**Практическое занятие № 2**

**«Решение задачи классификации.**

**Прогнозирование выживаемости пассажиров титаника»**

1. Для выполнения задания использовать данные – <https://www.kaggle.com/competitions/titanic/data> .
2. Загрузить и исследовать данные. Вывести размерность данных, основные описательные статистики, типы данных для признаков.
3. Осуществить проверку данных на наличие пропущенных значений. Если пропущенные значения есть – выбрать стратегию работы с пропусками для каждого отдельного признака и избавиться от них.
4. Удалить из данных признаки: **PassengerId**, **Name**, **Ticket**.
5. На основе признаков **SibSp** и **Parch** создать категориальный признак **TravelAlone**, который определяет путешествовал ли человек в одиночку или нет.
6. Используя функцию **value\_counts()**, посчитать количество человек каждого пола. С помощью **countplot** из **seaborn** визуализировать эти данные.
7. С помощью **barplot** из **seaborn** визуализировать влияние признака пол на выживаемость пассажиров. Построить гистограмму количества погибших и выживших пассажиров среди мужчин и женщин (использовать **countplot** из **seaborn**).
8. Посчитать количество пассажиров в зависимости от признака **Pclass**. Построить гистограмму по этим данным (**countplot**).
9. Исследовать влияние признака **Pclass** на выживаемость, визуализировать данные на основе **barplot**. Построить гистограмму количества погибших и выживших пассажиров, сгруппировав данные по признаку **Pclass** (использовать **countplot** из **seaborn**).
10. Исследовать выживаемость пассажиров в зависимости от возраста.
11. Построить **lmplot** из библиотеки **seaborn**, указав следующие аргумены (x='Age',y='Survived'). Сделать вывод по полученной визуализации.
12. Вывести количество пассажиров по признаку «**Emabarked**», визуализировать с помощью **countplot**.
13. Исследовать влияние признака «**Embarked**» на выживаемость, визуализировать с помощью **barplot**. Построить гистограмму количества погибших и выживших пассажиров, сгруппировав данные по признаку «**Embarked**» (использовать **countplot** из **seaborn**).
14. Вывести количество пассажиров на основе признака «**TravelAlone**», визуализировать данные с помощью **countplot**.
15. Исследовать влияние признака «**TravelAlone**» на выживаемость, визуализировать с помощью с помощью **barplot**.
16. По данным построить матрицу корреляций. Сделать выводы на основании полученных значений.
17. Для категориальных признаков «**Sex**» и «**Embarked**» выполнить **Label**-кодирование.
18. Разделить данные на обучающую и тестовую выборки.
19. Построить различные модели для прогноза выживаемости пассажира (не менее 3 моделей). Выбрать метрику и сравнить по ней построенные модели.
20. Для каждой модели построить матрицу ошибок (confusion\_matrix) и classification\_report по тестовой выборке.
21. Для лучшей модели подобрать оптимальные значения параметров с помощью поиска по сетке (**GridSearchCV**).
22. На основе лучшей модели с оптимальными параметрами сделать прогнозы на новых данных (файл **test.csv**).