Задания лабораторных работ по курсу «Системы искусственного интеллекта»

ЛАБОРАТОРНАЯ РАБОТА 1 Разведочный анализ данных (EDA) и подготовка к моделированию	
Цель	Научиться загружать данные, выявлять проблемы (пропуски, выбросы, типы данных), строить понятные визуализации, формулировать гипотезы и готовить датасет к моделированию.
Датасет	Использовать датасет согласно варианту (например, titanic.csv), структура столбцов стандартная: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked (+ возможные доп. поля).
Форма отчетности	• Jupyter Notebook (.ipynb) с кодом, графиками и пояснениями. • Отчёт с основными выводами и скриншотами ключевых графиков.
Требования к оформлению	Читаемые графики (подписи осей, легенды, единицы измерения), пояснительный текст к каждому блоку, отсутствие «немых» ячеек.
Инструменты	Python, pandas, numpy, matplotlib, seaborn, scikit-learn
Частые ошибки	• пустые/неподписанные графики; • перечисление цифр без выводов; • бездумное удаление строк с пропусками (покажите альтернативы); • изменение типа Pclass на float без необходимости (категория!); • закодировали категории – и забыли сохранить mapping/one-hot колонки.
Оценка:	**100 баллов**
Срок сдачи	**11 октября**

СПИСОК РЕКОМЕНДОВАННЫХ ДАТАСЕТОВ		
ЕДА		
Грибы: съедобный или ядовитый	8124×22, бинарная	https://www.kaggle.com/datasets/uciml/mushroom-classification
Семена пшеницы	210×7, 3 класса	https://www.kaggle.com/datasets/muratkokludataset/seeds-dataset
Качество красного вина	1599×11, мультикласс	https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009
ДЕНЬГИ		
Доходы населения	48k×14, бинарная	https://www.kaggle.com/datasets/uciml/adult-census-income
Маркетинг банковских услуг	45k×16, бинарная	https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing
Дефолты по кредитным картам	30k×23, бинарная	https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset
Кредитный скоринг	1k×20, бинарная	https://www.kaggle.com/datasets/uciml/german-credit
Подлинность банкнот	1372×4, бинарная	https://www.kaggle.com/datasets/ritesaluja/banknote-authentication-uci-data
Спам-письма (Spambase)	4601×57, бинарная	https://www.kaggle.com/datasets/uciml/spambase-data-set
ЗДОРОВЬЕ		
Диабет	768×8, бинарная	https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
Болезни сердца	303×13, бинарная	https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci
Рак груди	569×30, бинарная	https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Рубрика оценивания (на примере датасета «Титаник»)			
	Этап	**Баллы**	**Содержание**
1	Импорт и первичная диагностика	20	1. **Загрузка данных** o Прочитать CSV. o Показать первые/последние строки, размерность, типы столбцов. 2. **Качество данных** o Посчитать число и долю пропусков по каждому признаку. o Найти дубликаты (если есть) и решить, что с ними делать. o Проверить корректность типов (int/float/category/datetime) и при необходимости привести к нужным. В ноутбуке обязательно коротко объясняйте **почему** вы приняли то или иное решение (например, почему удалили дубликаты).
2	Описательная статистика и однофакторный анализ	20	3. **Числовые признаки** o Для количественных столбцов (`Age, Fare, ...`) вывести: `count, mean, std, min, 25%, 50%, 75%, max`. o Для каждого числового признака построить гистограмму (или KDE) и boxplot. o Коротко интерпретировать: где асимметрия, длинные хвосты, есть ли выбросы? 4. **Категориальные признаки** o Частоты категорий для `Sex, Pclass, Embarked` (таблица + столбчатая диаграмма). o Комментарий: какие категории доминируют? есть ли несбалансированность?
3	Двухфакторный анализ	20	5. **Целевой признак и группы** o Исследовать связь **`Survived`** с ключевыми факторами: ▪ `Survived` vs `Sex` (stacked barplot/percentage barplot). ▪ `Survived` vs `Pclass`. ▪ `Survived` vs `Embarked`. o Написать 2–3 наблюдения (напр., «доля выживших у женщин существенно выше»). 6. **Числовые признаки vs `Survived`** o Сравнить распределения `Age` и `Fare` в группах `Survived=0/1` (boxplot/violinplot + краткий комментарий). o Простой статистический тест (например, U-критерий Манна–Уитни) с пояснением результата. 7. **Корреляции** o Тепловая карта – только по числовым признакам; для `категория↔категория` допустимы хи-квадрат/Cramér's V. o Коротко отметить 1–2 самые сильные связи (по модулю).
4	Подготовка данных	20	8. **Работа с пропусками** o Показать выбранную стратегию: удаление/импьютация (среднее/медиана/мода/по группам). o Обосновать выбор хотя бы для `Age` и `Embarked`. 9. **Кодирование категорий** o Перевести `Sex, Embarked, Pclass` в числовой вид (Label/One-Hot Encoding). o Пояснить, почему выбран именно этот подход. 10. **Масштабирование числовых признаков** o Применить `StandardScaler` или `MinMaxScaler`. o Показать, как изменились распределения/статистики. Итог: получился чистый датасет + понятна структура признаков.

5	Анализ выбросов	10	**Провести анализ выбросов по IQR** o Отсортируйте выборку. o Найдите квартели o Посчитайте межквартильный размах o Границы «усов» boxplot o Все точки ниже нижней или выше верхней границы считаются выбросами (outliers). **11.** Не обязательно удалять**;** покажите влияние (до/после) и аргументируйте стратегию
6	Результаты и гипотезы	10	• 5–7 пунктов: что важно для выживаемости, какие проблемы были в данных, какие решения приняли. • Сформулировать 2 гипотезы для следующего шага (моделирования). Примеры: o «Женщины выживают чаще мужчин при прочих равных». o «Пассажиры 1 класса имеют больше шансов на выживание, чем 3 класса». «Увеличение тарифа (`Fare`) связано с большей вероятностью выживания».

Мини чек-лист

- Показал размерность, типы колонок, пропуски, дубликаты
- Сделал гистограммы и boxplot для всех числовых
- Посчитал частоты и сделал barplot для категориальных
- Провёл двухфакторный анализ**:** таргет vs 2–3 признака (категориальные/числовые)
- Построил heatmap корреляций только для числовых
- Обосновал стратегию пропусков, кодирования и масштабирования
- Проанализировал выбросы по IQR и показал влияние (до/после), без обязательного удаления
- Сформулировал 5–7 выводов и 2 гипотезы
- Зафиксировал random_state и вывел версии библиотек

ЛАБОРАТОРНАЯ РАБОТА 2 Задача классификации	
Цель	Построить и оценить модель классификации поверх очищенного датасета из ЛР-1; освоить базовые и продвинутые алгоритмы, метрики, интерпретацию, анализ ошибок.
Датасет	Использовать очищенный датасет из лабораторной работы 1
Форма отчетности	• Jupyter Notebook (.ipynb) с кодом, графиками и пояснениями. • Отчёт с основными выводами и скриншотами ключевых графиков.
Требования к оформлению	Читаемые графики (подписи осей, легенды, единицы), пояснительный текст к каждому блоку, отсутствие «немых» ячеек, фиксированный random_state.
Инструменты	Python, pandas, numpy, matplotlib, seaborn, scikit-learn (по желанию: shap, xgboost / lightgbm / catboost)
Частые ошибки	• Использование только accuracy (без F1/ROC-AUC/PR-AUC). • Нестратифицированный train/valid/test, отсутствует фиксированный seed. • Утечка таргета (препроцессинг по всему датасету или fit на test). • One-Hot/Scaler обучены на всём датасете, а не только на train (нет Pipeline/ColumnTransformer). • Отсутствует анализ ошибок.
Оценка:	**100 баллов**
Срок сдачи	**8 ноября**

Рубрика оценивания			
	Этап	**Баллы**	**Содержание**
1	Повторная загрузка и препроцессинг (Pipeline)	10	• Возьмите очищенный датасет из ЛР-1. • Соберите единый sklearn-pipeline (например, `ColumnTransformer` + `StandardScaler` для числовых + `OneHotEncoder` для категориальных). • Все преобразования fit только на train (чтобы избежать утечки). • Зафиксируйте список признаков и стратегию обработки пропусков (медиана/мода/константа – обосновать).
2	Разбиение и валидация	15	• Стратифицированный `train/valid/test = 60/20/20`, `random_state=42` (или заданный). • Для дисбалансных задач — проверьте стратификацию. • Допустимо CV (например, 5-fold) вместо фиксированного valid, но test оставить отложенным. • Test-набор не используем до финального выбора модели и порога; все решения принимаются по train/valid (или CV). • При CV используйте `StratifiedKFold(shuffle=True, random_state=42)`.
3	Базовые модели	15	• Обучите LogisticRegression и DecisionTreeClassifier (подберите `max_depth`). • Отчёт по метрикам на valid**:** F1**,** ROC-AUC или PR-AUC (если дисбаланс заметный). • Если доля позитивного класса < 30%**,** помимо F1 обязательно показать PR-AUC. • Краткий вывод: что и почему лучше как baseline.
4	Продвинутая модель и поиск гиперпараметров	30	• Одна из: CatBoost / LightGBM / XGBoost. • Поиск 5–10 конфигураций (Grid/Random/`SearchCV`); сохраните лог результатов. • Сравните с baseline, выберите «победителя».

5	Интерпретация и анализ ошибок	25	• Explainability: Permutation Importance (обязательно), SHAP – по желанию (если позволяет среда). • Error analysis**:** покажите минимум 3 FN и 3 FP с комментариями; сгруппируйте ошибки по одному признаку (например, `sex`/`class`/`length_bin`) и предложите 1–2 шага улучшения.
6	Репродуцируемость	5	• Зафиксировать `random_state/seed`. • Вывести версии библиотек (`sys.version`, `pandas.__version__`, `sklearn.__version__`). • Убедиться, что ноутбук запускается без ошибок.

Мини чек-лист

Данные и Pipeline

- Использую очищенный датасет из ЛР-1
- Все преобразования собраны в sklearn Pipeline/ColumnTransformer
- fit только на train (импьютация/скейлер/one-hot) – утечек нет
- Список признаков и стратегия пропусков зафиксированы и описаны

Разбиение / валидация

- Stratified `train/valid/test = 60/20/20`, `random_state=42`
- При CV: `StratifiedKFold(shuffle=True, random_state=42)`
- Test не трогаю до финального выбора модели/порога

Базовые модели

- Обучены LogisticRegression и DecisionTree (подобран `max_depth`)
- Метрики на valid**:** F1 + ROC-AUC (и PR-AUC, если позитивов < 30%)
- Краткий вывод: какая baseline-модель лучше и почему

Продвинутая модель

- Одна из: CatBoost / LightGBM / XGBoost
- Поиск 5–10 конфигураций (Grid/Random/SearchCV) с логом результатов
- Сравнение с baseline — выбран «победитель»

Интерпретация и ошибки

- SHAP (summary + 1–2 waterfall) или permutation importance
- Показаны минимум 3 FN и 3 FP с кратким разбором причин и идеями улучшения

Репродуцируемость

- Зафиксировал `random_state/seed`
- Вывел версии библиотек
- Убедился, что ноутбук запускается без ошибок

ЛАБОРАТОРНАЯ РАБОТА 3 Задачи машинного обучения	
Цель	Закрепить навыки применения методов машинного обучения. На выбор: • Построить и оценить модель регрессии. • Провести кластеризацию и интерпретировать её. • Построить простую рекомендательную систему**.**
Датасет	Любой открытый (по желанию можно взять тот же, что в ЛР-1/ЛР-2, если подходит).
Форма отчетности	• Jupyter Notebook (.ipynb) с кодом, графиками и пояснениями. • Отчёт с основными выводами и скриншотами ключевых графиков.
Требования к оформлению	читаемые графики (подписи осей, легенды, единицы), пояснительный текст к каждому блоку, отсутствие «немых» ячеек, фиксированный random_state.
Инструменты	Python, pandas, numpy, matplotlib, seaborn, scikit-learn (по желанию: surprise, implicit, xgboost/lightgbm/catboost, umap, shap)
Частые ошибки	• Использование только одной метрики без пояснений. • Отсутствие baseline-модели. • Нет сравнения методов. • Отсутствие анализа ошибок или интерпретации. • Нет пояснений к визуализациям.
Оценка:	**100 баллов**
Срок сдачи	**13 декабря**

Рубрика оценивания			
	Этап	**Баллы**	**Содержание**
1	Постановка задачи и данные	10	Кратко описать задачу (регрессия / кластеризация / рекомендации). Описание датасета, признаков и целевой переменной (если есть). Обоснование выбора.
2	Базовые методы	15	• Для регрессии: LinearRegression, Ridge/Lasso. • Для кластеризации: KMeans. • Для рекомендаций: baseline (например, популярность или user-average). Показать метрики/графики + краткий вывод.
3	Продвинутый метод	30	• Для регрессии: GradientBoosting/RandomForest/ XGBoost. • Для кластеризации: DBSCAN/Hierarchical. • Для рекомендаций: ALS/Matrix Factorization/Content-based. Сравнить с baseline, выбрать «победителя».
4	Метрики и сравнение	15	• Для регрессии: MSE/RMSE, MAE, R^2. • Для кластеризации: silhouette, inertia, ARI. • Для рекомендаций: Precision@10, Recall@10, MAP (или NDCG@10 по желанию). Краткий вывод о качестве.
5	Интерпретация и визуализация	15	Показать важность признаков (Permutation Importance / SHAP) или визуализацию (кластеров/сниженного пространства/топ-N рекомендаций). Объяснить результаты словами.
6	Анализ ошибок или результатов	10	• Для регрессии: примеры с наибольшими ошибками. • Для кластеризации: «неудачные» точки. • Для рекомендаций: примеры неудачных рекомендаций. Предложить идеи улучшения.
7	Репродуцируемость	5	Зафиксировать seed, вывести версии библиотек. Код запускается без ошибок.

Мини чек-лист

- Выбрал тип задачи (регрессия / кластеризация / рекомендации).
- Описал датасет и задачу.
- Реализовал baseline-метод.
- Реализовал продвинутый метод.
- Посчитал метрики, сделал сравнение.
- Добавил интерпретацию/визуализацию.
- Провёл анализ ошибок или ограничений.
- Зафиксировал seed, проверил воспроизводимость.