбора гиперпараметров. Какие гиперпараметры будут наиболее подходящими? Как изменилось качество модели?

Перечень необходимых инструментов

- Python
- scikit-learn
- pandas
- venv
- Jupiter Notebook
- IDE VS Code
- GigaIDE

Форма предоставления результата

1. В поле ссылки загрузить ссылку на удаленный репозиторий с доступом для наставника.
2. В поле файла загрузить архив с папкой, в которой разместить отчет со скриншотами по заданию и решение задачи. Решение должно быть представлено в формате .ipynb или .py.

Шкала оценивания

- 1.0 – отлично
- 0.7–0.9 – хорошо
- 0.5–0.6 – удовлетворительно
- Менее 0.5 – задание не выполнено