|  |  |
| --- | --- |
|  | Утверждаю:  Проректор по учебной работе  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Т.А. Ольховая  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_20\_\_ г. |

1. **Паспорт Образовательной программы**

**«**Основы анализа данных и машинного обучения**»**

|  |  |
| --- | --- |
| **Версия программы** | 1 |
| **Дата Версии** | 12**.**10.2020 |

1. **Сведения о Провайдере**

|  |  |  |
| --- | --- | --- |
| 1.1 | Провайдер | федеральное государственное бюджетное образовательное учреждение высшего образования «Оренбургский государственный университет» |
| 1.2 | Логотип образовательной организации |  |
| 1.3 | Провайдер ИНН | 5612001360 |
| 1.4 | Ответственный за программу ФИО | Шухман Александр Евгеньевич |
| 1.5 | Ответственный должность | Заведующий кафедрой геометрии и компьютерных наук |
| 1.6 | Ответственный Телефон | +79033987498 |
| 1.7 | Ответственный Е-mail | **s**hukhman@gmail.com |

1. **Основные Данные**

|  |  |  |
| --- | --- | --- |
| **№** | **Название** | **Описание** |
| 2.1 | Название программы | Основы анализа данных и машинного обучения |
| 2.2 | Ссылка на страницу программы | <http://www.osu.ru/doc/5023> |
| 2.3 | Формат обучения | онлайн |
|  | Подтверждение от ОО наличия возможности реализации образовательной программы с применением электронного обучения и (или) дистанционных образовательных технологий с возможностью передачи данных в форме элементов цифрового следа | имеется  <https://moodle.osu.ru/course/view.php?id=11344> |
| 2.4 | Уровень сложности | базовый |
| 2.5 | Количество академических часов | **72** |
|  | Практикоориентированный характер образовательной программы: не менее 50 % трудоёмкости учебной деятельности отведено практическим занятиям и (или) выполнению практических заданий в режиме самостоятельной работы (кол-во академических часов) | 48 |
| 2.6 | Стоимость обучения одного обучающегося по образовательной программе, а также предоставление ссылок на 3 (три) аналогичные образовательные программы иных организаций, осуществляющих обучение, для оценки объективности стоимости или обоснование уникальности представленной образовательной программы в случае отсутствия аналогичных образовательных программ на рынке образовательных услуг | 20000 руб.  1. <https://skillfactory.ru/ml-programma-machine-learning-online/> 43900 руб.  2. <https://netology.ru/programs/machine-learn#/presentation> 49000 руб.  3. <https://python-school.ru/courses/pyml-introduction-to-machine-learning/> 36000 руб.  4. <https://rtf.urfu.ru/ob-institute/dpo/machine-learning/> 20000 руб. |
| 2.7 | Минимальное количество человек на курсе | 15 |
| 2.8 | Максимальное количество человек на курсе | 40 |
| 2.9 | Данные о количестве слушателей, ранее успешно прошедших обучение по образовательной программе | нет |
| 2.10 | Формы аттестации | Тестирование, решение практических заданий, выполнение итогового проекта |
|  | Указание на область реализации компетенций цифровой экономики, к которой в большей степени относится образовательная программа, в соответствии с Перечнем областей | Искусственный интеллект |

1. **Аннотация программы**

Образовательная программа рассчитана на специалистов, имеющих среднее или высшее профессиональное образование.

Востребованность программы определяется активным применением технологий искусственного интеллекта и машинного обучения в производственных и бизнес процессах самых разных компаний и промышленных производств и потребностью подготовки специалистов, владеющих технологиями анализа данных и машинного обучения.

Длительность: Общее количество часов составляет 72 часа. Из них лекционные – 24 часа, практические – 24 часа, самостоятельная практическая работа слушателя – 24 часа.

Основные разделы программы:

1. Основы языка программирования Python для анализа данных.

2. Базовые задачи и методы машинного обучения.

3. Современные нейросетевые архитектуры.

4. Автоматизация сбора данных и доступа к моделям машинного обучения.

Результаты обучения: Содержание курса ориентировано на подготовку специалистов, способных использовать современные методы анализа данных и машинного обучения, а также средства языка программирования Python для постановки и решения практических задач в различных предметных областях. Обучение проводится на основе выполнения практических заданий.

Стоимость: 20 000 руб.

Выдаваемый документ: Удостоверение о повышении квалификации

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ   
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ   
«ОРЕНБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Утверждаю,

проректор по учебной работе

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Т.А. Ольховая

м.п.

Дополнительная профессиональная программа повышения квалификации

«ОСНОВЫ АНАЛИЗА ДАННЫХ И МАШИННОГО ОБУЧЕНИЯ»

72 часа

Оренбург, 2020

**ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ**

**1.Цель программы**

Цель программы – формирование компетенции цифровой экономики «Способность эффективно применять современные методы и средства анализа данных и машинного обучения для решения прикладных задач» у квалифицированных специалистов различных областей экономики.

**2.Планируемые результаты обучения:**

2.1. Знание (осведомленность в областях)

2.1.1. Основы языка программирования Python.

2.1.2. Основные этапы анализа данных.

2.1.3. Основные типы задач анализа данных и машинного обучения

2.1.4. Математические основы и методы решения задач классификации, регрессии, кластеризации, основные метрики качества алгоритмов машинного обучения

2.1.5. Математические основы и возможности нейросетевых методов

2.1.6 Методы предобработки и векторизации текстов

2.1.7 Методы сбора информации из интернет-источников

2.1.8 Методы создания API для доступа к моделям машинного обучения.

2.2. Умение (способность к деятельности)

2.2.1. Использование языка программирования Python, интерактивных блокнотов Jupiter Notebooks с библиотеками Pandas и Matplotlib для автоматизированного сбора, предварительной обработки, очистки и визуализации данных, конструирования и отбора признаков.

2.2.2 Обучение и использование моделей из библиотеки SkLearn для решения задач. машинного обучения с оценкой качества и подбором оптимальных гиперпараметров моделей.

2.2.3. Обучение и использование нейросетевых моделей на основе библиотеки. Keras+Tensorflow для задач классификации изображений и текстов, в том числе с использованием переноса обучения.

2.2.4 Создание API для доступа к моделям на основе каркаса Flask.

2.3.Навыки (использование конкретных инструментов)

2.3.1 Эффективное решение прикладных задач на основе данных из различных источников с подбором оптимальных гиперпараметров моделей машинного обучения.

2.3.2 Предобработка и классификация текстов и изображений.

2.3.3 Создание API для доступа к разработанным моделям.

**3.Категория слушателей**

* 1. Образование – высшее или среднее профессиональное образование.
  2. Квалификация - нет требований.
  3. Наличие опыта профессиональной деятельности – нет требований.
  4. Предварительное освоение иных дисциплин/курсов /модулей – требуется владение основами цифровой грамотности и алгоритмизации в объеме школьного курса информатики.

**4.Учебный план программы «Основы анализа данных и машинного обучения»**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Модуль** | **Всего, час** | **Виды учебных занятий** | | |
| **лекции** | **практические занятия** | **самостоятельная работа** |
| 1 | Основы языка программирования Python для анализа данных. | 22 | 8 | 8 | 6 |
| 2 | Базовые задачи и методы машинного обучения. | 22 | 8 | 8 | 6 |
| 3 | Современные нейросетевые архитектуры. | 16 | 6 | 6 | 4 |
| 4 | Автоматизация сбора данных и доступа к моделям машинного обучения | 6 | 2 | 2 | 2 |
| **Итоговая аттестация** | | **6** | **Указывается вид (экзамен, зачёт, реферат и т.д.)** | | |
| Выполнение итогового проекта | | 6 | Слушатели выполняют проект, связанный с решением практической задачи анализа данных и включающий все этапы анализа от предварительной обработки до развертывания модели | | |

**5.Календарный план-график реализации образовательной** программы

(дата начала обучения – дата завершения обучения) в текущем календарном году,

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п/п** | **Наименование учебных модулей** | **Трудоёмкость (час)** | **Сроки обучения** |
| 1 | Основы языка программирования Python для анализа данных. | 22 | 2.11-7.11 |
| 2 | Базовые задачи и методы машинного обучения. | 22 | 9.11-14.11 |
| 3 | Современные нейросетевые архитектуры. | 16 | 16.11-19.11 |
| 4 | Автоматизация сбора данных и доступа к моделям машинного обучения | 6 | 20.11-21.11 |
| Выполнение итогового проекта | | 6 | 22.11-23.11 |
| Всего: | | 72 | 2.11-23.11 |

**6. Учебно-тематический план программы «Основы анализа данных и машинного обучения»**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Модуль / Тема** | **Всего, час** | **Виды учебных занятий** | | | **Формы контроля** |
| **лекции** | **практические занятия** | **самостоятельная работа** |
| 1 | Основы языка программирования Python для анализа данных. | 22 | 8 | 8 | 6 |  |
| 1.1 | Базовые конструкции и типы данных Python | 11 | 4 | 4 | 3 | Тест, практические задания |
| 1.2 | Работа с наборами данных и визуализация | 11 | 4 | 4 | 3 | Тест, практические задания |
| 2 | Базовые задачи и методы машинного обучения | 22 | 8 | 8 | 6 |  |
| 2.1 | Этапы и основные задачи анализа данных | 2 | 2 | - | - | Тест |
| 2.2 | Задача регрессии, основные методы и оценка качества моделей | 6 | 2 | 2 | 2 | Тест, практические задания |
| 2.3 | Задача классификации, базовые методы, оценка качества, ансамблевые методы | 8 | 2 | 4 | 2 | Тест, практические задания |
| 2.4 | Задача кластеризации, основные методы и оценка качества | 6 | 2 | 2 | 2 | Тест, практические задания |
| 3 | Современные нейросетевые архитектуры | 16 | 6 | 6 | 4 |  |
| 3.1 | Искусственные нейронные сети. Полносвязные сети. Обучение нейросетей | 4 | 2 | 2 | - | Тест |
| 3.2 | Сверточные нейронные сети. Классификация изображений. | 6 | 2 | 2 | 2 | Тест, практические задания |
| 3.3 | Нейросети в обработке текстов. Предобработка, векторизация, классификация текстов | 6 | 2 | 2 | 2 | Тест, практические задания |
| 4 | Автоматизация сбора данных и доступа к моделям машинного обучения | 6 | 2 | 2 | 2 |  |
| 4.1 | Автоматизация сбора данных и доступа к моделям машинного обучения | 6 | 2 | 2 | 2 | Тест, практические задания |
|  | Выполнение итогового проекта | 6 |  |  | 6 | Проектное задание |
|  | **Итого** | **72** | **24** | **24** | **24** |  |

**7. Учебная (рабочая) программа повышения квалификации «Основы анализа данных и машинного обучения»**

Модуль 1. Основы языка программирования Python для анализа данных (22 часа)

Тема 1.1 Базовые конструкции и типы данных Python (11 часов)

Язык программирования Python. Переменные, операции, выражения. Целые, действительные числа. Условия, циклы. Строки. Списки. Словари. Описание функций, создание и использование классов. Отладка программ. Использование интерактивных блокнотов Jupiter Notebooks.

Тема 1.2 Работа с наборами данных и визуализация (11 часов)

Библиотека NumPy. Операции с векторами и матрицами. Библиотека Pandas. Операции с наборами данных. Загрузка и обработка данных. Конструирование признаков. Фильтрация и сортировка данных. Визуализация данных, основные виды графиков и диаграмм. Библиотекb Matplotlib, Seaborn.

Модуль 2. Базовые задачи и методы машинного обучения (22 часа)

Тема 2.1 Этапы и основные задачи анализа данных (6 часов)

Этапы анализа данных. Виды признаков. Анализ данных и машинное обучение. Обучение с учителем, без учителя, с подкреплением. Основные задачи машинного обучения. Принципы оценки качества моделей машинного обучения. Библиотека SkLearn.

Тема 2.2 Задача регрессии, основные методы и оценка качества моделей (6 часов)

Задача регрессии. Линейная и полиномиальная регрессия. Функции ошибок. Переобучение и регуляризация.

Тема 2.3 Задача классификации, базовые методы, оценка качества, ансамблевые методы (8 часов)

Задача классификации. Метрики качества классификации. Метод к ближайших соседей. Логистическая регрессия. Метод опорных векторов. Решающие деревья. Ансамблевые методы. Случайный лес. Градиентный бустинг. Подбор оптимальных гиперпараметров.

Тема 2.4 Задача кластеризации, основные методы и оценка качества (6 часов)

Задача кластеризации. Метрики кластеризации. Метод к-средних. Метод DBSCAN. Метод t-SNE.

Модуль 3. Современные нейросетевые архитектуры (16 часов)

Тема 3.1 Искусственные нейронные сети. Многослойный персептрон. Обучение нейросетей (4 часа)

Искусственный нейрон. Функции активации. Многослойный персептрон. Обучение нейросетей. Классификация с помощью многослойного персептрона. Библиотека Keras+Tensorflow.

Тема 3.2 Сверточные нейронные сети. Классификация изображений. (6 часов)

Сверточные нейронные сети. Глубокое обучение. Перенос обучения. Современные архитектуры нейросетей для компьютерного зрения.

Тема 3.3 Нейросети в обработке текстов. Предобработка, векторизация, классификация текстов. (6 часов)

Предобработка, векторизация текстов. Классификация текстов. Рекуррентные нейронные сети для классификации текстов. Нейросетевые методы обработки текстов.

Модуль 4. Автоматизация сбора данных и доступа к моделям машинного обучения (6 часов)

Тема 4.1 Автоматизация сбора данных и доступа к моделям машинного обучения (6 часов)

Автоматизация загрузки данных из сети. Разбор HTMl, поиск и выделение информации с веб-страниц. Доступ к данным через API. Разбор XML, JSon. Создание API с помощью фреймворка Flask.

**Описание практико-ориентированных заданий и кейсов**

|  |  |  |
| --- | --- | --- |
| **Номер темы/модуля** | **Наименование практического занятия** | **Описание** |
| 1.1 | Основы Python | Изучение основных конструкций и типов данных языка Python, создание функций и классов |
| 1.2 | Анализ и визуализация данных | Изучение основных возможностей Pandas и Maplotlib |
| 2.2 | Линейная регрессия | Изучение методов линейной регрессии, улучшение моделей. |
| 2.3 | Методы классификации | Изучение методов и метрик классификации, подбор оптимальных гиперпараметров моделец |
| 2.4 | Методы кластеризации | Изучение методов и метрик кластеризации |
| 3.1 | Нейросети. Классификация с помощью нейросетей | Изучение классификации с помощью многослойного персептрона |
| 3.2 | Классификация изображений. Перенос обучения. | Изучение классификации изображений на основе обученных моделей с использованием переноса обучения |
| 3.3 | Предобработка, векторизация и классификация текстов | Исследование методов предобработки, векторизации и классификации текстов |
| 4.1 | Доступ к API и создание собственных API | Сбор данных через API и создание собственных API для доступа к моделям |

8.Оценочные материалы по образовательной программе

8.1. Вопросы тестирования по модулям

|  |  |  |  |
| --- | --- | --- | --- |
| № модуля | Вопросы входного тестирования | Вопросы промежуточного тестирования | Вопросы итогового тестирования |
| 1 | Входное тестирование включает 10 вопросов на оценку уровня цифровой грамотности слушателя | Тесты включают вопросы по конструкциям языка Python и библиотекам анализа и визуализации | Тест включает вопросы по результатам выполнения самостоятельных практических заданий |
| 2 | - | Тесты включают вопросы по задачам и методам анализа данных и машинного обучения | Тест включает вопросы по результатам выполнения самостоятельных практических заданий |
| 3 | - | Тесты включают вопросы по нейросетевым методам | Тест включает вопросы по результатам выполнения самостоятельных практических заданий |
|  | - | - | Тест включает вопросы по результатам выполнения самостоятельных практических заданий |

8.2. Описание показателей и критериев оценивания, шкалы оценивания.

По каждому модулю курса в автоматизированной системе управления обучением Moodle есть тесты двух видов. Промежуточные тесты выполняются после прослушивания лекций и оценивают понимание теоретического материала. Тест содержит от 4 до 6 заданий с выбором варианта ответа или вводом ответа.

Для оценки умений используются практические задания на использование конкретных методов анализа данных и машинного обучения, которые необходимо самостоятельно выполнить в системе Jupiter или Google Colaboratory. Практические задания разработаны в интерактивных блокнотах, которые содержат формулировку заданий и место для вставки кода решения. Практические задания проверяются автоматизированно с помощью итоговых тестов, вопросы которых соответствуют ответам на каждый пункт задания. Кроме того, практические задания прикрепляются в систему Moodle и анализируются преподавателем. По каждому тесту вычисляется средний процент правильных ответов.

Для оценки навыков используется итоговое задание, включающее в себя все этапы решения прикладной задачи – от сбора и предобработки данных до создания API к обученной модели. Итоговое задание прикрепляется в систему Moodle и анализируется преподавателем. Итоговое задание оценивается по 100 балльной шкале – как отношение суммы набранных баллов за каждое задание к общему количеству баллов.

8.3. Пример итогового проектного задания по образовательной программе.

В задании предлагается провести предварительную обработку данных и построение модели для классификации клиентов банка по набору данных Marketing, представляющего данные маркетингового исследования клиентов банка, проведенного в 2011 г. Основной целью задания является построение максимально точной модели для предсказания отклика клиентов на маркетинговые исследования по новым банковским продуктам.

Описание данных:

TARGET - наличие отклика на маркетинговую кампанию (1 - отклик был зарегистрирован, 0 - отклика не было). Тип переменной - категориальная, бинарная.

AGE - возраст клиента. Тип переменной - количественная, дискретная.

SOCSTATUS\_WORK\_FL - социальный статус клиента относительно работы (1 - работает, 0 - не работает). Тип переменной - категориальная, бинарная.

SOCSTATUS\_PENS\_FL - социальный статус клиента относительно пенсии (1 - пенсионер, 0 - не пенсионер). Тип переменной - категориальная, бинарная.

GENDER - пол клиента (0 - мужской, 1 - женский). Тип переменной - категориальная, бинарная.

EDUCATION - образование клиента. Тип переменной - категориальная, небинарная, порядковая.

MARITAL\_STATUS - семейное положение клиента. Тип переменной - категориальная, небинарная, номинальная.

ORG\_TP\_STATE - форма собственности компании. Тип переменной - категориальная, небинарная, номинальная.

JOB\_DIR - направление деятельности внутри компании. Тип переменной - категориальная, небинарная, номинальная.

PERSONAL\_INCOME - личный доход клиента (в рублях). Тип переменной - количественная, непрерывная.

OWN\_AUTO - кол-во автомобилей в собственности. Тип переменной - количественная, дискретная.

WORK\_TIME - время работы на текущем месте (в месяцах). Тип переменной - количественная, непрерывная.

LOAN\_MAX\_DLQ\_AMT - максимальная сумма просрочки (в рублях). Тип переменной - количественная, непрерывная.

Задание 1. (1 балл)

Как вы видите, в данных имеются пропуски (NaN). Любым способом посчитайте количество пропусков в каждом из столбцов (но мы должны видеть как вы это делали).

Для тех переменных, в которых есть пропуски, выведите количество пропущенных значений в следующем формате: "В переменной {название переменной} {количество} пропущенных значений"

Задание 2. (1 балл)

Пропуски в столбцах ORG\_TP\_STATE, JOB\_DIR, WORK\_TIME могут быть связаны с тем, что клиент является пенсионером и не работает. Заполните у неработающих пенсионеров пустые значения в столбцах ORG\_TP\_STATE и JOB\_DIR значением "Не работает", в столбце WORK\_TIME значением 0. Удалите оставшиеся строки с пустыми значениями. Сохраните изменения в тот же набор данных.

Задание 3. (0.5 балла)

Для решения задачи классификации необходимо сформировать список категориальных небинарных признаков, которые далее будем кодировать One-hot кодированием. Сделайте два списка: категориальных признаков categorical\_columns и числовых признаков numeric\_columns на основе анализа типа данных столбца.

Задание 4. (0.5 балла)

В Pandas One-hot кодирование выполняется функцией pd.get\_dummies. Сгенерируйте One-hot признаки для нашего датасета. Сохраните полную матрицу объекты признаки в переменную X.

Задание 5. (0.5 балла)

Постройте матрицу и гистограмму распределения целевой переменной TARGET. Сбалансирован ли набор данных относительно целевой переменной? Какие метрики нужно использовать для измерения качества классификации в нашем наборе?

Задание 6. (0.5 балла)

Удалите из набора Х целевой столбец и положите его в y. Разбейте наборы на обучающую и тестовую выборки в соотношении 80:20, сохранив распределение целевой переменной и используя random\_state = 56.

Задание 7. (1 балл)

Обучите на обучающей выборке базовые классификаторы с параметрами по умолчанию: KNeighborsClassifier, LogisticRegression, DecisionTreeClassifier, RandomForestClassifier. На тестовой выборке вычислите основные показатели качества - долю правильных ответов, точность, полноту и F1-меру для каждого классификатора. Сделайте выводы о качестве модели

Задание 8. (0.5 балла)

Попробуем выполнить очистку данных. Выведите основные статистики (минимум, максимум, среднее, дисперсию, квантили) для всех столбцов.  
Постройте гистограммы распределения признаков в датасете Определите, есть ли в каких-то количественных столбцах выбросы - искаженные данные, имеющие слишком малое или слишком большое значение.

Задание 9. (0.5 балла)

В столбце PERSONAL\_INCOME есть малые значения. Отсортируйте набор данных по этому столбцу и выдайте 10 минимальных значений. Определите индекс записи с минимальным доходом и сохраните его в переменную min\_index.

Задание 10. (0.5 балла)

Замените значения, меньшие 1000, в столбце PERSONAL\_INCOME, увеличив их в 1000 раз. Убедитесь, что запись min\_index изменилась.

Задание 11. (0.5 балла)

В столбце WORK\_TIME есть выбросы - нереальные значения трудового стажа. Определите количество человек, у которых этот столбец содержит нереальные значения - количество месяцев превышает возраст человека в месяцах с момента начала трудовой деятельности (в 16 лет). Обратите внимание, что возраст 30 лет в наборе данных может на самом деле означать 30 лет 11 месяцев. Удалите данные об этих людях из набора и положите результат в clear\_data. Все дальнейшие задания выполняйте с набором clear\_data.

Задание 12. (0.5 балла)

Добавьте новый категориальный признак INCOME\_LEVEL на основе столбца PERSONAL\_INCOME:

низкий < 6000

ниже среднего = 6000 - 11999

средний = 12000 – 24999

выше среднего = 25000 - 39999

высокий >= 40000

Постройте столбчатую диаграмму по этому признику. Обратите внимание, на диаграмме метки оси х должны быть упорядочены по возрастанию дохода.

Задание 13. (0,5 балла)

Заново постройте наборы признаков, закодируйте категориальные признаки One-hot кодированием, числовые - масштабируйте c помощью StandardScaler(). Сделайте обучающую и тестовую выборку и проверьте работу наших классификаторов. Сделайте выводы.

Задание 14. (1 балл)

С помощью перебора по указанной ниже сетке подберите оптимальный параметр для RandomForest. parameters = {'n\_estimators':[10, 20, 30, 50, 75, 100, 125, 150, 200]} Для выбора наилучшей модели используйте F1-меру. Выведите показатели лучшей модели на тестовой выборке и оптимальное значение параметра. Сделайте выводы.

Задание 15. (1 балл)

Для работы с несбалансированными данными разработана специальная библиотека imbalanced-learn <https://imbalanced-learn.org/stable/index.html>.

Попробуйте использовать для обучения нашей модели ансамблевые классификаторы BalancedRandomForestClassifier и RUSBoostClassifierиз этой библиотеки. Проверьте работу классификаторов на тестовой выборке. Сделайте выводы. Какие показатели стали хуже, какие улучшились?

Задание 16. (1 балл)

Сделайте удаленный доступ к обученной модели с помощью фреймворка Flask.

8.4. Тесты и обучающие задачи (кейсы), иные практикоориентированные формы заданий

Пример промежуточного теста.

1. Метеоролог хочет построить модель, прогнозирующую температуру воздуха на завтра. К какому типу относится данная задача?

* Классификация
* Ранжирование
* Кластеризация
* Регрессия

1. Врач хочет построить модель, прогнозирующую возможность развития осложнений у пациента после операции - все ли будет благополучно в течение нескольких следующих месяцев или нет. К какому типу относится данная задача?

* Классификация
* Ранжирование
* Кластеризация
* Регрессия

1. Астроном хочет построить модель, которая сможет разбить известные науке звезды на группы по их характеристикам, чтобы лучше изучить их особенности. К какому типу относится данная задача?

* Классификация
* Ранжирование
* Кластеризация
* Регрессия

1. В задаче какого типа в обучающей выборке для объектов НЕТ ответов?

* Регрессия
* Во всех этих задачах в обучающей выборке есть ответы
* Классификация
* Кластеризация

1. Выберите все верные утверждения:

* Модель машинного обучения, по сути, является отображением пространства ответов в пространство объектов
* Элементами обучающей выборки являются объекты, характеристики которых являются значениями признаков
* Функционал ошибки показывает, насколько плохое качество имеют данные, используемые для решения задачи
* Процесс обучения модели заключается в минимизации функционала ошибки

1. Стоимость акции некоторой компании на бирже является…

* категориальным признаком
* числовым признаком
* бинарным признаком
* порядковым признаком

1. Наличие балкона в квартире является…

* порядковым признаком
* категориальным признаком
* бинарным признаком
* числовым признаком

1. Возраст человека…

* может считаться как числовым, так и порядковым признаком
* может считаться только порядковым признаком
* может считаться только категориальным признаком
* может считаться только числовым признаком

Пример практического задания.

Набор данных – информация о студентах математических школ.

Найдите ответы на вопросы:

1. Какая причина выбора школы была самой частой? В качестве ответа приведите соответствующее значение признака.

2. Найдите количество студентов, у родителей которых нет никакого образования.

3. Найдите минимальный возраст учащегося школы Mousinho da Silveira.

4. Найдите количество студентов, имеющих нечетное число пропусков.

5. Найдите разность между средними итоговыми оценками студентов, состоящих и не состоящих в романтических отношениях. В качестве ответа приведите число, округленное до двух значащих цифр после запятой.

6. Сколько занятий пропустило большинство студентов с самым частым значением наличия внеклассных активностей?

Пример практического задания.

1. Загрузите данные load\_wine из sklearn.datasets. Из обучающей части исключите объекты класса 2. Обучите на этих данных случайный лес, используйте гиперпараметр n\_estimators = 5, зафиксируйте random\_state=0. Оцените качество по метрике MSE на тестовой части. Ответ разделите на 1000 и округлите до целой части по математичестким правилам округления.

2.Загрузите данные load\_wine из sklearn.datasets. Из обучающей части исключите объекты класса 2. Обучите случайный лес, задав только гиперпараметры n\_estimators=100 и random\_state=0. Оцените важность признаков. Укажите название двух наиболее важных признаков.

3.Загрузите данные load\_wine из sklearn.datasets. Из обучающей части исключите объекты класса 2. Отмасштабируйте признаки, используя класс StandardScaler с гиперпараметрами по умолчанию. Обучите случайный лес, задав только гиперпараметры n\_estimators = 100 и random\_state=0. Оцените важность признаков. Укажите название двух наиболее важных признаков.

8.5. Описание процедуры оценивания результатов обучения.

Оценка за курс вычисляется по 100 балльной шкале как средний процент правильных ответов по всем тестам курса. Слушатель аттестуется если оценка за курс больше либо равна 50 и итоговое задание выполнено не менее чем на 50 баллов из 100.

Уровень сформированности компетенций: 50-60 баллов – начальный, 61-80 баллов – базовый, 81- 100 баллов – продвинутый.

**9.Организационно-педагогические условия реализации программы**

**9.1. Кадровое обеспечение программы**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| №  п/п | Фамилия, имя, отчество (при наличии) | Место основной работы и должность, ученая степень и ученое звание (при наличии) | Ссылки на веб-страницы с портфолио (при наличии) | Фото в формате jpeg | Отметка о полученном согласии на обработку персональных данных |
| 1 | Шухман Александр Евгеньевич | Зав. кафедрой геометрии и компьютерных наук, к.п.н., доцент | [http://osu.ru/doc/1041/kaf/6349 /prep/12564](http://osu.ru/doc/1041/kaf/6349/prep/12564) |  | получено |
| 2 | Морковина Эльвира Фаридовна | Доцент кафедры геометрии и компьютерных наук, к.п.н., доцент | [http://osu.ru/doc/1041/kaf/6349 /prep/14713](http://osu.ru/doc/1041/kaf/6349/prep/14713) |  | получено |
| 3 | Парфенов Денис Игоревич | Доцент кафедры пркладной математики, к.т.н. |  |  | получено |

**9.2.Учебно-методическое обеспечение и информационное сопровождение**

|  |  |
| --- | --- |
| **Учебно-методические материалы** | |
| Методы, формы и технологии | Методические разработки,  материалы курса, учебная литература |
| Электронный курс в системе moodle.osu.ru |  |
| Учебная литература | 1.  Вандер Плас Дж. Python для сложных задач. Наука о данных и машинное обучение. Питер, 2018  2. Рашка С. Python и машинное обучение. Машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow. Вильямс, 2019. |

|  |  |
| --- | --- |
| **Информационное сопровождение** | |
| Электронные  образовательные ресурсы | Электронные  информационные ресурсы |
| МООК «Введение в машинное обучение» https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie | MachineLearning.ru  Профессиональный информационно-аналитический ресурс, посвященный  [машинному обучению](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5), [распознаванию образов](http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%B0%D1%82%D0%B5%D0%B3%D0%BE%D1%80%D0%B8%D1%8F:%D0%A0%D0%B0%D1%81%D0%BF%D0%BE%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2) и [интеллектуальному анализу данных](http://www.machinelearning.ru/wiki/index.php?title=%D0%98%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D1%83%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85). |
| МООК «Основы программирования на Python»  <https://www.coursera.org/learn/python-osnovy-programmirovaniya> | <https://scikit-learn.org/stable/>  scikit-learn Machine Learning in Python |
| МООК «Быстрый старт в ИИ»  https://stepik.org/course/80782/syllabus |  |

**9.3.Материально-технические условия реализации программы**

|  |  |
| --- | --- |
| Вид занятий | Наименование оборудования,  программного обеспечения |
| Лекция | Персональный компьютер или ноутбук с ОС Windows, Mac OS или Linux.  Zoom для проведения лекций, Yandex disk для выкладывания записей, moodle для ссылок на лекции и тестирования. |
| Практическое занятие | Персональный компьютер или ноутбук с ОС Windows, Mac OS или Linux.  Jupiter Notebook или Google Colaboratory для выполнения заданий, Zoom для проведения занятия, Yandex disk для выкладывания записей, moodle для ссылок на занятия и тестирования |
| Самостоятельная работа | Персональный компьютер или ноутбук с ОС Windows, Mac OS или Linux.  Jupiter Notebook или Google Colaboratory для выполнения заданий, moodle для загрузки заданий и тестирования |
| Итоговое задание | Персональный компьютер или ноутбук с ОС Windows, Mac OS или Linux.  Jupiter Notebook или Google Colaboratory для выполнения заданий, moodle для загрузки решений |

**III.Паспорт компетенций (Приложение 2)**

**ПАСПОРТ КОМПЕТЕНЦИИ**

**«Основы анализа данных и машинного обучения»**

федеральное государственное бюджетное образовательное учреждение высшего образования «Оренбургский государственный университет»

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | Наименование компетенции | Способность эффективно применять современные методы и средства анализа данных и машинного обучения для решения прикладных задач | |
| 2. | Указание типа компетенции | профессиональная | |
| 3. | Определение, содержание и основные сущностные характеристики компетенции | Под компетенцией понимается способность использования методов анализа данных и машинного обучения, а также средств языка программирования Python для постановки и решения задач предварительной обработки, классификации, регрессии, кластеризации и визуализации данных в различных предметных областях.  Выпускник курса должен:  знать: основы языка программирования Python, основные этапы анализа данных, основные типы задач анализа данных и машинного обучения, математические основы и методы решения задач классификации, регрессии, кластеризации, основные метрики качества алгоритмов машинного обучения, математические основы и возможности нейросетевых методов, методы предобработки и векторизации текстов, методы конструирования признаков, методы отбора признаков для понижения размерности задачи, методы сбора информации из интернет-источников, методы создания API для доступа к моделям машинного обучения.  уметь: использовать язык программирования Python, интерактивные блокноты Jupiter Notebooks с библиотеками Pandas и Matplotlib для автоматизированного сбора, предварительной обработки, очистки и визуализации данных, конструирования и отбора признаков, обучать и использовать модели из библиотеки SkLearn для решения задач машинного обучения с оценкой качества и подбором оптимальных гиперпараметров моделей, обучать и использовать нейросетевые модели на основе библиотеки Keras+Tensorflow для задач классификации изображений и текстов, в том числе с использованием переноса обучения, создавать API для доступа к моделям на основе каркаса Flask.  владеть: навыками эффективного решения прикладных задач на основе данных из различных источников с подбором оптимальных гиперпараметров моделей машинного обучения, навыками предобработки и классификации текстов и изображений, навыками создания API для доступа к разработанным моделям. | |
| 4. | Дескриптор знаний, умений и навыков по уровням | Уровни сформированности компетенции обучающегося | Индикаторы сформированности компетенции в зависимости от уровня |
| Начальный уровень (Компетенция недостаточно развита. Частично проявляет навыки, входящие в состав компетенции. Пытается, стремится проявлять нужные навыки, понимает их необходимость, но у него не всегда получается) | Знает: основы языка программирования Python, основные этапы анализа данных, основные типы задач анализа данных и машинного обучения, базовые методы решения задач классификации, регрессии, кластеризации, основные метрики качества алгоритмов машинного обучения, области применения нейросетей для анализа изображений и текстов.  Умеет: использовать язык программирования Python, интерактивные блокноты Jupiter Notebooks с библиотеками Pandas и Matplotlib для предварительной обработки и визуализации данных на основе заданных сценариев, обучать и использовать модели из библиотеки SkLearn для решения задач машинного обучения с оценкой качества модели по аналогии с решенными задачами  Владеет: навыками использования  методов и средств анализа данных и машинного обучения для различных прикладных задач с данными, заданными таблично |
| Базовый уровень (Уверенно владеет навыками, способен, проявлять соответствующие навыки в ситуациях с элементами неопределённости сложности) | Знает: математические основы методов машинного обучения, методы подбора оптимальных гиперпараметров моделей машинного обучения, математические основы нейросетевых методов, методы предобработки и векторизации текстов.  Умеет: самостоятельно проводить предобработку и очистку данных, выполнять выбор и сравнительный анализ моделей машинного обучения для конкретных прикладных задач, выполнять подбор оптимальных гиперпараметров моделей машинного обучения, обучать и использовать нейросетевые модели на основе библиотеки Keras+Tensorflow для задач классификации изображений и текстов, в том числе с использованием переноса обучения.  Владеет: навыками эффективного решения прикладных задач с подбором  оптимальных гиперпараметров моделей машинного обучения,  навыками предобработки и классификации текстов и изображений |
| Продвинутый (Владеет сложными навыками, способен активно влиять на происходящее, проявлять соответствующие навыки в ситуациях повышенной сложности) | Знает: методы конструирования признаков, методы отбора признаков для понижения размерности задачи, методы сбора информации из интернет-источников, возможности современных нейросетевых архитектур, методы создания API для доступа к моделям машинного обучения.  Умеет: проводить отбор и конструирование признаков, автоматизировать сбор информации из интернет-источников, использовать для решения задач современные нейросетевые архитектуры, создавать API для доступа к моделям на основе каркаса Flask.  Владеет: навыками эффективного решения прикладных задач на основе данных из различных источников, навыками создания API для доступа к разработанным моделям |
| 5. | Характеристика взаимосвязи данной компетенции с другими компетенциями/ необходимость владения другими компетенциями для формирования данной компетенции | Требует владения следующим компетенциями: цифровая грамотность, основы алгоритмизации и программирования | |
| 6. | Средства и технологии оценки | Для оценки знаний используются тесты в автоматизированной системе управления обучением Moodle. Тесты разработаны по каждой лекции курса. Тест содержит от 4 до 6 заданий с выбором варианта ответа или вводом ответа.  Для оценки умений используются практические задания на использование конкретных методов анализа данных и машинного обучения, которые необходимо самостоятельно выполнить в системе Jupiter или Google Colaboratory. Практические задания разработаны в интерактивных блокнотах, которые содержат формулировку заданий и место для вставки кода решения.  Практические задания проверяются автоматизировано с помощью тестов, вопросы которых соответствуют ответам на каждый пункт задания. Кроме того, практические задания прикрепляются в систему Moodle и анализируются преподавателем.  Для оценки навыков используется итоговое задание, включающее в себя все этапы решения прикладной задачи – от сбора и предобработки данных до создания API к обученной модели. Итоговое задание прикрепляется в систему Moodle и анализируется преподавателем. | |

**IV.Иная информация о качестве и востребованности образовательной программы**

нет

**V.Рекомендаций к программе от работодателей**:

Рекомендации от АО»Синимекс-информатика» и ООО «Ребис».

**VI.Указание на возможные сценарии профессиональной траектории граждан** по итогам освоения образовательной программы (в соответствии с приложением)

Сценарии профессиональной траектории граждан

|  |  |
| --- | --- |
| Цели получения персонального цифрового сертификата | |
| **Текущий статус** | **Цель** |
| **Трудоустройство** | |
| состоящий на учете в Центре занятости безработный | трудоустроенный, самозанятый (фриланс), ИП/бизнесмен |
| **Развитие компетенций в текущей сфере деятельности** | |
| работающий по найму в организации, на предприятии | развитие профессиональных качеств |
| работающий по найму в организации, на предприятии | повышение заработной платы |
| работающий по найму в организации, на предприятии | смена работы без изменения сферы профессиональной деятельности |
| временно отсутствующий на рабочем месте (декрет, отпуск по уходу за ребенком и др.) | сохранение и развитие квалификации |
| **Переход в новую сферу занятости** | |
| освоение новой сферы занятости | самозанятый, ИП/бизнесмен, расширение кругозора |
| освоение смежных профессиональных областей | повышение уровня дохода, расширение профессиональной деятельности |

**VII.Дополнительная информация**

**нет**

**VIII.Приложенные Скан-копии** Утвержденной рабочей программа (подпись, печать, в формате pdf)