Лабораторная работа №3 — Деревья решений

Дано
датасет DT_titanic.csv, содержащий следующие данные

survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Gender	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Таргет переменная — survival. Задача состоит в том, чтобы понять, выживет ли с Титаника пассажир, имеющий определенный набор вводных.

Задание
1) Проанализировать распределение и взаимосвязь (визуальную) фичей. Есть ли выбросы, и как с ними поступить?
2) Привести категориальные фичи к численному виду посредством OneHotEncoder.
3) Разбить датасет в соотношении 80% - тренировочный, 20% - тестовый.
4) На тестовой выборке натренировать дерево решений [DecisionTreeClassifier], подобрав оптимальные параметры дерева. Обосновать (в комментарии) выбор данных параметров.
5) Рассчитать ключевые характеристики классификации на тренировочных и на тестовых данных.
6) Вывести фичи в порядке их важности в полученной модели.
7) Повторить пункты [4]-[6] для случайного леса [RandomForestClassifier].

Система оценивания
Задание будет оценено по 5-балльной шкале.
За позднюю сдачу задания будет снят 1 балл.

Формат выполнения
.ipynb файл, без ошибок прогоняемый сверху вниз с любого компьютера
Файл выслать преподавателю посредством мессенджера Telegram : @evgeny_zavalnyuk

Срок выполнения [время фиксируется исходя из отправки решения преподавателю]
21 апреля 2024г. до 23:59:59 МСК