МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО «СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ:
Директор НОЦ «Институт
непрерывного образования»
_____Е.В. Мошкина
«11» января 2023 г.

ДОПОЛНИТЕЛЬНАЯ ПРОФЕССИОНАЛЬНАЯ ПРОГРАММА
ПОВЫШЕНИЯ КВАЛИФИКАЦИИ

«Анализ больших данных на Python»

Красноярск 2023

I. ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ

1.1. Аннотация программы

Программа повышения квалификации разработана, чтобы дать слушателям понимание концепций науки о данных и основ машинного обучения. Вы узнаете, как использовать язык программирования Python для анализа данных, визуализации данных, машинного обучения и научных исследований. Курс предоставляет информацию об особенностях больших наборов данных и работе с ними с использованием различных библиотек Python: применении matplotlib для визуализации, линейной алгебре в numpy, теории вероятностей с помощью scipy, введении в pandas и, наконец, о машинном обучении с помощью scikit-learn.

Программа повышения квалификации предназначена для широкого круга слушателей, кто хочет начать и/или продолжить карьеру в области наук о данных, машинного обучения или смежных областях; для преподавателей технических направлений, цифровой гуманитаристики. А также для преподавателей, работающих в рамках проекта «Цифровая кафедра Сибирского федерального университета» и др.

1.2. Цель программы

Цель программы повышения квалификации — формирование и(или) совершенствование у слушателей компетенций в области обработки больших данных.

1.3. Компетенции (трудовые функции) в соответствии с Профессиональным стандартом (формирование новых или совершенствование имеющихся)

Программа разработана на основе квалификационных характеристик должностей руководителей и специалистов высшего профессионального и дополнительного профессионального образования, утвержденных приказом Минздравсоцразвития РФ от 11 января 2011 г. № 1н (ЕКСД РФ) и соответствует требованиям Порядка организации и осуществления образовательной деятельности по дополнительным профессиональным программам, утвержденного приказом Минобрнауки России от 1 июля 2013 г. № 499, приказа Минобрнауки России от 29 марта 2019 г. № 178, а также с учетом прогноза научно-технологического развития Российской Федерации до 2030 года.

Программа направлена на совершенствование компетенций (совершенствование способов и средств исполнения должностных обязанностей в соответствии с ЕКСД РФ) в части III «Должности профессорско-преподавательского состава»:

- организация и осуществление учебной и учебно-методической работы по преподаваемой дисциплине или отдельным видам учебных занятий;
- организация и планирование методического и технического обеспечения учебных занятий.

В соответствии с профессиональным стандартом 06.046 «Специалист по моделированию, сбору и анализу данных цифрового следа» (утв. приказом

Минтруда России от 04.02.2021 г. № 39н), программа направлена на формирование и(или) совершенствование следующих трудовых функций:

- B/02.5 Проверка гипотез, представленных в модели деятельности человека (группы людей) и ИКС, поиск закономерностей.
- B/03.5 Визуализация данных анализа цифрового следа в соответствии с моделью деятельности человека (группы людей) и ИКС.

1.4. Планируемые результаты обучения

Слушатель в результате освоения программы повышения квалификации сможет достичь следующих результатов:

РО1. Знать методы структурирования наборов данных.

РО2. Проводить очистку данных цифрового следа (поиск аномалий, корректировка, подсказка, автоматизация/уменьшение объема ручной работы, поиск дубликатов) записей и пользователей.

РО3. Проводить сравнительный анализ для проверки гипотез, представленных в модели деятельности человека (группы людей) и ИКС, относящиеся к различным прикладным областям.

РО4. Применять библиотеки и фреймворки языка программирования Python для визуализации данных.

1.5. Категория слушателей

Профессорско-преподавательский состав, учебно-вспомогательный персонал, а также административно-управленческий персонал университета.

1.6. Требования к уровню подготовки поступающего на обучение

Высшее образование.

1.7. Продолжительность обучения: 48 часов, из них 22 контактных.

1.8. Форма обучения: заочная с использованием дистанционных образовательных технологий.

1.9. Требования к материально-техническому обеспечению

Наличие у каждого слушателя компьютера и телефона с доступом к Интернету, микрофон.

1.10. Особенности (принципы) построения дополнительной профессиональной программы повышения квалификации

Особенности построения программы повышения квалификации «Аналитика больших данных на Python»:

- модульная структура программы;
- в основу проектирования программы положен компетентностный подход;

- выполнение комплексных (сквозных) учебных заданий, требующих практического применения знаний и умений, полученных в ходе изучения логически связанных дисциплин (модулей);
- использование информационных и коммуникационных технологий, в том числе современных систем технологической поддержки процесса обучения, обеспечивающих комфортные условия для обучающихся, преподавателей;
- применение электронных образовательных ресурсов (дистанционное, электронное, комбинированное обучение и пр.).

В поддержку дополнительной профессиональной программы повышения квалификации разработан электронный курс в системе электронного обучения СФУ «е-Курсы».

1.10. Документ об образовании: удостоверение о повышении квалификации установленного образца.

II. ОСНОВНОЕ СОДЕРЖАНИЕ ПРОГРАММЫ

2.1. Учебно-тематический план

№ п/п	Наименование и содержание разделов и тем программы	Всего часов	В том числе:		Использование средств ЭО и ДОТ	Результаты обучения
			Контактная работа	Самостоятельная работа		
1	**Основы языка Python**	**12**	**6**	**6**		
1.1	Введение в Python: основной синтаксис и среды разработки	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО4
1.2	Работа с функциями и файлами в Python	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО4
1.3	Основные библиотеки языка Python. Библиотеки numpy, pandas, matplotlib	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО4, РО2
2	**Основы теории вероятностей и математическая статистика**	**16**	**8**	**8**		
2.1	Случайные события и величины	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО1
2.2	Методы статистического описания результатов наблюдения	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО1, РО3
2.3	Статистическое оценивание параметров	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО1, РО3
2.4	Проверка статистических гипотез	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО1, РО3
3	**Анализ многомерных данных для решения прикладных задач**	**16**	**8**	**8**		
3.1	Первичная обработка больших данных	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО2, РО3, РО4
3.2	Восстановление регрессии	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО2, РО3, РО4

№ п/п	Наименование и содержание разделов и тем программы	Всего часов	В том числе:		Использование средств ЭО и ДОТ	Результаты обучения
			Контактная работа	Самостоятельная работа		
3.3	Классификация. Кластеризация	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО2, РО3, РО4
3.4	Ключевые задачи в подготовке датасетов и их важность	4	2	2	Система электронного обучения СФУ «е-Курсы»	РО2, РО3, РО4
4	**Итоговая аттестация**	**4**	**–**	**4**	Система электронного обучения СФУ «е-Курсы»	РО1–РО4
	Всего	**48**	**22**	**26**		

2.2. План учебной деятельности

Результаты обучения	Учебные действия/ формы текущего контроля	Используемые ресурсы/ инструменты/ технологии
РО1. Знать методы структурирования наборов данных	Тестовые задания	Электронный курс: https://e.sfu-kras.ru/course/view.php?id=33803
РО2. Проводить очистку данных цифрового следа (поиск аномалий, корректировка, подсказка, автоматизация/уменьшение объема ручной работы, поиск дубликатов) записей и пользователей	Задания по предварительной обработке данных	Anaconda Individual Edition with jupyter notebook, Google colab
РО3. Проводить сравнительный анализ для проверки гипотез, представленных в модели деятельности человека (группы людей) и ИКС, относящиеся к различным прикладным областям	Кейс на построение гипотез и применение методов обработки данных	Электронный курс: https://e.sfu-kras.ru/course/view.php?id=33803
РО4. Применять библиотеки и фреймворки языка программирования Python для визуализации данных	Задания на визуализацию данных с использованием библиотек языка программирования Python	Anaconda Individual Edition with jupyter notebook, Google colab

2.3. Виды и содержание самостоятельной работы

Выполнение самостоятельной работы слушателями предполагается в дистанционном режиме в рамках электронного курса, размещенного в системе электронного обучения. В дополнение к синхронным занятиям, слушателями самостоятельно изучаются представленные теоретические материалы в форме презентаций, обучающих ноутбуков с кодом, сохранённых видеозаписей лекций и в текстовом варианте. Также слушатели самостоятельно проводят анализ и систематизацию материала в рамках выполнения практических заданий. Для оценки уровня усвоения изученного учебного материала, слушатели проходят контрольные тесты и задания.

III. УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ

3.1. Учебно-методическое обеспечение, в т.ч. электронные ресурсы в корпоративной сети СФУ и сети Интернет

1. Бурков А. Машинное обучение без лишних слов. – СПб: Питер, 2020. – 192 с.
2. Грас Д. Data Science. Наука о данных с нуля: пер. с англ. – СПб.: БХВ-Петербург, 2017. – 336 с.
3. Джоши П. Искусственный интеллект с примерами на Python. – СПб., 2019. – 448 с.
4. Кукарцев, Владислав Викторович. Теория баз данных: учебник / В.В Кукарцев, Р.Ю. Царев, О.А. Антамошкин; Сиб. федер. ун-т, Ин-т космич. И информ. технологий. – Красноярск: СФУ, 2017. – 178 с.
5. Маккинли У. Python и анализ данных. – Саратов: Профобразование, 2019. – 482 с.
6. Мерков А.Б. Распознавание образов: введение в методы статистического обучения. 2-е изд., испр. – М., 2019 – 256 с.
7. Мирджалили В., Рашка С. Python и машинное обучение. Машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow. – М. – СПб., 2020. – 848 с.
8. Мыльников Л.А. Статистические методы интеллектуального анализа данных. – СПб.: БХВ-Петербург, 2021. – 240 с.
9. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. – М, 2017. – 480 с.
10. Плас Дж. В. Python для сложных задач: наука о данных и машинное обучение. – СПб.: Петербург, 2018. – 576 с.
11. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data, Python и наука о данных. – СПб: Питер, 2017. – 336 с.
12. Скиена С. Наука о данных: учебный курс / Пер. с англ. – СПб.: ООО «Диалектика», 2020. – 544 с.
13. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / Пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.

14. Хасти Т., Тибришани Р. Основы статистического обучения: интеллектуальный анализ данных, логический вывод и прогнозирование. – М. – СПб., 2020. – 768 с.
15. Чубукова И.А. Data Mining: учеб. пособие. – 2-е изд., испр. – М.: БИНОМ. Лаб. знаний, 2008. – 382 с.
16. Элбон К. Машинное обучение с использованием Python. Сборник рецептов. – СПб.: БХВ-Петербург, 2020. – 384 с.

3.2. Программное обеспечение (информационные обучающие системы, системы вебинаров, сетевые ресурсы хостинга видео, изображений, файлов, презентаций и др.)

1. Электронный ресурс «Аналитика больших данных». – Режим доступа: https://e.sfu-kras.ru/course/view.php?id=33803.
2. Интернет-браузер, обновленный до последней версии: Google Chrome, Opera, Microsoft Edge, Safari, Mozilla FireFox.
3. Anaconda Individual Edition (with Python 3.7 и выше) с jupyter notebook.
4. Google colab: https://colab.research.google.com/.
5. Приложения для связи: Zoom, Telegram.

IV. ОЦЕНКА КАЧЕСТВА ОСВОЕНИЯ ПРОГРАММЫ

4.1. Формы аттестации, оценочные материалы, методические материалы

Обучение на программе повышения квалификации заключается в поэтапном выполнении текущих заданий, размещенных в электронном курсе «Аналитика больших данных» и предполагающих применение соответствующих онлайн- и оффлайн-сервисов.

Текущий контроль осуществляется путем опроса, тестирования, выполнения заданий по темам курса в дистанционном режиме.

Все методические материалы и рекомендации представлены в электронном курсе «Аналитика больших данных».

Примеры тестовых заданий

Разделы «Основы теории вероятностей» и «Математическая статистика»

1. При проверке статистической гипотезы, ошибка первого рода — это:
   1) принятие нулевой гипотезы, которая в действительности является неверной;
   2) отклонение альтернативной гипотезы, которая в действительности является верной;
   3) принятие альтернативной гипотезы, которая в действительности является неверной;
   4) **отклонение нулевой гипотезы, которая в действительности является верной.**

2. При увеличении объема выборки n и одном и том же уровне значимости ширина доверительного интервала

1) может как уменьшиться, так и увеличиться;
2) **уменьшается;**
3) не изменяется;
4) увеличивается.

3. Какие из названных распределений используются при проверке гипотезы о числовом значении математического ожидания при неизвестной дисперсии?

1) **распределение Стьюдента;**
2) распределение Фишера;
3) нормальное распределение;
4) распределение хи-квадрат.

Примеры практических заданий

Раздел «Основы языка Python»

Задание:

1. Сгенерировать массивы numpy, распределенные по следующим законам распределения: равномерное, нормальное, экспоненциальное.
2. Посчитать среднее значение (mean), максимальное (max), минимальное (min), стандартное отклонение (std).
3. Сравнить результаты при создании массива из 3, 10, 100, 1000, 1000000 элементов.

В качестве ответа на дополнительное задание загрузить файл с кодом или ссылку на Google Colab.

Раздел «Анализ многомерных данных»
Тема «Предварительная обработка данных»

Задание:

1. Загрузить исходные данные из файла в память в виде объекта Pandas.DataFrame.
2. Построить визуальное представление для каждого столбца (признака) в исходном наборе данных.
3. Провести проверку правдоподобности исходных данных (проверка типов исходных данных, лишних пропусков, невозможных значений и др.).
4. Провести поиск значений в исходном наборе данных, резко отличающихся от других значений (выбросов).
5. Провести поиск пропущенных значений в исходных данных.
6. Привести числовые признаки к стандартному виду. Для категориальных признаков выполнить их кодировку.

4.2. Требования и содержание итоговой аттестации

Основанием для аттестации является выполнение не менее 60 % заданий, размещенных в электронном курсе «Аналитика больших данных».

Итоговая аттестация по программе представляет собой решение кейса, включающего построение гипотез по собственным данным, а также их описание и обработку с использованием инструментов, рассмотренных в курсе.

Программу составили:

Доктор технических наук,
профессор кафедры информатики Института космических
и информационных технологий СФУ,
заведующий лаборатории искусственного интеллекта О.А. Антамошкин

Инженер-исследователь лаборатории
искусственного интеллекта Департамента науки
и инновационной деятельности СФУ А.К. Сомов

Старший преподаватель
научно-учебной лаборатории программного
обеспечения Института космических
и информационных технологий СФУ А.С. Михалев

Ассистент
научно-учебной лаборатории программного
обеспечения Института космических
и информационных технологий СФУ Е.О. Пересунько

Старший преподаватель
научно-учебной лаборатории программного
обеспечения Института космических
и информационных технологий СФУ П.В. Пересунько

Руководитель программы:

Доктор технических наук,
профессор кафедры информатики Института космических
и информационных и технологий,
заведующий лаборатории искусственного интеллекта О.А. Антамошкин