

МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технологический университет «СТАНКИН»
(ФГБОУ ВО «МГТУ «СТАНКИН»)

УТВЕРЖДАЮ

Проректор по
образовательной деятельности и
молодежной политике

«29» 02.2021

М.В. Бильчук



ДОПОЛНИТЕЛЬНАЯ ПРОФЕССИОНАЛЬНАЯ ПРОГРАММА
(повышение квалификации)

Большие данные: проектная деятельность учащихся IT классов

Направление программы: проект
«IT-класс в московской школе»

Разработчик:
Нежметдинов Рамиль Амирович

Москва
2021

Раздел 1. Характеристика программы

1.1. Цель реализации программы – совершенствование профессиональных компетенций, обучающихся в области проектной деятельности учащихся IT классов при работе с большими данными.

В процессе обучения слушатели научатся обеспечивать информационно-методическое сопровождение организации учебного процесса, в том числе с применением дистанционных технологий, используя образовательные ресурсы Интернет-пространства.

Совершенствуемые компетенции

№	Компетенция	Направление подготовки 44.03.01 Педагогическое образование
		Код компетенции
1.	Способен организовывать совместную и индивидуальную учебную и воспитательную деятельность обучающихся, в том числе с особыми образовательными потребностями, в соответствии с требованиями федеральных государственных образовательных стандартов	ОПК – 3
2.	Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	ОПК – 9

1.2. Планируемые результаты обучения

№	Уметь-знать	Направление подготовки 44.03.01 Педагогическое образование
		Код компетенции
1.	Уметь: - разрабатывать программные решения по работе с данными используя основные конструкции языка Python. - разрабатывать программы на языке Python с использованием библиотеки pandas для работы с массивами данных. - проектировать решения с использованием метрика north star (NSM, метрика «полярной звезды»). - реализовывать задачи классификации с применением тензорных вычислений. - разрабатывать решения статистического анализа с использованием модуля Scipy.Stats. Знать: - алгоритмы применения основных конструкций языка	ОПК – 9

	Python; - стандартные библиотеки языка Python и особенности работы с массивами данных на языке Python; - методику применения метрики north star (NSM, метрика «полярной звезды»); - методику применения тензорных вычислений при работе с большими данными; - особенности реализации программ статистического анализа на языке Python с использованием модуля Scipy.Stats.	
2.	Уметь: разрабатывать план организации проектной деятельности учащихся IT классов в области работы с большими данными Знать: - особенности проектной деятельности учащихся IT классов в области работы с большими данными; - алгоритм разработки плана организации проектной деятельности учащихся IT классов в области работы с большими данными	ОПК-3, ОПК – 9

1.3. Категория обучающихся: уровень образования – ВО, область профессиональной деятельности – обучение информационным технологиям в образовательных организациях.

1.4. Программа может быть реализована с применением дистанционных образовательных технологий. Форма обучения очная.

1.5. Режим занятий: доступ к образовательной платформе организации круглосуточно при соблюдении установленных сроков обучения.

1.6. Трудоемкость программы: 36 часов.

Раздел 2. Содержание программы

2.1. Учебный (тематический) план

№	Наименование разделов (модулей) и тем	Трудоемкость и виды занятий			Формы контроля
		Трудоемкость	Лекции, презентации	Практические занятия	
1.	Большие данные в современном мире	2	2	0	
2.	Python для анализа данных	6	2	4	Тест № 1

3.	Работа с массивами данных в Python	6	2	4	
4.	Метрики данных	4	2	2	Тест № 2
5.	Машинное обучение при аналитике больших данных	6	2	4	
6.	Статистический анализ в Python	4	2	2	Тест № 3
7.	Проектная деятельность учащихся IT классов в области работы с большими данными	4	2	2	Проект
8.	Итоговая аттестация	4	-	4	Зачет на основании совокупности выполненных работ и защита проекта
	Итого	36	14	22	

2.2. Учебная программа

№ п/п Наименование разделов (модулей) и тем	Виды учебных занятий/работ, час.	Содержание
Тема 1. Большие данные в современном мире	Лекции с элементами дискуссии – 2 ч.	Что такое большие данные. Сферы применения больших данных. Перспективы использования в новых сферах. Основные этапы развития технологии сбора и анализа больших данных. Уровень развития современных аналитических систем. Облачные вычисления и их связь с большими данными.
Тема 2. Python для анализа данных	Лекция – 2 ч.	Стандартные библиотеки Python. Функции взаимодействия с операционной системой. Аргументы командной строки. Регулярные выражения. Базовые математические функции, случайные числа и статистика в Python. Модули доступа в глобальную сеть Интернет. Модуль datetime, управление параметрами даты и времени. Сжатие данных. Измерение производительности. Многопоточность в Python. Алгоритм применения основных конструкций языка Python по работе с данными.
	Практическое занятие – 4 ч.	Разработка программ по работе с данными используя основные конструкции языка Python. Рассматриваемые вопросы: кортежи,

		списки и словари в Python; работа с массивами данных; циклы; цикл for; функция range(); работа с классами; экземпляры классов; наследование; указатели и словари классов; модули в Python; файловый ввод/вывод; функции работы с файлами: open(), tell(), readline(), readlines(), write() и close(); обработка исключений. Тест №1 Практическая работа №1 Разработка программного решения по работе с данными, используя основные конструкции языка Python (условия задаются преподавателем)
Тема 3. Работа с массивами данных в Python	Лекция – 2 ч.	Работа с массивами данных. Создание и инициализация массивов, базовые операции над массивами; индексы, срезы, итерации; изменение формы (shape) массива; объединение массивов; разбиение массива; копии и представления. Алгоритм разработки программ на языке Python с использованием библиотеки pandas для работы с массивами данных.
	Практическое занятие – 4 ч.	Разработка программ на языке Python с использованием библиотеки pandas для работы с массивами данных. Рассматриваемые вопросы: структура данных Series; создание Series из списка Python; создание Series из ndarray массива из numpy; создание Series из словаря (dict); создание Series с использованием константы; работа с элементами Series; структура данных DataFrame; создание DataFrame из словаря; создание DataFrame из списка словарей; создание DataFrame из двумерного массива; работа с элементами DataFrame; доступ к данным в структурах pandas; получение случайного набора из структур pandas; индексация с использованием логических выражений; использование isin для работы с данными в pandas. Практическая работа №2 Разработка программы на языке Python с использованием библиотеки pandas для работы с массивами данных (условия задаются преподавателем)
Тема 4. Метрики данных	Лекция – 2 ч.	Использование библиотеки SQLite. Формирование запросов к данным. Создание таблиц. Типы ячеек, поддерживаемые в SQLite. Выбор минимальных и максимальных значений. Отбор по пороговым значениям.

		Визуализация отобранных данных. Формирование гистограмм по запросам. Алгоритм разработки программ на языке Python по методике применения метрики north star (NSM, метрика «полярной звезды»)
	Практическое занятие – 2 ч.	Проектирование и разработка решений с использованием метрика north star (NSM, метрика «полярной звезды»). Практический пример реализации метрик в розничной торговле: ежемесячный доход, ежемесячная активность клиентов, ежемесячное количество заказов, средний доход на один заказ, коэффициент активности новых клиентов, скорость притока/оттока клиентов. Тест №2. Практическая работа №3 Разработка программы на языке Python по методике применения метрики north star (NSM, метрика «полярной звезды», условия задаются преподавателем).
Тема 5. Машинное обучение при анализе больших данных	Лекция – 2 ч.	Изучение основ работы с библиотеками TensorFlow и Keras. Среды тензорных вычислений. Решение математических задач, анализ трендов с применением TensorFlow Keras. Функции активации, виды функций активации. Слои в нейронных сетях. Обучение нейронных сетей, «эпохи» обучения. Сравнение предсказанных результатов с реальными. Средне квадратичная ошибка. Алгоритм разработки программ на языке Python по методике применения тензорных вычислений с использованием библиотек TensorFlow и Keras.
	Практическое занятие – 4 ч.	Реализация задачи классификации на примере распознавания рукописных цифр с использованием датасета MNIST и библиотеки TensorFlow, Keras. Практическая работа №4 Разработка программы на языке Python по методике применения тензорных вычислений с применением библиотеки Tensor Flow, Keras. (условия задаются преподавателем).
Тема 6. Статистический анализ в Python	Лекция – 2 ч.	Статистические функции в библиотеке numpy. Нахождение максимума и минимума массива в NumPy. Нахождение среднего, медианы, стандартного отклонения и дисперсии в NumPy. Функция NumPy Average. Гармоническое Среднее. Z-статистика и p-value. Алгоритм разработки программ на языке Python для статистических вычислений с

		использованием модуля Scipy.Stats.
	Практическое занятие – 2 ч.	Разработка решений статистического анализа с использованием модуля Scipy.Stats. Случайные переменные. Генерация случайных чисел. Сдвиг и масштабирование. Параметры формы. Тест №3. Практическая работа №5 Разработка программы на языке Python для статистических вычислений с использованием модуля Scipy.Stats (условия задаются преподавателем).
Тема 7. Проектная деятельность учащихся IT классов в области работы с большими данными	Лекция – 2 ч.	Особенности проектной деятельности учащихся IT классов в области сбора и анализа больших данных Алгоритм разработки плана организации проектной деятельности учащихся IT классов в области сбора и анализа больших данных
	Практическое занятие – 2 ч.	Проект Выбор проекта осуществляет обучающийся и разрабатывает его на основе полученных знаний и умений. Проект 1.1 Разработка приложения, для распознавание рукописных цифр на основе датасета MNIST. Проект 1.2 Разработка приложения для определения кредитной платежеспособности (кредитного скрининга)
Итоговая аттестация	Зачет – 4 ч.	Зачет на основании совокупности выполненных работ и защита проекта.

Раздел 3. Формы аттестации и оценочные материалы

3.1. Текущий контроль осуществляется в форме: тестирования с автоматической проверкой и выполнения практических работ.

Варианты тестовых вопросов

Отметка «зачтено» выставляется при правильном выполнении не менее 50% заданий теста.

Тест №1 – Большие данные в современном мире. Python для анализа данных.

Какой язык программирования наиболее популярен при разработке приложений для анализ данных и нейронных сетей?

- A. C
- B. Java
- C. Ruby
- D. Python +

К какому типу языков программирования относится язык Python?

- A. Интерпретируемые +
- B. Компилируемые
- C. Невыполнимые
- D. Исключительные

Что в Python начинается с ключевого слова def?

- A. Переменная
- B. Функция +
- C. Массив
- D. Начало программы

Что делает оператор import в Python?

- A. Создает CSV файл
- B. Компилирует код
- C. Подключает библиотеки +
- D. Очищает кэш

Популярная библиотека Python для обработки и анализа данных

- A. Numeric
- B. Tensorflow
- C. NumPy
- D. Pandas +

Тест №2 – Работа с массивами данных в Python. Метрики данных.

Популярная библиотека Python для числовых вычислений

- E. Numeric
- F. Sum
- G. Tensorflow
- H. NumPy +

[1, 3, 10, 15] что это за структура данных?

- A. Список
- B. Строка
- C. Массив +
- D. Простое число

К какой задаче относится задача распределения данных по параметрам?

- A. Распознавание
- B. Классификация +
- C. Предсказание

Какой термин задан следующим определением «процентная величина, отражающая расхождение между ожидаемым и полученным ответами»?

- A. Процент расхождения
- B. Ошибка +
- C. Качество нейронной сети

За что отвечает функция Average библиотеки NumPy?

- A. Вычисляет средневзвешенное значение элементов в массиве +
- B. Находит элемент массива с минимальным значением
- C. Вычисляет сумму элементов массива
- D. Вычисляет произведение элементов массива

**Тест №3 – Машинное обучение при аналитике больших данных.
Статистический анализ в Python.**

Что из перечисленного ниже не является типовой задачей нейронных сетей?

- A. Классификация
- B. Прогнозирование
- C. Распознавание
- D. Математические вычисления +

Как называются нейронные сети с большим количеством слоев?

- A. Глубокие +
- B. Длинные
- C. Широкие
- D. Масштабные

Как называется самый популярный набор данных рукописных цифр?

- A. OIST
- B. GIST
- C. SKLAD
- D. MNIST +

На чём базируется библиотека tensorflow?

- A. Техническая поддержка
- B. Тензоры +
- C. Понятный интерфейс
- D. Кроссплатформенность

Как называется конечный набор входных сигналов (иногда вместе с правильными выходными сигналами), по которым происходит обучение сети?

- A. Тестовая выборка
- B. Эталонная выборка
- C. Обучающая выборка +

3.2. Промежуточная аттестация

Проект – реализация ИТ решения на базе выбранной технологии (выбор осуществляет учитель).

Темы проекта:

1. Создать сеть распознавания рукописных цифр

Создайте собственное консольное/декстопное/веб приложение, для распознавание рукописных цифр на основе датасета MNIST. Датасет можно скачать по ссылке: <http://yann.lecun.com/exdb/mnist/>

Необходимо реализовать следующие возможности:

1) сделать модуль, для предварительной подготовки изображения перед распознаванием

2) нужно обучить сеть распознавать цифры

Рекомендуемый язык программирования для реализации проекта – Python.

Пояснение: MNIST - это набор данных, разработанный Янн ЛеКун, Коринной Кортес и Кристофером Берджесом для оценки моделей машинного обучения по задаче классификации рукописных цифр. Набор данных был составлен из нескольких наборов данных отсканированных документов. Изображения цифр были взяты из различных отсканированных документов, нормализованы по размеру и центрированы. Каждое изображение представляет собой квадрат 28 на 28 пикселей (всего 784 пикселя). Стандартный набор данных используется для оценки и сравнения моделей, где 60000 изображений используются для обучения модели, а отдельный набор из 10 000 изображений используется для ее проверки. Это задача распознавания цифр. Таким образом, есть 10 цифр (от 0 до 9) или 10 классов для прогнозирования.

Форма представления результатов

- Запускаемое приложение (*.exe) или ссылка на web-ресурс
- Исходный код программного продукта в виде текстового файла (файл *.txt, *.c, *.cpp, *.h, *.hxx или иного расширения; должен быть указан язык программирования)

- Отчет (*.doc/ *.docx / *.pdf)
- Презентация(*.ppt / *.pptx/ *.pdf)
- Видео с демонстрацией работы программного кода (до 1 мин.)

2. Создать сеть прогнозирования кредитного скоринга

Создайте собственное консольное/десктопное/веб приложение для определения кредитной платежеспособности (кредитного скрининга). Пусть имеются данные о клиентах, обратившихся за кредитом. Здесь объектами являются клиенты, а признаками могут являться их пол, уровень дохода, образование, информация о том, имеется или отсутствует у них положительная кредитная история и т.д.

В качестве выделенного признака (ответа) выступает информация о том, вернул ли клиент кредит в банк или нет. По этим данным требуется научиться предсказывать, вернет ли кредит новый клиент, обратившийся в банк.

Таким образом, речь идет о задаче классификации: требуется определить, какому классу: положительному (кредит будет возвращен) или отрицательному (кредит не будет возвращен) – принадлежит клиент.

В качестве модельных данных необходимо использовать данные Credit Approval Data Set (<https://archive.ics.uci.edu/ml/datasets/Credit+Approval>) из коллекции UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.html>).

Данные имеют формат csv и их можно использовать следующим образом (на языке Python):

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/credit
screening/crx.data' data = pd.read_csv(url, header=None, na_values='?')
```

Форма представления результатов

- Запускаемое приложение (*.exe) или ссылка на web-ресурс
- Исходный код программного продукта в виде текстового файла (файл *.txt, *.c, *.cpp, *.h, *.hxx или иного расширения; должен быть указан язык программирования)

- Отчет (*.doc/ *.docx / *.pdf)
- Презентация(*.ppt / *.pptx/ *.pdf)
- Видео с демонстрацией работы программного кода (до 1 мин.)

Требования к проекту: реализация проекта осуществляется на базе предложенных программных решений, в рамках выбранной технологии. Проект должен содержать пояснительную записку на 2-3 страницах, которая должна отражать тему проекта, цель проекта, описание освоенной технологии и выбранного программного решения, а также пояснение предложенного решения.

Критерии оценивания: все требования к проекту выполнены.

Оценивание: зачет/незачет.

3.3. Итоговая аттестация: Зачет на основании совокупности выполненных работ и защита проекта.

Критерии оценивания:

1. Программная реализация проекта соответствует поставленным целям и задачам.
2. Проект содержит пояснительную записку с описанием предложенного решения.
3. По проекту подготовлена презентация решения.

Раздел 4. Организационно-педагогические условия реализации программы

4.1. Учебно-методическое и информационное обеспечение программы

Электронные ресурсы:

1. Московская электронная школа. Информационно-методическая поддержка проекта [Электронный ресурс]// URL: <http://mes.mosedu.ru/> (дата обращения: 07.04.2020).
2. Общегородская платформа электронных образовательных материалов [Электронный ресурс]. – режим доступа: <https://uchebnik.mos.ru/> – (дата обращения: 07.04.2020).
3. Нейронные сети для начинающих. Часть 1. [Электронный ресурс]// URL: <https://habr.com/ru/post/312450/>– (дата обращения: 24.05.2020).
4. Нейронные сети [Электронный ресурс]// URL: <https://postnauka.ru/video/99268> – (дата обращения: 24.05.2020).
5. Как работает нейронная сеть: алгоритмы, обучение, функции активации и потери [Электронный ресурс]// URL: <https://neurohive.io/ru/osnovy-data-science/osnovy-nejronnyh-setej-algoritmy-obuchenie-funkcii-aktivacii-i-poteri/> – (дата обращения: 24.05.2020).
7. Лекция «Задачи машинного обучения» [Электронный ресурс] // <http://www.machinelearning.ru/wiki/images/a/a1/BayesML-2009-1.pdf>
8. Введение в машинное обучение [Электронный ресурс] // <https://habr.com/ru/post/448892/>
9. Задачи машинного обучения в ML.NET [Электронный ресурс] // <https://docs.microsoft.com/ru-ru/dotnet/machine-learning/resources/tasks>

4.2. Список литературы

1. Вайгенд Андреас BIG DATA. Вся технология в одной книге / Вайгенд Андреас – Москва: Эксмо., 2017. – 480 с.: ISBN 978-5-04-094117-9
2. Фрэнкс Билл «Укрощение больших данных» / Билл Фрэнкс – Москва: «Манн, Иванов и Фербер»., 2014. – 340 с.: ISBN 978-5-00057-146-0
3. Майер-Шенбергер Виктор «Большие данные» / Виктор Майер-Шенбергер, Кеннет Кукьер – Москва: АСТ., 2014 - 450 с.: ISBN: 978-5-17-104034-

4. Рашка Себастьян «Python и машинное обучение» / Рашка Себастьян – Москва: ДМК Пресс., 2017. – 420 с.: ISBN: 978-5-97060-409-0
5. Кухарев Г. А. «Методы обработки и распознавания изображений лиц в задачах биометрии» / Кухарев Г.А., Каменская Е.И., Матвеев Ю.Н., Щеголева Н.Л. – Санкт-Петербург: Политехника., 2013. – 388 с.: ISBN: 978-5-7325-1028-7
6. Флах Петер «Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных» / Флах Петер – Москва: ДМК Пресс., 2017. – 402 с.: ISBN: 978-5-97060-273-7.
7. Благирев Алексей Big data простым языком» / Благирев Алексей, Хапаева Наталья – Москва: АСТ., 2019. – 250 с.: ISBN: 978-5-17-111829-7.
8. Сенько Александр Работа с BigData в облаках / Александр Сенько – Санкт-Петербург: Издательство «Питер»., 2019. – 447 с.: ISBN: 978-5-4461-0578-6.
9. Анналин Ын Теоретический минимум по Big Data / Анналин Ын, Кеннет Су - Санкт-Петербург: Издательство «Питер»., 2019. – 447 с.: ISBN: 978-5-4461-0578-6.
10. Андреас Мюллер Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными / Андреас Мюллер, Сара Гвидо – Москва: Диалектика., 2017. – 480 с.: ISBN: 978-5-9908910-8-1
11. Джефф Танг Умные мобильные проекты с TensorFlow / Джефф Танг – Москва: ДМК-Пресс, 2019, 386 с.: ISBN: 978-5-97060-706-0

4.3. Материально-технические условия реализации программы

Для успешной реализации программы необходимы: компьютерное и мультимедийное оборудование для использования видео- и аудиовизуальных средств обучения с подключением к сети Интернет, пакет слайдовых презентаций, видеоролики (по темам учебной программы).

Кадровое обеспечение программы

Программа реализуется сотрудниками и преподавателями «МГТУ «СТАНКИН».