Экзамен по дисциплине «Технологии анализа данных и машинное обучение», состоит из теоретической и практической частей. Данная модель устанавливает следующую схему выставления баллов промежуточной аттестации:

* 40 баллов текущего контроля определяются по итогам изучения всей дисциплины;
* 20 баллов из 60 отводится для выполнения теоретического задания;
* 40 баллов выставляются за выполнение практического задания.

Практическая часть представляет собой задачу по анализу конкретного датасета. Для каждого задания будет необходимо:

1. Загрузить датасет в Python.
2. Описать набор данных и решаемую задачу.
3. Выделить целевую переменную и факторные переменные.
4. Удалить ненужные данные, проанализировать отсутствующие значения.
5. Прокомментировать количественные параметры датасета.
6. Разбить выборку на обучающую и тестовую.
7. Работа по вариантам.

Вариант 1. Очистка данных и обучение моделей.

Данный вариант предполагает фокусировку на обучении нескольких видов моделей обучения с учителем. В зависимости от набора данных, может предполагаться задача классификации и регрессии. Необходимо после минимальной подготовки датасета к обучению обучить несколько моделей и сравнить их эффективность.

Вариант 2. Описательный анализ и визуализация данных.

Данный вариант предполагает фокусировку на исследовании данных и визуализации. При решении этого варианта следует провести как можно более подробный описательный анализ данных с использованием максимального спектра средств визуализации. При этом следует делать значимые выводы об обнаруженных в данных закономерностях.

Вариант 3. Построение модели и оптимизация гиперпараметров.

Данный вариант предполагает фокусировку на процессе улучшения эффективности модели обучения с учителем. Студенту следует подготовить датасет к обучению, обучить одну из моделей с учителем со значениями гиперпараметров по умолчанию, получить значение эффективности. После этого вручную или автоматически подобрать значения гиперпараметров таким образом, чтобы получить максимальный прирост эффективности.

Вариант 4. Выбор признаков.

Данный вариант предполагает фокусировку на улучшении модели путем ввода новых признаков в модель. Следует подготовить модель к обучению, обучить модель и зафиксировать начальный уровень эффективности. Затем следует исследовать влияние исключения существующих и введения новых признаков в модель на эффективность. Как вариант можно рассматривать введение полиномиальных признаков. Следует стремиться к максимальному увеличению эффективности модели.

Вариант 5. Исследование влияния обучения без учителя на эффективность обучения.

Данный вариант предполагает фокусировку на использовании методов обучения без учителя для ускорения или повышения эффективности обучения с учителем. Следует подготовить модель к обучению, обучить модель и зафиксировать начальный уровень эффективности. Затем следует попробовать применить понижение размерности, обнаружение аномалий или кластеризацию (в любой комбинации) для трансформации исходного датасета. В конце работы следует сделать значимый вывод об изменении скорости и эффективности обучения с учителем.

Для выполнения кейса задачи, рекомендуется использовать язык программирования Python и специализированные библиотеки для анализа данных и машинного обучения.

Студент должен прислать рабочий notebook с решением кейса задачи, комментариями кода и аналитическими выводами до конца экзамена на электронную почту преподавателя.

Критерии оценки практической экзаменационной работы:

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. Визуализация. Работа должна демонстрировать навыки студента визуализировать информацию. Особенно на этапах описательного анализа и анализа обучаемости модели. Оценивается разнообразие, наглядность и информативность визуализации.
4. 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.
5. Валидность результатов. Студент должен продемонстрировать умение оценивать достоверность измерения метрик моделей и повышать ее с использованием перекрестной проверки (кросс-валидации). Использование k-fold cross validation является предпочтительным методом измерения эффективности модели. Если происходит выбор модели, то ее итоговая эффективность должна измеряться на чистом наборе данных.

Список датасетов, использующихся на экзамене:

1. <https://www.kaggle.com/uciml/mushroom-classification>
2. <https://www.kaggle.com/lodetomasi1995/income-classification>
3. <https://www.kaggle.com/uciml/glass>
4. <https://www.kaggle.com/uciml/german-credit>
5. <https://www.kaggle.com/zaurbegiev/my-dataset>
6. <https://www.kaggle.com/kaushiksuresh147/customer-segmentation>
7. <https://www.kaggle.com/deepu1109/star-dataset>
8. <https://www.kaggle.com/vinesmsuic/star-categorization-giants-and-dwarfs>
9. <https://www.kaggle.com/shebrahimi/financial-distress>
10. <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>
11. <https://www.kaggle.com/amir75/caesarean-section-classification>
12. <https://www.kaggle.com/sachinsharma1123/performance-prediction>
13. <https://www.kaggle.com/ninzaami/loan-predication>
14. <https://www.kaggle.com/caparrini/beatsdataset>
15. <https://www.kaggle.com/zhiruo19/covid19-symptoms-classification>
16. <https://www.kaggle.com/kunalvsingh93/banking-modelclassification>
17. <https://www.kaggle.com/mansoordaku/ckdisease>
18. <https://www.kaggle.com/mnassrib/telecom-churn-datasets>
19. <https://www.kaggle.com/akshayksingh/kidney-disease-dataset>
20. <https://www.kaggle.com/henriqueyamahata/bank-marketing>
21. <https://www.kaggle.com/maajdl/yeh-concret-data>
22. <https://www.kaggle.com/hellbuoy/car-price-prediction>
23. <https://www.kaggle.com/rhuebner/human-resources-data-set>
24. <https://www.kaggle.com/loveall/appliances-energy-prediction>
25. <https://www.kaggle.com/elikplim/forest-fires-data-set>
26. <https://www.kaggle.com/nandvard/microsoft-data-science-capstone>
27. <https://www.kaggle.com/shebrahimi/financial-distress>
28. <https://www.kaggle.com/aungpyaeap/beauty>
29. <https://www.kaggle.com/vbmokin/ammonium-prediction-in-river-water>
30. <https://www.kaggle.com/veer06b/marrket-mix-dataset>

Вторая часть экзамена - теоретическая. В ходе оценки ответа студента преподаватель оценивает понимание студентом пройденного материала и формирует итоговую оценку за экзамен.

Вопросы по теоретической части:

1. Понятие машинного обучения. Отличие машинного обучения от других областей программирования.
2. Классификация задач машинного обучения. Примеры задач из различных классов.
3. Основные понятия машинного обучения: набора данных, объекты, признаки, атрибуты, модели, параметры.
4. Структура и представление данных для машинного обучения.
5. Инструментальные средства машинного обучения.
6. Задача регрессии: постановка, математическая формализация.
7. Метод градиентного спуска для парной линейной регрессии.
8. Понятие функции ошибки: требования, использование, примеры.
9. Множественная и нелинейная регрессии.
10. Нормализация признаков в задачах регрессии.
11. Задача классификации: постановка, математическая формализация.
12. Метод градиентного спуска для задач классификации.
13. Логистическая регрессия в задачах классификации.
14. Множественная и многоклассовая классификация. Алгоритм “один против всех”.
15. Метод опорных векторов в задачах классификации.
16. Понятие ядра и виды ядер в методе опорных векторов.
17. Метод решающих деревьев в задачах классификации.
18. Метод k ближайших соседей в задачах классификации.
19. Однослойный перцептрон в задачах классификации.
20. Метрики эффективности и функции ошибки: назначение, примеры, различия.
21. Понятие набора данных (датасета) в машинном обучении. Требования, представление. Признаки и объекты.
22. Шкалы измерения признаков. Виды шкал, их характеристика.
23. Понятие чистых данных. Определение, очистка данных.
24. Основные этапы проекта по машинному обучению.
25. Предварительный анализ данных: задачи, методы, цели.
26. Проблема отсутствующих данных: причины, исследование, пути решения.
27. Проблема несбалансированных классов: исследование, пути решения.
28. Понятие параметров и гиперпараметров модели. Обучение параметров и гиперпараметров. Поиск по сетке.
29. Понятие недо- и переобучения. Определение, пути решения.
30. Диагностика модели машинного обучения. Методы, цели.
31. Проблема выбора модели машинного обучения. Сравнение моделей.
32. Измерение эффективности работы моделей машинного обучения. Метрики эффективности.
33. Метрики эффективности моделей классификации. Виды, характеристика, выбор.
34. Метрики эффективности моделей регрессии. Виды, характеристика, выбор.
35. Перекрестная проверка (кросс-валидация). Назначение, схема работы.
36. Конвейеры в библиотеке sklearn. Назначение, использование.
37. Использование методов визуализации данных для предварительного анализа.
38. Исследование коррелированности признаков: методы, цели, выводы.
39. Решкалирование данных. Виды, назначение, применение. Нормализация и стандартизация данных.
40. Преобразование категориальных признаков в числовые.
41. Методы визуализации данных для машинного обучения.
42. Задача выбора модели. Оценка эффективности, валидационный набор.
43. Кривые обучения для диагностики моделей машинного обучения.
44. Регуляризация моделей машинного обучения. Назначение, виды, формализация.
45. Проблема сбора и интеграции данных для машинного обучения.
46. Понятие чистых данных и требования к данным.
47. Основные задачи описательного анализа данных.
48. Полиномиальные модели машинного обучения.
49. Основные виды преобразования данных для подготовки к машинному обучению.
50. Задача выбора признаков в машинном обучении.
51. Ансамблевые модели машинного обучения. Виды ансамблирования.
52. Конвейеризация моделей машинного обучения.
53. Методы векторизации текстов для задач машинного обучения.
54. Представление графической информации в моделях машинного обучения.
55. Задачи без учителя. Кластеризация. Метод k средних.
56. Задачи без учителя. Обнаружение аномалий.
57. Задачи без учителя. Понижение размерности. Метод PCA.
58. Воспроизводимость алгоритма преобразования данных в машинном обучении.
59. Случайный лес как ансамблевая модель машинного обучения.
60. Частичное обучение с учителем.