Курсовой проект по дисциплине
«Анализ данных и машинное обучение»

Исходные данные.

Исходные данные для выполнения курсового проекта представляют собой набор данных в формате csv. Набор данных содержит определенный набор признаков и целевой признак.

Задание на курсовой проект.

1. Визуальный анализ данных.

С использованием полученного набора данных необходимо провести визуальный анализ данных. Всего должно быть построено не менее 15 различных графиков. Обязательно должны быть использованы графики:

- Stackplot

С использованием pandas plot графики типов:

- pie;
- density;

-bar.

С использованием Seaborn:

- boxplot
- countplot
- distplot
- pairplot

Возможно использовать также любые графики из любых доступных библиотек. Ссылка на хорошую статью с применением графиков: https://habr.com/ru/post/468295/

2. Построение моделей машинного обучения.

Используя полученный набор данных необходимо построить модели машинного обучения. Целевой признак отдельно указан для каждого набора данных. В качестве набора для обучения выбирается 80% данных, в качестве тестовой выборки 20% данных. Параметр random_state при разбиении указывается 15.

Для полученного набора данных требуется построить следующие модели машинного обучения:

- дерево решений;
- метод ближайших соседей;
- логистическая регрессия;
- случайный лес;
- градиентный бустинг.

Обучение моделей провести в 3 этапа:

1. С минимальной подготовкой данных. Выполнить только преобразование значений только в случае если это требуется для обучения модели.

2. Выполнить чистку и предподоготовку данных. Удалить «выбросы», заполнить недостающие значения и т.д. Удалить не более 5% данных.

3. Выполнить подбор гиперпараметров модели.

Для каждого из этапов записать результат работы модели на тестовой выборке используя метрики:accuracy, precision, recall, roc auc.

Jupyter Notebook с полным кодом анализа данных и построения моделей прилагается к курсовому проекту.

Критерии оценивания курсового проекта.

Отлично. Выполнены все задачи курсового проектирования. Пояснительная записка написана грамотным языком и отражает все этапы выполнения курсового проекта.

Хорошо. Построено 15 графиков, но использовано только 5 из 8 обязательных. Применено 4 из 5 обязательных моделей машинного обучения, но выполнены все этапы обучения.

Удовлетворительно. Построено 10 графиков, но использовано только 3 из 8 обязательных. Применено 3 из 5 обязательных моделей машинного обучения, выполнено 2 из 3 этапов обучения.

Варианты для выполнения курсового проекта.

Вариант для выполнения курсового проекта выбирается в соответствии с порядковым номером в списке группы.

№	Целевой признак	DataSet
1	income	https://www.kaggle.com/code/riccardobollati01/85-accuracy-ensemble-learning-model-data-viz/data
2	hospital_death	https://www.kaggle.com/datasets/mitishaagarwal/patient
3	class	https://www.kaggle.com/datasets/kukuroo3/body-performance-data
4	Target	https://www.kaggle.com/datasets/krantiswalke/bankfullcsv
5	Star type	https://www.kaggle.com/datasets/deepu1109/star-dataset
6	is_safe	https://www.kaggle.com/datasets/mssmartypants/water-quality
7	depressed	https://www.kaggle.com/datasets/diegobabativa/depression
8	specie	https://www.kaggle.com/datasets/bertiemackie/sloth-species
9	Survived	https://www.kaggle.com/datasets/yasserh/titanic-dataset
10	label	https://www.kaggle.com/datasets/praveengovi/credit-risk-classification-dataset
11	TravelInsurance	https://www.kaggle.com/datasets/tejashvi14/travel-insurance-prediction-data
12	satisfied	https://www.kaggle.com/datasets/mohamedharris/employee-satisfaction-index-dataset
13	final_payment_status	https://www.kaggle.com/datasets/vizdom/customer-orders
14	status	https://www.kaggle.com/datasets/debasisdotcom/parkinson-disease-detection
15	status	https://www.kaggle.com/datasets/qusaybtoush1990/banks-loan
16	target	https://www.kaggle.com/datasets/yasserh/heart-disease-dataset
17	Class	https://www.kaggle.com/datasets/mssmartypants/rice-type-classification
18	rating	https://www.kaggle.com/datasets/qusaybtoush1990/the-cars
19	Your level of satisfaction in Online Education	https://www.kaggle.com/datasets/sujaradha/online-education-system-review
20	BAD	https://www.kaggle.com/datasets/ajay1735/hmeq-data