## Контрольная работа по дисциплине “Технологии анализа данных и машинное обучение”

### Цель работы

Продемонстрировать владение основными навыками работы с методами машинного обучения с учителем и без учителя, владение основными инструментальными средствами библиотек языка программирования Python, методами и приемами подготовительного и описательного анализа данных, средствами визуализации данных, использования и усовершенствования обучаемых моделей, умение делать выводы из проведенного анализа.

### Задания для выполнения

1. Выбрать задачу (классификация, кластеризация, регрессия). Выбрать набор данных для анализа. Описать этот набор и решаемую задачу.
2. Провести предварительный анализ и очистку данных. Этот этап включает в себя вывод информации о количественных характеристиках датасета, информацию об отсутствующих значениях, характеристиках и физическом смысле каждого атрибута данных, его значимости для предсказания целевой переменной, вывод нескольких точек данных для иллюстрации структуры данных.
3. При необходимости преобразовать атрибуты исходного датасета в числовые признаки. Этот этап сильно зависит от типа исследуемых данных и может включать в себя векторизацию текста, извлечение признаков их аудио и видео данных, преобразование изображений в плоский численный массив и другие преобразования.
4. Провести описательный анализ данных. Сделать выводы. Этот этап включает в себя определение шкалы измерения каждого признака, выявление аномальных значений, визуализацию распределения каждого признака, при необходимости - проверка на нормальность, построение кореллограмм и совместных распределений каждого признака с целевой переменной, выявление коррелированных признаков и признаков, не несущих информации для данной задачи.
5. Разделить набор данных на обучающую и тестовую выборки. Обосновать количественные характеристики и метод разделения (временной, случайный, последовательный).
6. Обучить несколько моделей для решения выбранной задачи (для задач классификации - не менее 7 различных алгоритмов). Проанализировать результаты, сделать выводы.
7. Выбрать наиболее перспективную модель для решения поставленной задачи. Изменить гиперпараметры модели. Найти оптимальные гиперпараметры.
8. С учетом сделанных выводов провести усовершенствование моделей. Это можно осуществить с помощью введения регуляризации, изменение параметров модели (для параметрических моделей), введением суррогатных признаков, отбором признаков, нормализацией данных, изменением алгоритма предварительной обработки данных. Сравнить результаты.
9. Попробовать изменить порядок предобработки данных для повышения эффективности модели. Попробовать применить понижение размерности для создания суррогатных признаков. Сравнить результаты, сделать выводы.
10. Представить результаты моделирования в наглядном виде (графики, линии обучения, таблицы сравнения моделей, таблицы классификации, и другие). Сделать выводы, сравнить с существующими аналогичными решениями, порассуждать о перспективах решения проблемы.

### Методические указания

1. Работа выполняется в виде программного ноутбука Python Jupyter. Пояснительная записка выполняется в виде текстового документа и должна включать в себя: текстовое описание проблемы, по необходимости пример кода для каждого этапа работы, текстовые выводы по каждому этапу и сформулированное заключение с результатами работы и их интерпретация.
2. Все пояснения, выводы и замечания, на которые необходимо обратить внимание должны присутствовать в работе в виде ячеек документации либо (менее предпочтительно) программных комментариев.
3. Работа должна выполняться студентом самостоятельно и индивидуально.
4. Оценка качества моделирования должна производиться с использованием определенных метрик. Их выбор должен быть описан и обоснован до начала моделирования. Плюсом работы является широкий набор метрик эффективности моделей.

### Критерии оценки

1. Структурированность отчета. В работе должна прослеживаться четкая структура - подготовительный этап, анализ данных, построение простых моделей, сравнение и анализ моделей, выводы, построение моделей с учетом выводов, итоговый результат.
2. Наличие выводов. Работа должна содержать текстовые замечания, поясняющие каждый шаг работы студента: что делается, зачем и какую информацию это нам дает. Оценивается полнота и адекватность выводов.
3. (рекомендуется) Замеры времени. В целях анализа временной сложности алгоритмов. Все инструкции, запускающие цикл обучения модели должны содержать замер времени обучения. Замер можно производить с помощью магических инструкций Jupyter или (более предпочтительно) с использованием стандартной библиотеки Python. Сравнение моделей должно учитывать и время обучения.
4. Визуализация. Работа должна демонстрировать навыки студента визуализировать информацию. Особенно на этапах описательного анализа и анализа обучаемости модели. Оценивается разнообразие, наглядность и информативность визуализации.
5. (в зависимости от темы) Разнообразие моделей. Студент должен продемонстрировать умение работать с разнообразными моделями обучения, применимыми к одной задаче. Например, в задачах классификации существует как минимум десять наиболее применимых моделей. Оценивается число алгоритмов, примененных студентом для одной и той же задачи.
6. (в зависимости от темы) Улучшение модели. Студент должен продемонстрировать умение анализировать обученную модель и искать пути для ее совершенствования. Оценивается количество итераций совершенствования модели и их эффективность.
7. Предобработка данных. Работа должна содержать исчерпывающий алгоритм предварительной обработки данных. Он служит для того, чтобы исправить все несовершенства в данных и сделать набор данных как можно более пригодным для машинного обучения. Оценивается сложность и воспроизводимость процедуры предварительной обработки данных.
8. Использование метрик эффективности. Оценивается разнообразие и адекватность задаче примененных метрик эффективности (включая время обучения) а также полнота сравнения и правильность выводов из сравнения моделей по разным метрикам.
9. Валидность результатов. Студент должен продемонстрировать умение оценивать достоверность измерения метрик моделей и повышать ее.