Московский государственный технический университет им. Н.Э. Баумана

Факультет «Головной учебно-исследовательский и методический центр

профессиональной реабилитации лиц с ограниченными возможностями здоровья (инвалидов)»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа 2**

**по дисциплине «Методы машинного обучения в АСОИУ»**

**"** **Обработка признаков (часть 1) "**

СТУДЕНТ:

студент группы ИУ5Ц-21М

Москалик А.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

Москва, 2024

# Цель лабораторной работы: изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

# Задание:

# Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.) Просьба не использовать датасет, на котором данная задача решалась в лекции.

# Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:

# А. устранение пропусков в данных;

# Б. кодирование категориальных признаков;

# В. нормализация числовых признаков.

# Ход работы

# Подготовка

# Подключение библиотек и загрузка датасета

# 

# Здесь представлена таблица датасета, который, относится к оценке учебной производительности студентов.

# Представленные данные:

# -gender: Пол ученика — женский (female) или мужской (male).

# -race\_ethnicity: Этническая принадлежность учеников, классифицированная по группам от A до C (возможно, существуют и другие группы в полном датасете).

# -parental\_level\_of\_education: Уровень образования родителей, который варьируется от "some college" (некоторое количество колледжного образования), до "bachelor's degree" (степень бакалавра) и "master's degree" (степень магистра).

# -lunch: Тип питания, которое предоставляется ученику в школе — "standard" (стандартное) или "free/reduced" (бесплатное/со скидкой).

# -test\_preparation\_course: Указывает на то, закончил ли ученик курс подготовки к тестам ("completed") или нет ("none").

# -math\_score: Оценка ученика по математике.

# -reading\_score: Оценка ученика по чтению.

# -writing\_score: Оценка ученика по письму. Эти данные могут использоваться для анализа связи между социально-экономическими факторами (такими как этническая принадлежность, образование родителей и тип питания) и учебной производительностью студентов, а также влиянием курса подготовки на оценки.

# Проверка на наличие пропусков

# 

# В каждом из столбцов датасета отсутствуют пропуски, поскольку количество пропущенных значений для каждого признака равно 0.

# Т.к. для ЛР2 мы работаем с пропусками, следовательно создадим код, который создаст пропуски в датасете.

# Обработка пропусков

# 

# Для числовых признаков (например, math\_score, reading\_score, writing\_score), обычно пропуски заполняют средним или медианным значением столбца:

# 

# 

# Math Score (Оценки по математике):Если гистограмма показала пик в районе среднего значения, это говорит о том, что значительное количество пропущенных значений было заменено на среднее, что уменьшает разнообразие исходного распределения оценок. Среднее значение до и после замены остается неизменным, но стандартное отклонение может уменьшиться, что указывает на снижение изменчивости оценок.

# Reading Score (Оценки по чтению):Аналогично, гистограмма могла показать увеличение частоты в районе среднего значения, что также свидетельствует о замене пропущенных данных средними значениями. Распределение оценок после замены может выглядеть менее "естественно", особенно если пропущенных значений было много.

# Writing Score (Оценки по письму):Как и в предыдущих случаях, замена пропусков средними значениями создает пик на уровне среднего значения. Это может быть проблематично при анализе распределения, так как искусственно создает более однородный набор данных. Меры центральной тенденции, такие как среднее и медиана, не изменяются, но меры разброса, включая дисперсию и стандартное отклонение, могут быть искажены из-за уменьшения разнообразия.

# Во всех трех случаях, важно помнить, что замена пропусков средними значениями может исказить истинное распределение оценок, если пропусков было много. Это также может повлиять на результаты статистического анализа, включая корреляционный и регрессионный анализы, так как данные становятся менее вариативными и менее представительными.

# Для категориальных признаков (например, gender, race\_ethnicity, parental\_level\_of\_education, lunch, test\_preparation\_course), часто используют моду (наиболее часто встречающееся значение):

# 

# 

# 

# В этом коде мы сначала заполняем пропущенные значения в каждом категориальном признаке самым частым значением (модой) для этой колонки. Затем для каждой колонки создается столбчатая диаграмма, показывающая количество наблюдений для каждой категории.

# 

# Результат, который получили после использования data.isnull().sum(), показывает, что в каждом столбце датасета теперь 0 пропусков. Это означает, что успешно обработали все пропуски в данных. Теперь датасет полностью заполнен, и можно перейти к следующим шагам обработки данных, таким как кодирование категориальных признаков и нормализация числовых признаков.

# Кодирование категориальных признаков

# Категориальные переменные обычно нужно преобразовать в числовые значения, чтобы их можно было использовать в машинном обучении. Вот два распространённых метода:

# One-Hot Encoding — создаёт новые столбцы для каждого уникального значения в категориальной переменной, где 1 означает наличие этого значения и 0 его отсутствие.

# Label Encoding — присваивает каждому уникальному значению в категориальной переменной целочисленный идентификатор. Выбор метода зависит от модели и данных. Например, для линейных моделей предпочтительнее использовать One-Hot Encoding, так как Label Encoding может ввести ложное понятие порядка, которого на самом деле нет. Для One-Hot Encoding можно использовать следующий код:

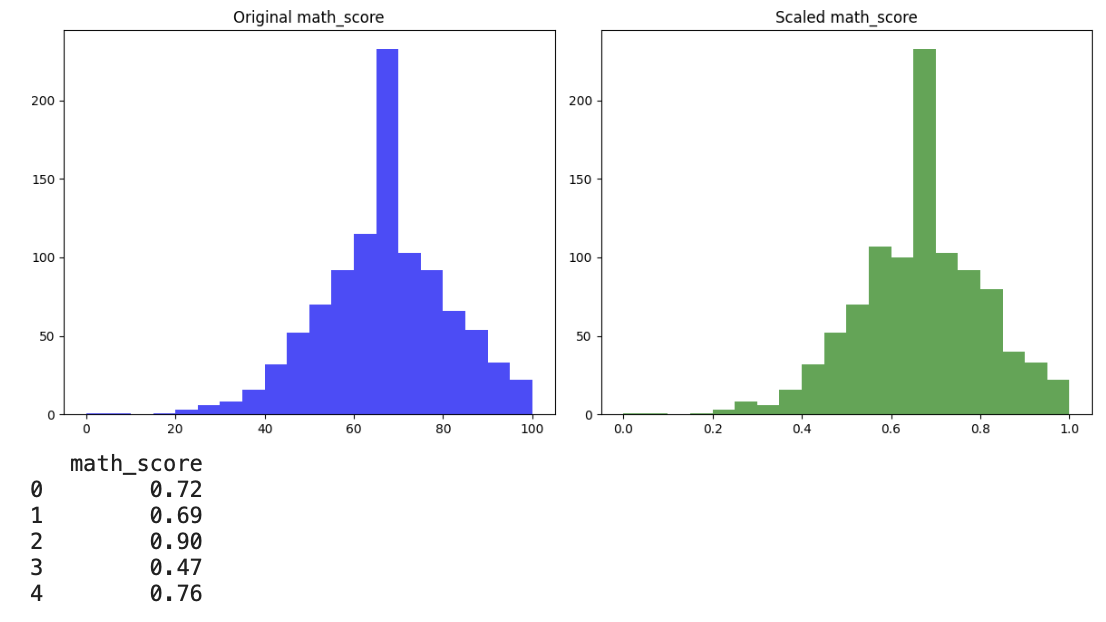
# 

Эта тепловая карта корреляции показывает степень линейной взаимосвязи между различными переменными в вашем наборе данных. Числовые значения на карте представляют коэффициент корреляции Пирсона, который может варьироваться от -1 до +1: Значение +1 указывает на совершенно положительную линейную корреляцию между двумя переменными, т.е. когда одна переменная увеличивается, другая тоже увеличивается. Значение 0 означает отсутствие линейной корреляции. Значение -1 указывает на совершенно отрицательную линейную корреляцию, т.е. когда одна переменная увеличивается, другая уменьшается. На графике видны следующие ключевые моменты: -Очень сильная корреляция между оценками по математике (math\_score), чтению (reading\_score) и письму (writing\_score). Это ожидаемо, так как успеваемость учащихся обычно коррелирует между разными предметами. -Противоположные значения корреляции между gender\_female и gender\_male показывают, что это противоположные группы — если учащийся не мужского пола, то он женского (и наоборот), что обусловливает корреляцию -1. -Для других переменных значительных корреляций не наблюдается, многие значения близки к нулю, что указывает на отсутствие сильной связи между этими категориями.

**Нормализация числовых признаков**

Нормализация числовых признаков необходима для того, чтобы все они имели одинаковый масштаб, что особенно важно для моделей, чувствительных к величине данных, например, градиентного спуска. Min-Max Scaling — приводит все числа к шкале от 0 до 1. Стандартизация (Z-score normalization) — приводит данные к распределению с нулевым средним и единичным стандартным отклонением. Для Min-Max Scaling можно использовать следующий код:





Оригинальная гистограмма (Original math\_score): Показывает исходное распределение оценок по математике, где большинство значений сосредоточены в определенном диапазоне. Это может указывать на нормальное распределение оценок или на наличие определенного уровня достижений среди учащихся.

Масштабированная гистограмма (Scaled math\_score): Показывает те же данные после применения масштабирования MinMaxScaler, который приводит все значения к диапазону от 0 до 1. Форма распределения остается той же, что указывает на то, что относительные различия между значениями сохраняются, но абсолютный масштаб изменен. Также приведенный фрагмент таблицы показывает первые пять значений масштабированного столбца math\_score, где каждое значение теперь находится между 0 и 1. Это свидетельствует о том, что масштабирование было успешно применено.

**Выводы:** Масштабирование полезно, когда данные используются в моделях машинного обучения, которые чувствительны к масштабу признаков, например, в алгоритмах на основе расстояний. Применение MinMaxScaler не изменяет форму распределения данных, но переводит все значения в унифицированный масштаб от 0 до 1, что может помочь в стандартизации данных для сравнительного анализа или в улучшении производительности моделей. Изменение масштаба не должно влиять на интерпретацию формы распределения данных; распределение оценок по математике сохраняет свои характеристики независимо от масштаба.