Проектная работа по курсу
«Системы искусственного интеллекта»
на тему «Анализ данных социальных сетей»

ЭТАП 1 **Разведочный анализ данных (EDA) и подготовка к моделированию на реальных данных**	
Цель	• Научиться работать с «живыми» данными соцсетей: находить ботов, определять качество данных, выявлять закономерности в самораскрытии пользователей. • Подготовить датасет для последующего анализа (кластеризация, классификация).
Датасет	• Основной: VK dataset • Дополнительно: 2 датасета реальных пользователей (550 аккаунтов с метаданными).
Задачи	1. **Загрузка и первичный осмотр** o Вывести размерность, типы признаков. o Посчитать долю пропусков по каждому признаку. o Найти дубликаты и аномалии. 2. **Удаление ботов / «пустых» профилей** o Предложить критерии (отсутствие друзей, фото, постов, подписок). o Построить визуализации: распределение «живых» и «пустых» профилей. o Сделать обоснованный выбор стратегии фильтрации (правила, кластеры, аномалии). 3. **Анализ заполненности профиля** o Построить распределения по ключевым полям (образование, возраст, семейное положение, число фото, друзей). o Сравнить мужчин и женщин: кто чаще указывает образование, семейное положение, кто активнее выкладывает фото. o Сформулировать и проверить гипотезы, например: • «Люди с образованием имеют более заполненный профиль» • «Люди с большим числом фото чаще пишут посты» 4. **Гендерные различия в самораскрытии** o Провести дисперсионный или регрессионный анализ: зависят ли показатели активности от пола и образования. o Визуализировать результаты (boxplot, barplot, heatmap). 5. **Подготовка данных для моделирования** o Очистить признаки (убрать нерелевантные, обработать пропуски, привести категории к единому виду). o Создать новые признаки (например, «плотность профиля» = число заполненных полей / общее число полей). o Подготовить итоговый датасет для следующих этапов.
Форма отчетности	• Jupyter Notebook (.ipynb) с кодом, графиками и пояснениями. • Отчёт с основными выводами и скриншотами ключевых графиков.
Требования к оформлению	• Графики с подписями, легендами, читаемыми осями. • Каждый блок сопровождается кратким пояснением. • Должны быть сформулированы минимум **3 гипотезы** о поведении пользователей.
Инструменты	Python, pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels.
Частые ошибки	• «Механическое» удаление строк с пропусками. • Подсчёт статистик без пояснения. • Неполные/неподписанные графики. • Нет попытки сформулировать гипотезы.
Оценка	**100 баллов**
Срок сдачи	**18 октября**

ЭТАП 2 Статистический и факторный анализ профилей пользователей социальных сетей	
Цель	• Освоить методы статистического анализа для поиска закономерностей в поведении пользователей. • Научиться выявлять скрытые факторы и строить кластеры пользователей по паттернам активности. • Проверить гипотезы о взаимосвязях между характеристиками профиля.
Датасет	• Подготовленный на Этапе 1 очищенный VK dataset. • Дополнительно: 2 датасета реальных пользователей (550 аккаунтов с метаданными).
Задачи	1. **Дисперсионный анализ (ANOVA)** o Проверить различия в заполненности профиля по полу, возрасту, образованию. o Сформулировать статистически подтверждённые выводы. 2. **Регрессионный анализ** o Построить модели зависимости активности пользователя (число постов, фото, подписчиков) от признаков профиля. o Интерпретировать коэффициенты регрессии: какие признаки вносят наибольший вклад. 3. **Факторный анализ** o Выделить скрытые факторы самораскрытия (например: «социальность», «медиа-активность», «академичность»). o Интерпретировать каждый фактор и дать названия. 4. **Кластерный анализ** o Разделить пользователей на группы по поведенческим признакам. o Визуализировать кластеры (heatmap, PCA/UMAP проекция). o Дать характеристику каждому кластеру («активные постеры», «фото-ориентированные», «минималисты» и т.п.). 5. **Проверка гипотез** o Найти и подтвердить/опровергнуть не менее 2–3 интересных взаимосвязей. o Примеры: ▪ «Люди, указывающие образование, на X% чаще заполняют семейное положение». ▪ «Пользователи с большим числом фото публикуют больше постов». 6. **Обогащение датасета (по желанию, бонус)** o Найти и подключить дополнительные данные (например, поле «образование», «город» или «работа» из VK API либо вспомогательных источников). o Описать, как это влияет на полноту анализа и качество проверяемых гипотез. o Примеры: ▪ добавить образование и проверить, как оно связано с самораскрытием; ▪ добавить город и проверить, есть ли различия между мегаполисами и малыми городами.
Форма отчетности	• Jupyter Notebook (.ipynb) с кодом, графиками и пояснениями. • Отчёт с основными выводами и скриншотами ключевых графиков.
Требования к оформлению	• Каждая модель/тест сопровождается визуализацией и текстовой интерпретацией. • Выводы должны быть связаны с поставленными гипотезами. • Минимум 2–3 статистических вывода и 1 интерпретация кластеров.
Инструменты	Python, pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels.
Частые ошибки	• Проведение тестов без проверки условий применимости. • Интерпретация коэффициентов без учёта значимости. • Формальные кластеры без попытки их осмыслить. • Отсутствие гипотез или формулировка без численных доказательств.
Оценка	**100 баллов**
Срок сдачи	**8 ноября**

ЭТАП 3 Построение среднего профиля пользователя и сравнение с результатами анализа	
Цель	• Сформировать «типичный» профиль пользователя по ключевым признакам. • Сравнить усреднённые профили разных групп (мужчины/женщины, по образованию, возрасту). • Сопоставить полученные средние профили с результатами статистического и кластерного анализа.
Датасет	• Подготовленный очищенный датасет из Этапа 1. • Результаты анализа и кластеризации из Этапа 2.
Задачи	2. **Построение усреднённых профилей** o Рассчитать «средний профиль» отдельно для мужчин и женщин. o По желанию: построить средний профиль по типу образования (технари, гуманитарии, естественники) и по возрастным группам (18–25, 26–40, 40+). **3. Сравнение с кластеризацией (Этап 2)** o Проверить: совпадают ли усреднённые профили с «центрами кластеров». o Если есть различия — интерпретировать (например: «Средний женский профиль соответствует кластеру 2, но кластер 2 делится на активных и неактивных»). **4. Интерпретация профилей** o Описать в словах полученные средние портреты (например: «типичный мужчина 18–25 лет указывает образование, редко семейное положение, имеет 50–70 друзей и мало фото»). o Сравнить с гипотезами, сформулированными ранее. **5. Визуализация** o Построить графики (барплоты, boxplot, spider/radar chart) для наглядного сравнения профилей. o Сделать минимум 2 визуализации для разных групп.
Форма отчетности	• Jupyter Notebook (.ipynb) с кодом, графиками и пояснениями. • Отчёт с основными выводами и скриншотами ключевых графиков.
Требования к оформлению	• У профилей должны быть как числовые характеристики, так и словесная интерпретация. • Сравнение должно содержать минимум 2–3 выводов. • Визуализации должны быть читаемыми (подписи, легенды).
Инструменты	Python, pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels.
Частые ошибки	• Формальное усреднение без интерпретации («средний возраст = 23.7» без комментариев). • Игнорирование сравнения с результатами предыдущего этапа. • Отсутствие визуализаций. • Слишком общие выводы без цифр.
Оценка	**100 баллов**
Срок сдачи	**29 ноября**

ЭТАП 4 Разработка программного продукта для анализа профиля пользователя	
Цель	• Создать инструмент, который позволяет автоматически анализировать профиль пользователя ВКонтакте. • Реализовать инструмент, доступный через интерфейс (бот или скрипт), который сравнивает конкретный профиль со «средним» и оценивает его особенности. • Ввести критерий «удалённости профиля от среднего».
Датасет	• Подготовленный и очищенный датасет (Этап 1). • Средние профили пользователей (Этап 3).
Задачи	**1. Разработка функционала анализа профиля** o Реализовать функцию/бот-команду analyze_profile(id). o Должно выполняться: ▪ вывод заполненных полей профиля, ▪ сравнение с «средним» профилем, ▪ отображение отличий. **2. Критерий «удалённости» от среднего профиля** o Определить и реализовать метрику (например, % совпадения, косинусное расстояние). o Выводить её при анализе. **3. Интерфейс взаимодействия** o Вариант A: Jupyter Notebook или скрипт с функцией и текстовым выводом. o Вариант B (для продвинутых): Telegram-бот с командами `/analyze <id>`, `/avg`, кнопками и визуализациями. **4. Визуализация результатов** o Добавить хотя бы один способ графической демонстрации различий: ▪ radar chart (сравнение с «средним»), ▪ таблица совпадений/отличий. **5. Документация и тестирование** o Описать работу функции/бота. o Проверить на нескольких примерах профилей. o Сделать выводы о том, какие профили ближе или дальше от «среднего».
Форма отчетности	• Код (Notebook / скрипт / бот). • Отчёт с основными выводами и скриншотами ключевых графиков.
Требования к оформлению	• Код должен быть читаемым и воспроизводимым. • Все шаги сопровождены пояснениями. • Результаты визуализированы.
Инструменты	Python, pandas, numpy, matplotlib, seaborn, scikit-learn. (По желанию: Streamlit/Flask для интерфейса).
Частые ошибки	• Формальная реализация без проверки на реальных профилях. • Нет визуализации различий. • Отсутствие метрики «удалённости» или её некорректная интерпретация. • Неполная документация к коду.
Оценка	**100 баллов**
Срок сдачи	**20 декабря**