

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное
образовательное учреждение высшего образования
«Казанский (Приволжский) федеральный университет»
Институт вычислительной математики и информационных технологий

УТВЕРЖДАЮ

Проректор

(подпись) И.А. Хайруллин
(инициалы, фамилия)

« ____ » _____ 20__ г.

**Дополнительная профессиональная программа
повышения квалификации
«Профессиональное программирование на python.
Анализ данных и введение в машинное обучение»
(наименование программы)**

Утверждена учебно-методической комиссией института вычислительной математики и информационных технологий КФУ (протокол № ____ от « ____ » _____ 20__ г.)

Председатель комиссии _____

(подпись)

Директор ИВМиИТ

(подпись)

Егорчев А.А.
(инициалы, фамилия)

« ____ » _____ 20__ г.

Казань, 2025

1. ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ

1.1. Цель реализации программы

Целью реализации программы является совершенствование и получение новой компетенции, необходимой для профессиональной деятельности.

1.2. Планируемые результаты обучения

Слушатель в результате освоения программы должен обладать следующими знаниями:

- Структуру и возможности библиотеки NumPy для создания и обработки многомерных массивов, включая основы линейной алгебры;
- Основные структуры данных библиотеки Pandas и приёмы предварительного анализа, очистки и агрегирования табличных данных;
- Основы графического представления данных и возможности библиотек Matplotlib (настройка стилей, создание анимаций) и Seaborn (статистические графики);
- Математические модели и сферы применения ключевых алгоритмов машинного обучения – линейной / логистической регрессии, деревьев решений и случайных лесов;
- Концепции алгоритмов кластеризации, включая алгоритмы k-means и DBSCAN, а также принципы интерпретации кластеров;
- Методы оценки и валидации моделей: перекрёстная валидация, матрица ошибок, ROC-кривая и выбор подходящих метрик для задач регрессии, и классификации.

Слушатель должен уметь:

- Создавать и модифицировать массивы NumPy, выполнять векторизованные вычисления и простейшие линейно-алгебраические операции;
- Загружать данные в Pandas, выполнять фильтрацию, группировку, агрегирование и формировать краткие статистические отчёты;
- Создавать публикационно-качественные графики в Matplotlib и статистические визуализации в Seaborn, включая настройку стилей и построение анимаций;
- Строить, обучать и интерпретировать модели линейной / логистической регрессии, деревьев решений и случайных лесов с помощью scikit-learn;
- Применять и настраивать алгоритмы кластеризации (*k-means*, DBSCAN) для обнаружения скрытых закономерностей и визуализировать результаты;
- Реализовывать полный ML-конвейер: разделение данных, кросс-валидация, подбор метрик, визуальный анализ качества моделей и документирование процесса в Jupyter-ноутбуке.

1.3. Требования к уровню подготовки поступающего на обучение (при необходимости)

К освоению дополнительных профессиональных программ допускаются:

1) лица, имеющие среднее профессиональное и (или) высшее образование;

1.4. Программа разработана на основе:

Профессионального стандарта «Программист» (утвержденного приказом Минтруда России от 20.07.2022 N 424н «Об утверждении профессионального стандарта «Программист»).

1.5. Форма обучения – заочная с применением дистанционных технологий.

2. СОДЕРЖАНИЕ ПРОГРАММЫ

2.1. Учебный план

Наименование раздела	Тру- до- ем-	Дистанционные занятия	СРС, час
-------------------------	--------------------	-----------------------	-------------

		Всего, час.	в том числе		
			лекции	прак. занятия, семинары	
1	2	3	4	5	6
NumPy: массивы и базовые операции	8	4	2	2	4
NumPy углублённо: индексация, срезы и линейная алгебра	8	4	2	2	4
Pandas: от сырого файла к структурированному DataFrame	8	4	2	2	4
Matplotlib без секретов: от базовых графиков к интерактивной анимации	8	4	2	2	3
Seaborn для статистической визуализации: категории, регрессионные модели и дизайн	8	4	2	2	3
Первые шаги в ML: модели линейной и логистической регрессий	6	4	2	2	4
Деревья решений и случайный лес: интерпретируемые и ансамблевые методы	8	4	2	2	4
Обучение без учителя: методы кластеризации k-means и DBSCAN	8	4	2	2	4
Как измерить качество модели: кросс-валидация, метрики и ROC-кривая	8	4	2	2	4
Всего	70	36	18	18	34
Итоговая аттестация	2	-	-	-	-
Итого	72	36	18	18	34

2.2. Календарный учебный график

Период обучения (дни, недели) ¹⁾	Наименование раздела
1-й день	NumPy: массивы и базовые операции. NumPy углублённо: индексация, срезы и линейная алгебра.
2-й день	Pandas: от сырого файла к структурированному DataFrame.
3-й день	Matplotlib без секретов: от базовых графиков к интерактивной анимации. Seaborn для статистической визуализации: категории, регрессионные модели и дизайн.
4-й день	Первые шаги в ML: модели линейной и логистической регрессий. Деревья решений и случайный лес: интерпретируемые и ансамблевые методы.
5-й день	Обучение без учителя: методы кластеризации k-means и DBSCAN.

	Как измерить качество модели: кросс-валидация, метрики и ROC-кривая.
¹⁾ Даты обучения будут определены в расписании занятий при наборе группы на обучение	

2.3. Рабочие программы разделов

№, наименование темы	Содержание лекций (количество часов)	Наименование практических занятий или семинаров (количество часов)	Виды СРС (количество часов)
1	2	3	4
NumPy: массивы и базовые операции	Вводятся nd-массивы NumPy, способы их создания, основные арифметические и логические операции, преимущества векторизации (2 часа)	Генерация двумерных массивов, выполнение поэлементных и агрегирующих операций, измерение выигрыша по скорости против чистого Python (2 часа)	Реализовать скрипт статистического анализа (среднее, медиана, стандартное отклонение) случайно сгенерированного массива данных (3 часа)
NumPy углублённо: индексация, срезы и линейная алгебра	Подробно разбираются расширенная индексация, маскирование, reshape/transpose, а также базовые линейно-алгебраические функции dot, eig, inv (2 часа)	Решение системы линейных уравнений методом NumPy, сравнение точности и скорости разных способов обращения к элементам массива (2 часа)	Составить и проанализировать матрицу корреляций 10-мерного датасета; визуализировать тепловую карту любым удобным инструментом, подготовить мини-отчет (4 часа)
Pandas: от сырого файла к структурированному DataFrame	Ознакомление со структурами Series и DataFrame, загрузка/сохранение данных, первичный обзор (head, info, describe), работа с пропусками (2 часа)	Импорт CSV-файла, очистка пропусков, фильтрация и группировка данных с расчётом агрегированных метрик; вывод сводной таблицы (2 часа)	Подготовить самостоятельный анализ реального набора данных (выбор набора из открытых

			источников), сформировать отчёт в Jupyter с графиками и выводами (4 часа)
Matplotlib без секретов: от базовых графиков к интерактивной анимации	Рассматривается устройство объектной модели Matplotlib (Figure → Axes), порядок построения линий, столбчатых и точечных диаграмм, а также продвинутое возможности – стили, возможности компоновки графиков, интерактивные элементы и создание анимаций (2 часа).	Построение многоосевых графиков по реальному набору данных, настройка цветовых схем и добавление анимированной линии, экспорт результатов в PNG и GIF (2 часа).	Подготовить графический отчёт о собственном датасете, включив более 3 разных типов графиков и простую анимацию или интерактивный элемент, оформить всё в Jupyter-ноутбуке (3 часа).
Seaborn для статистической визуализации: категории, регрессионные модели и дизайн	Рассматриваются высокоуровневые функции Seaborn для категориальных (bar/box/violin) и регрессионных (regplot, lmpplot) графиков, настройка тем и палитр, автоматическое добавление статистических подсказок (2 часа).	Исследование датасета tips/iris (или других): построение pairplot, catplot и regplot, доработка дизайна до публикационного качества и сохранение в SVG (2 часа).	Выполнить EDA на предоставленном датасете (или собственном), сформировать не менее четырёх визуализаций Seaborn, снабдить краткими текстовыми выводами (3 часа).
Первые шаги в ML: модели линейной и логистической регрессии	Введение в задачи регрессии и классификации; линейная регрессия: функция потерь, градиентный спуск, метрики MSE/R ² ; логистическая регрессия: сигмоида, вероятность класса, метрики Accuracy/Recall/F1 (2 часа).	Обучение линейной и логистической регрессии; анализ коэффициентов, построение графиков ошибок (2 часа).	Реализовать многоклассовую логистическую регрессию и оформить отчет (4 часа).
Деревья решений и случайный	Алгоритм дерева решений: критерии разделения, переобучение, обрезка; случайный лес как ансамбль, оценка	Построение и визуализация дерева классификации, затем	Сравнить выбранные модели по

лес: интерпретируемые и ансамблевые методы	важности признаков, плюсы и минусы (2 часа).	обучение случайного леса с GridSearchCV, вывод TOP-10 важных признаков (2 часа).	качеству и скорости, представить краткий отчёт (4 часа).
Обучение без учителя: методы кластеризации k-means и DBSCAN	Основы кластеризации; алгоритм k-means, выбор k, инерция, силуэтометрика; DBSCAN как метод плотностной кластеризации, параметры eps/min_samples, работа с выбросами (2 часа).	Сегментация клиентов методом k-means с визуализацией PCA-проекций; применение DBSCAN к геоаннотациям, сравнение кластеров и «шума» (2 часа).	Исследовать иерархическую кластеризацию на выбранном датасете, построить дендрограмму и сопоставить кластеры с результатами k-means (4 часа).
Как измерить качество модели: кросс-валидация, метрики и ROC-кривая	Разделение данных и k-fold кросс-валидация; метрики классификации (точность, полнота, F1, AUC) и регрессии (MAE, RMSE); построение ROC-кривой и интерпретация площади под кривой (2 часа).	Реализация полного CV-конвейера для случайного леса, вывод матрицы ошибок и ROC-кривой для десятка моделей; сравнение метрик (2 часа).	Написать скрипт, автоматически рассчитывающий выбранные метрики и строящий ROC-кривые для разных алгоритмов, результаты оформить в виде Jupyter-ноутбука (4 часа).

2.4. Оценка качества освоения программы (формы аттестации, оценочные и методические материалы)

2.4.1. Форма(ы) промежуточной и итоговой аттестации

Итоговая аттестация (зачёт) проводится в форме онлайн-тестирования, включающего как теоретические вопросы, так и практические, содержащие короткие фрагменты кода по материалам изученных тем. К тесту допускаются слушатели, предъявившие все выполненные задания для самостоятельной работы (скрипты / Jupyter ноутбуки с отчётами).

2.4.2. Примеры вопросов теста

Слушателю необходимо выбрать 1 или несколько правильных ответов.

Вопрос	Ответ №1	Ответ №2	Ответ №3	Ответ №4
Для получения информации о	bar	scatter	hist	heatmap

зависимости значений одного столбца датафрейма от другого лучше всего подойдет график типа				
Какая из следующих функций потерь чаще всего используется в задачах регрессии?	Categorical cross-entropy	Mean squared error (MSE)	Binary cross-entropy	Hinge loss
Предположим, у вас есть DataFrame df со столбцом 'age'. Какой из приведённых вариантов корректно выбирает строки, в которых значение 'age' больше 25 и меньше 40?	df['age'] > 25 and df['age'] < 40	df[df['age'] > 25 & df['age'] < 40]	df[(df['age'] > 25) & (df['age'] < 40)]	df.query('age > 25 and age < 40')

Оценка результатов:

Результаты итоговой аттестации определяются следующими оценками: «зачтено», «не зачтено».

Оценка «не зачтено» ставится, если:

- слушатель верно ответил менее чем на 60% вопросов итогового тестирования;

Оценка «зачтено» ставится, если:

- слушатель верно ответил на 60% и более вопросов итогового тестирования;

3. ОРГАНИЗАЦИОННО-ПЕДАГОГИЧЕСКИЕ УСЛОВИЯ

3.1. Материально-технические условия

Наименование специализированных учебных помещений	Вид занятий	Наименование оборудования, программного обеспечения
Не требуется (дистанционный формат обучения)	Лекция, практическое занятие	Компьютер, интернет, гарнитура

3.2. Учебно-методическое и информационное обеспечение

Список литературы

1. Златопольский, Д. М. Основы программирования на языке Python : учебник / Златопольский Д. М. - Москва : ДМК Пресс, 2017. - 284 с. - ISBN 978-5-97060-552-3. - Текст: электронный // ЭБС 'Консультант студента' : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785970605523.html> (дата обращения: 10.01.2025). - Режим доступа : по подписке.
2. Жуков, Р. А. Язык программирования Python: практикум : учебное пособие / Р.А. Жуков. - Москва : ИНФРА-М, 2024. - 216 с. - (Высшее образование). - DOI 10.12737/textbook_5cb5ca35aaa7f5.89424805. - ISBN 978-5-16-018516-3. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/2139862> (дата обращения: 10.01.2025). - Режим доступа: по подписке.
3. Постолиит А. В. Основы искусственного интеллекта в примерах на Python / А.В. Постолиит. - Санкт-Петербург : БХВ-Петербург, 2021. - 448 с. - ISBN 978-5-9775-6765-7. - URL: <https://ibooks.ru/bookshelf/380050> (дата обращения: 10.01.2025). - Текст: электронный.
4. Васильев Юлий. Python для data science. - (Серия 'Библиотека программиста'). - Санкт-Петербург : Питер, 2023. - 272 с. - ISBN 978-5-4461-2392-6. - URL:

<https://ibooks.ru/bookshelf/390133/reading> (дата обращения: 10.01.2025). - Текст: электронный.

3.3. Кадровые условия

Обеспечение образовательной программы осуществляется преподавательским составом из числа докторов и кандидатов наук КФУ (или имеющих уникальный опыт по направлению образовательной программы).

3.4. Условия для функционирования электронной информационно-образовательной среды (при реализации программ с использованием дистанционных образовательных технологий)

Электронные информационные ресурсы	Вид занятий	Наименование оборудования, программного обеспечения
Яндекс.360: Телемост	Лекция, практическое занятие	Компьютер, интернет, гарнитура

4. РУКОВОДИТЕЛЬ И АВТОР(Ы) ПРОГРАММЫ

Руководитель:

Егорчев Антон Александрович, директор Института вычислительной математики и информационных технологий.

Авторы:

Тумаков Дмитрий Николаевич, заведующий кафедрой прикладной математики и искусственного интеллекта Института вычислительной математики и информационных технологий,

Гиниятова Динара Халиловна, старший преподаватель кафедры прикладной математики и искусственного интеллекта Института вычислительной математики и информационных технологий