СОГЛАСОВАНО:

Главный эксперт

_____ Гордеев Никита Николаевич

СОГЛАСОВАНО:

Индустриальный партнер

АО «Научно-исследовательский институт «Вектор»,

заместитель директора Центра защиты информации

_____ Д.В. Магницкий

М.П

КОНКУРСНОЕ ЗАДАНИЕ КОМПЕТЕНЦИИ «МАШИННОЕ ОБУЧЕНИЕ И БОЛЬШИЕ ДАННЫЕ»

2023

Конкурсное задание разработано экспертным сообществом и утверждено Менеджером компетенции, в котором установлены нижеследующие правила и необходимые требования владения профессиональными навыками для участия в соревнованиях по профессиональному мастерству.

Конкурсное задание включает в себя следующие разделы:

Используемые сокращения

ML- это класс методов, в которых машину, алгоритм, учат думать и действовать как человек на основе полученного опыта или данных.

MNIST - Модифицированный Национальный институт стандартов и технологий (база данных). База данных MNIST является одним из самых известных эталонных наборов данных в машинном обучении. Благодаря простоте использования и относительно небольшому размеру, он часто используется для сравнения производительности различных алгоритмов.

NLP - это область машинного обучения, которая посвящена тому, как компьютер анализирует естественный, то есть человеческий язык.

DL - совокупность методов машинного обучения (с учителем, с частичным привлечением учителя, без учителя, с подкреплением), основанных на обучении представлениям (англ. feature/representation learning), а не специализированных алгоритмах под конкретные задачи.

OCR – механический или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные, использующиеся для преставления символов в компьютере (например, в текстовом редакторе).

CNN - это алгоритм глубокого обучения, который может принимать входное изображение, присваивать важность (изучаемые веса и смещения) аспектам или объектам изображении и отличать одно от другого.

1. Основные требования компетенции

1.1. Общие сведения о требованиях компетенции

Требования компетенции (ТК) «Машинное обучение и большие данные» определяют знания, умения, навыки и трудовые функции, которые лежат в основе наиболее актуальных требований работодателей отрасли.

Целью соревнований по компетенции является демонстрация лучших практик и высокого уровня выполнения работы по соответствующей рабочей специальности или профессии.

Требования компетенции являются руководством для подготовки конкурентоспособных, высококвалифицированных специалистов / рабочих и участия их в конкурсах профессионального мастерства.

В соревнованиях по компетенции проверка знаний, умений, навыков и трудовых функций осуществляется посредством оценки выполнения практической работы.

Требования компетенции разделены на четкие разделы с номерами и заголовками, каждому разделу назначен процент относительной важности, сумма которых составляет 100.

1.2. Перечень профессиональных задач специалиста по компетенции «Машинное обучение и большие данные»

Таблица №1

Перечень профессиональных задач специалиста

№ п/п	Раздел	Важность в %
1	**Подготовка данных для проведения аналитических работ по исследованию больших данных**	21
	Специалист должен знать и понимать: – Теоретические и прикладные основы анализа больших данных; – Современные методы и инструментальные средства анализа больших данных; – Современный опыт использования анализа больших данных; – Типы больших данных: метаданные, полуструктурированные, структурированные, неструктурированные; – Методы извлечения информации и знаний из гетерогенных, мультиструктурированных, неструктурированных источников, в том числе при потоковой обработке;	

	– Технологии хранения и обработки больших данных в организации: базы данных, хранилища данных, распределенная и параллельная обработка данных, вычисления в оперативной памяти; – Облачные технологии, облачные сервисы	
	Специалист должен уметь: – Использовать инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников, в том числе в режиме реального времени; – Производить очистку данных для проведения аналитических работ; – Проводить интеграцию и преобразование больших объемов данных; – Оценивать соответствие наборов данных задачам анализа больших данных; – Оценивать стоимость данных для проведения аналитических работ	
2	**Планирование и организация аналитических работ с использованием технологий больших данных**	24
	Специалист должен знать и понимать: – Возможности использования свободно распространяемого программного обеспечения для анализа больших данных; – Основы планирования аналитических работ; – Стандарты проведения анализа данных; – Методы и инструментальные средства управления аналитическими проектами по исследованию больших данных; – Содержание и последовательность выполнения этапов аналитического проекта по исследованию больших данных; – Типы анализа больших данных, виды аналитики; – Теоретические и прикладные основы анализа больших данных; – Современные методы и инструментальные средства анализа больших данных; – Теория вероятностей и математическая статистика; – Методы интерпретации и визуализации анализа больших данных	
	Специалист должен уметь: – Представлять содержание и результаты работ по анализу больших данных; – Планировать аналитические работы с использованием технологий больших данных; – Проводить аналитические работы с использованием технологий больших данных; – Проводить анализ больших данных; – Осуществлять интеграцию и преобразование данных в ходе работ по анализу больших данных	
3	**Проведение аналитического исследования с применением технологий больших данных в соответствии с требованиями заказчика Наименование раздела (знания, умения, трудовые функции)**	24

	Специалист должен знать и понимать: – Технологии анализа данных: статистический анализ, семантический анализ, анализ изображений, машинное обучение, методы сравнения средних, частотный анализ, анализ соответствий, кластерный анализ, дискриминантный анализ, факторный анализ, деревья классификации, многомерное шкалирование, моделирование структурными уравнениями, методы анализа выживаемости, временные ряды, планирование экспериментов, карты контроля качества; – Статистический анализ: метод многовариантного тестирования, корреляционный анализ, регрессионный анализ; – Статистические методы: параметрические, непараметрические, управляемые, неуправляемые, полууправляемые, кластеризация; – Семантический анализ: обработка естественного языка, сентиментный анализ, анализ текста; – Алгоритмы машинного обучения: обучение с учителем, обучение без учителя, полууправляемое обучение, обучение с подкреплением; – Машинное обучение: классификация, кластеризация, обнаружение выбросов, фильтрация; – Методы и модели классификации: логистическая регрессия, деревья решений, предредукция, постредукция, модели, основанные на правилах, вероятностные классификаторы, усиление энтропии информации; – Фильтрация шумовых выбросов, виды шумовых выбросов: глобальный, контекстуальный, коллективный; – Анализ изображений, анализ сетей, анализ пространственных данных, анализ временных рядов; – Методы идентификации шаблонов;Методы оценки моделей: оценка качества построенной модели по тестовой выборке и анализ обобщающих способностей алгоритма; – Правила деловой переписки – Методы разработки отчетной аналитической документации	
	Специалист должен уметь: – Проводить сравнительный анализ методов и инструментальных средств анализа больших данных; – Разрабатывать и оценивать модели больших данных; – Программировать на языках высокого уровня, ориентированных на работу с большими данными: для статистической обработки данных и работы с графикой, для работы с разрозненными фрагментами данных в больших массивах, для работы с базами структурированных и неструктурированных данных; – Адаптировать и развертывать модели в предметной среде; – Решать задачи классификации, кластеризации, регрессии, прогнозирования, снижения размерности и ранжирования данных; – Решать проблемы переобучения и недообучения алгоритма; – Формировать предложения по использованию результатов анализа; – Оформлять результаты аналитического исследования для представления заказчику	

4	**Разработка продуктов на основе встроенной аналитики больших данных**	21
	Специалист должен знать и понимать: – Локальные и глобальные потребности в создании новых и модернизации существующих продуктов на основе встроенной аналитики больших данных; – Существующие и перспективные методы и программный инструментарий технологий больших данных; – Существующий опыт разработки и использования продуктов и услуг на основе технологий больших данных; – Современные и перспективные методы сбора, хранения и передачи данных из гетерогенных источников; – Источники больших данных, интенсивность генерации данных источниками; – Технические средства и среды сбора, хранения и обработки больших данных; – Современные и перспективные средства визуализации и интерпретации больших данных; – Системная инженерия; – Машинное обучение; – Математическое моделирование; – Методы сравнительного анализа; – Основы инновационной деятельности и управления инновациями в сфере информационных технологий; – Основы управления информационно-технологическими проектами; – Показатели эффективности технологий больших данных; – Основы охраны авторских прав и интеллектуальной собственности в сфере информационных технологий; – Правила деловой переписки	
	Специалист должен уметь: – Проводить аналитические и поисковые исследования по тематике информационных технологий, технологий больших данных; – Проводить маркетинговые исследования в области информационных продуктов и услуг; – Разрабатывать конкурсную, проектную и рабочую документацию на разработку новых продуктов; – Проводить технико-экономическое обоснование разработки новых продуктов; – Оценивать экономические параметры технологий больших данных; – Осуществлять разработку программно-аппаратных компонентов и систем; – Осуществлять математическое и информационное моделирование; – Проводить аналитические работы на основе технологий больших данных; – Разрабатывать научно-техническую документацию; – Проводить презентации, подготавливать публикации по итогам проектных работ	

5	**Разработка инфраструктурных решений на основе аналитики больших данных**	10
	Специалист должен знать и понимать: – Производители программного обеспечения и инфраструктуры технологий больших данных; – Принципы и методы управления защитой и обеспечением конфиденциальности больших данных; – Функциональные возможности существующих программного обеспечения и инфраструктуры технологий больших данных, условия их приобретения и использования; – Методы выявления требований на создание инфраструктурных решений на основе больших данных; – Технологии подготовки и проведения презентаций; – Методы анализа деятельности организации; – Основы организационного дизайна; – Методы управления проектами в области больших данных; – Исследование операций; – Методы и инструменты бизнес-аналитики; – Источники информации, в том числе информации, необходимой для обеспечения деятельности в предметной области, условия их использования; – Высокопроизводительные и распределенные вычисления; – Показатели эффективности технологий больших данных	
	Специалист должен уметь: – Проводить аналитические и поисковые исследования по тематике информационных технологий, технологий больших данных; – Проводить презентации; – Анализировать бизнес-процессы и функции подразделений организации, выделять проблемные места и возможности совершенствования информационно-технологической инфраструктуры для анализа больших данных; – Проводить сравнительный анализ методов и программного обеспечения функций и поддержки бизнес-процессов, процессов принятия решений и аналитических задач на основе технологий больших данных; – Разрабатывать проектную документацию на разработку проектов в области больших данных; – Выполнять технико-экономическое обоснование разработки проектов в области больших данных и информационно-технологической инфраструктуры организации; – Оценивать экономические параметры технологий больших данных; – Разрабатывать проектную документацию по проектам инфраструктурных решений на основе больших данных; – Проводить приемо-сдаточные испытания проектов инфраструктурных решений на основе больших данных – Проводить презентации по итогам проектных работ	

1.3. Требования к схеме оценки

Сумма баллов, присуждаемых по каждому аспекту, должна попадать в диапазон баллов, определенных для каждого раздела компетенции, обозначенных в требованиях и указанных в таблице №2.

Таблица №2

Матрица пересчета требований компетенции в критерии оценки

Критерий/Модуль							Итого баллов за раздел ТРЕБОВАНИЙ КОМПЕТЕНЦИИ
Разделы ТРЕБОВАНИЙ КОМПЕТЕНЦИИ		А	Б	В	Г	Д	
	1	21					21
	2		24				24
	3			24			24
	4				21		21
	5					10	10
Итого баллов за критерий/модуль		21	24	24	21	10	**100**

1.4. Спецификация оценки компетенции

Оценка Конкурсного задания будет основываться на критериях, указанных в таблице №3:

Таблица №3

Оценка конкурсного задания

Критерий		Методика проверки навыков в критерии
А	**Парсинг и предобработка данных**	Оценка на соревнованиях попадает в одну из двух категорий: объективное и судейское решение. Для обеих категорий оценки использование точных эталонов для сравнения, по которым оценивается каждый аспект, является существенным для гарантии качества.
Б	**Разведочный анализ данных**	Оценка на соревнованиях попадает в одну из двух категорий: объективное и судейское решение. Для обеих категорий оценки использование точных эталонов для сравнения, по которым оценивается каждый аспект, является существенным для гарантии качества.
В	**Построение, обучение и оптимизация модели**	Оценка на соревнованиях попадает в одну из двух категорий: объективное и судейское решение. Для обеих категорий оценки использование точных эталонов для сравнения, по которым оценивается каждый аспект, является существенным для гарантии качества.
Г	**Разработка программного продукта**	Оценка на соревнованиях попадает в одну из двух категорий: объективное и судейское решение. Для обеих категорий оценки использование точных эталонов для

		сравнения, по которым оценивается каждый аспект, является существенным для гарантии качества.
Д	**Разработка средств интеграции и поддержки готового решения**	Оценка на соревнованиях попадает в одну из двух категорий: объективное и судейское решение. Для обеих категорий оценки использование точных эталонов для сравнения, по которым оценивается каждый аспект, является существенным для гарантии качества.

1.5. Конкурсное задание

Общая продолжительность Конкурсного задания1: 15 ч.

Количество конкурсных дней: 3 дней

Вне зависимости от количества модулей, КЗ должно включать оценку по каждому из разделов требований компетенции.

Оценка знаний участника должна проводиться через практическое выполнение Конкурсного задания. В дополнение могут учитываться требования работодателей для проверки теоретических знаний / оценки квалификации.

1.5.1. Разработка/выбор конкурсного задания

Конкурсное задание состоит из 5 модулей, включает обязательную к выполнению часть (инвариант) - 4 модуля, и вариативную часть - 1 модуль. Общее количество баллов конкурсного задания составляет 100.

Обязательная к выполнению часть (инвариант) выполняется всеми регионами без исключения на всех уровнях чемпионатов.

Количество модулей из вариативной части, выбирается регионом самостоятельно в зависимости от материальных возможностей площадки соревнований и потребностей работодателей региона в соответствующих специалистах. В случае если ни один из модулей вариативной части не подходит под запрос работодателя конкретного региона, то вариативный (е) модуль (и) формируется регионом самостоятельно под запрос работодателя. При этом, время на выполнение модуля (ей) и количество баллов в критериях оценки по аспектам не меняются.

---

[1] *Указывается суммарное время на выполнение всех модулей КЗ одним конкурсантом.*

Таблица №4

Матрица конкурсного задания

Обобщенная трудовая функция	Трудовая функция	Нормативный документ/ЗУН	Модуль	Константа/ вариатив	ИЛ	КО
1	2	3	4	5	6	7
Анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры	Подготовка данных для проведения аналитических работ по исследованию больших данных	ПС: 06.042; ФГОС СПО 09.02.07 Информационные системы и программирование	Модуль А – Парсинг и предобработка данных	Инвариант	Раздел ИЛ 1	21
Анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры	Планирование и организация аналитических работ с использованием технологий больших данных	ПС: 06.042; ФГОС СПО 09.02.07 Информационные системы и программирование	Модуль Б- Разведочный анализ данных	Инвариант	Раздел ИЛ 2	24
Анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры	Проведение аналитического исследования с применением технологий больших данных в соответствии с требованиями заказчика	ПС: 06.042; ФГОС СПО 09.02.07 Информационные системы и программирование	Модуль В – Построение, обучение и оптимизация модели	Инвариант	Раздел ИЛ 3	24

Управление разработкой продуктов, услуг и решений на основе больших данных	Разработка продуктов на основе встроенной аналитики больших данных	ПС: 06.042; ФГОС СПО 09.02.07 Информационные системы и программирование	Модуль Г – Разработка программного продукта	Инвариант	Раздел ИЛ 4	21
Управление разработкой продуктов, услуг и решений на основе больших данных	Разработка инфраструктурных решений на основе аналитики больших данных	ПС: 06.042; ФГОС СПО 09.02.07 Информационные системы и программирование	Модуль Д – Разработка средств интеграции и поддержки готового решения	Вариатив	Раздел ИЛ 5	10
						100

Инструкция по заполнению матрицы конкурсного задания **(Приложение № 1)**

1.5.2. Структура модулей конкурсного задания (инвариант/вариатив)

Модуль А. Парсинг и предобработка данных

Время на выполнение модуля 3 часа.

Описание модуля:

Ежегодно в России проводится Национальная премия за вклад в развитие российского сегмента сети Интернет. Каждый раз экспертное жюри осматривает вручную все заявки и выбирает из них самых достойных, а экспертный совет путем голосования выбирает в каждой номинации победителя. Премия Рунета является желаемой наградой не только для бизнеса, но и для представителей госорганов и ведомств. Данная премия является престижнейшей в Рунете и самой значимой, ежегодно собирая более 10 000 зрителей.

Задача: на основе аналитических статей, связанных с информационными технологиями, бизнесом и интернетом с Хабра (системы тематических коллективных блогов с элементами новостного сайта) обобщить информацию по публичной активности каждой организации-номинанта и разработать алгоритм, способный самостоятельно определять номинацию компании.

При выполнении модуля ставятся следующие цели:

Использовать инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников.

При выполнении данного модуля ставятся следующие задачи:

1. Выполнить парсинг данных для сбора информации о компаниях с указанных веб-ресурсов;
2. Формирование структуры набора данных;
3. Провести предварительную обработку данных;
4. Выполнить построение и отбор признаков.

Требования к оформлению письменных материалов

Письменный материал отсутствует.

Представление результатов работы

Результат выполнения Модуля А «Парсинг и предобработка данных»: предобработанные данные (архив Data.zip), отчет о проделанной работе (Report_M1.html, Report_M1.ipynb), дополнительные комментарии коду (Readme.txt).

Необходимые приложения

Приложение 1: Архив, содержащий статьи о компаниях (Data.zip)

Приложение 2: Список номинантов конкурса (Condidates.doc)

ЗАДАНИЕ

1.1 Парсинг данных

На основе имеющихся аналитических статей, связанных с информационными технологиями, бизнесом и интернетом необходимо построить исходный набор данных (.csv или .xml). Набор данных должен включать названия, описание, рейтинг и сферу деятельности компаний, дату публикации, а также текст статей из Интернет-ресурсов. Подготовленный набор данных должен содержать сведения о всех номинантах конкурса.

Можно дополнить набор какими-либо другими данными, если они могут быть полезны для дальнейшего исследования.

1.2 Формирование структуры набора данных

Задача заключается в определении класса(кластера) – номинации премии Рунета. Исходя из этого, необходимо определить, какие атрибуты имеют наибольшее влияние на определение классов(кластеров), оставить только их для последующего обучения. Также необходимо обосновать выбор дополнительных атрибутов и причину исключения каких-либо данных из исходного набора документов.

1.3 Предварительная обработка текстовых данных

Проведите предварительную обработку текста с помощью методов NLP: токенизацию, лемматизацию, выделение значимых частей речи, а также удаление стоп-слов, пунктуации, спецсимволов. Обоснуйте выбор методов предварительной обработки данных.

1.4 Подготовка отчета

Подготовьте отчет о проделанной работе, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip.

Модуль Б. Разведочный анализ данных

Время на выполнение модуля 3 часа.

Описание модуля:

В модуле необходимо провести анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, в том числе с использованием инструментов визуализации.

При выполнении модуля ставятся следующие цели:

Провести аналитическое исследование набора данных.

При выполнении данного модуля ставятся следующие задачи:

1. Выполнить семантический анализ: обработку естественного языка, сентиментный анализ, анализ текста;
2. Провести кластерный анализ данных.

Требования к оформлению письменных материалов

Письменный материал отсутствует.

Представление результатов работы

Результат выполнения Модуля Б «Разведочный анализ данных»: предобработанные данные (архив Data.zip), отчет о проделанной работе (Report_M2.html, Report_M2.ipynb), дополнительные комментарии коду (Readme.txt).

Необходимые приложения

Приложение 1: Архив, содержащий статьи о компаниях (Data.zip)

Приложение 2: Список номинантов конкурса (Condidates.doc)

ЗАДАНИЕ

2.1 Поиск ключевых слов/n-грамм. Векторизация текстов

Выполните поиск ключевых слов/биграмм/триграмм в тексте различными способами. Обоснуйте выбор алгоритмов поиска ключевых слов/биграмм/триграмм. Добавьте ключевые слова/биграммы/триграммы, как новые признаки в набор данных.

Преобразовать документы в векторные представления, к которым можно применить численное машинное обучение.

2.2 Тематическое моделирование

Выполните тематическое моделирование различными способами (не менее трех) и визуализируйте его результаты. Обоснуйте выбор алгоритмов тематического моделирования.

2.3 Кластеризация

Выполнить кластеризацию данных по сходству компаний несколькими способами (не менее трех). Выберите метрику оценки качества кластеризации. Обоснуйте выбор методов и приемов. Выполните визуальный анализ кластерных структур и оценки качества кластеризации. Определите лучший алгоритм кластеризации на основе выбранной метрики.

2.4 Разведочный анализ

Проведите анализ плотности распределения атрибутов и целевой переменной набора данных. Дайте интерпретацию полученных результатов.

Выполнить визуализацию пространства текстовых признаков различными способами. Визуализация должна отражать зависимости темы от временных признаков, рейтинга и ключевых слов/n-грамм. Также провести визуальный анализ статистики публикаций.

2.5 Подготовка отчета

Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip.

Модуль В. Построение, обучение и оптимизация модели

Время на выполнение модуля 3 часа.

Описание модуля:

В этом модуле продолжается работа с данными, подготовленными в предыдущей сессии. Требуется осуществить выбор алгоритма классификации, построить модель и провести оптимизацию полученной модели машинного обучения в контексте исследуемой задачи.

При выполнении модуля ставятся следующие цели:

Построение классификатора номинаций компаний-конкурсантов Премии Рунета.

При выполнении модуля 2 ставятся следующие задачи:

1. Выполнить разбиение выборки на обучающую и валидационную;
2. Осуществить построение моделей классификации;

3. Оценить качество полученных моделей в соответствии со спецификой решаемой задачи;

4. Выполнить оптимизацию лучшей модели.

Требования к оформлению письменных материалов

Письменный материал отсутствует.

Представление результатов работы

Результат выполнения Модуля «Построение, обучение и оптимизация модели»: результирующие файлы (архив Data.zip), отчет о проделанной работе (Report_M3.html, Report_M3.ipynb), дополнительные комментарии коду (Readme.txt).

Необходимые приложения

Приложение 1: Архив, содержащий статьи о компаниях (Data.zip)

Приложение 2: Список номинантов конкурса (Condidates.doc)

Приложение 3: Информация по номинациям для представленных организаций (Target.json)

ЗАДАНИЕ

3.1 Разбиение выборки

Выполните разбиение полученной выборки на обучающую и тестовую. Проведите обучение моделей, основанных на различных алгоритмах. Сделайте предсказание номинации организации на тестовой выборке. Выполните оценку моделей разной степени сложности в соответствии с выбранной метрикой. Определите модель, показавшую лучшее качество.

3.2 Оптимизация модели

Выполните настройку полученной модели уменьшив вычислительную сложность модели (выбор значимых признаков, понижение размерности). Оцените качество полученной модели, сделайте вывод.

Необходимо оптимизировать полученную модель под решаемую задачу, настраивая гиперпараметры. Выполните оценку данной модели после настройки гиперпараметров.

Построить для данной модели кривые валидации и обучения, интерпретируйте полученные результаты.

3.3 Подготовка отчета

Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков

кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip.

Модуль Г. Разработка программного продукта

Время на выполнение модуля 3 часа.

Описание модуля:

В этом модуле вы продолжаете работать с данными, подготовленными в предыдущей сессии. Вам предстоит выполнить прогноз для тестовой выборки. Также необходимо выполнить развертывание модели машинного обучения в рабочей среде в качестве API.

Какая-либо работа, обусловленная задачами предыдущей сессии, выполненная в ходе текущей, оцениваться не будет, поэтому проделывайте её только в случае необходимости.

При выполнении модуля ставятся следующие цели:

Выполнить развертывание модели машинного обучения, то есть обеспечение доступности модели на производстве, где веб-приложения, корпоративное программное обеспечение и API-интерфейсы могут использовать обученную модель, предоставляя новые данные и генерируя прогнозы.

При выполнении данного модуля ставятся следующие задачи:

1. Выполнить предсказание целевой переменной для контрольной выборки;
2. Разработать программный интерфейс для выполнения прогнозирования номинации компании-конкурсанта;
3. Разработать клиентское приложение с графическим интерфейсом.

Требования к оформлению письменных материалов

Письменный материал отсутствует.

Представление результатов работы

Результат выполнения Модуля Г «Разработка программного продукта»: результирующие файлы (архив Data.zip), отчет о проделанной работе

(Report_M4.html, Report_M4.ipynb), дополнительные комментарии коду (Readme.txt).

Необходимые приложения

Приложение 1: Архив, содержащий статьи о компаниях (Data.zip)

Приложение 2: Список номинантов конкурса (Condidates.doc)

Приложение 3: Информация по номинациям для представленных организаций (Target.json)

Приложение 4: Контрольная выборка (Control.xlsx)

ЗАДАНИЕ

4.1 Предсказание целевой переменной

Необходимо выполнить прогнозирование для контрольной выборки с помощью полученной модели. Запишите ответы, содержащие id компании-номинанта и соответствующие предсказания целевой переменной, в файл.

4.2 Разработка прикладного решения

Разработайте программный интерфейс для итоговой модели машинного обучения. API должен позволить приложению пользователя получать доступ к модели классификации для прогнозирования номинации конкурсанта на основе данных из Интернет-источников.

Разработать приложение с графическим интерфейсом, которое должно с помощью разработанного API генерировать прогнозы по новым данным в режиме реального времени. Приложение должно предоставлять справку по имеющимся командам и их параметрам.

4.3 Подготовка отчета

Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков

кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip.

Модуль Д. Разработка средств интеграции и поддержки готового решения

Время на выполнение модуля 3 часа.

Описание модуля 4: «Разработка средств поддержки готового решения»

В этом модуле вы продолжаете работать с программным продуктом, разработанным в предыдущей сессии. Вам предстоит разработать программную документацию к системе и доклад с презентацией о проделанной работе.

Какая-либо работа, обусловленная задачами предыдущей сессии, выполненная в ходе текущей, оцениваться не будет, поэтому проделывайте её только в случае необходимости.

При выполнении модуля 4 ставятся следующие цели:

Подготовка документации, сопровождающей разработанное программное обеспечение (ПО).

При выполнении данного модуля 4 ставятся следующие задачи:

1. Провести оформление технической документации;
2. Разработать пользовательскую документацию к системе;
3. Разработать презентацию по результатам работы

Требования к оформлению письменных материалов

Письменный материал отсутствует.

Представление результатов работы

Результат выполнения Модуля «Разработка средств поддержки готового решения»: техническая документация по системе (Report.doc), руководство пользователя (Rucovodstvo.doc), презентация результатов работы (Project.pptx).

Необходимые приложения

Необходимые приложения смотреть в папке «КОД 2.1 Приложения к вариантам».

Приложение 1: Архив, содержащий статьи о компаниях (Data.zip)

Приложение 2: Список номинантов конкурса (Condidates.doc)

Приложение 3: Информация по номинациям для представленных организаций (Target.json)

Приложение 4: Контрольная выборка (Control.xlsx)

ЗАДАНИЕ

4.1 Разработка технической документации по системе

При создании программы, одного лишь кода, недостаточно. Должен быть предоставлен некоторый текст, описывающий различные аспекты того, что именно делает код. Такая документация должна быть включена непосредственно в исходный код или предоставляется вместе с ним.

Необходимо разработать техническую документацию для определения и описания API, структур данных и алгоритмов.

Для стилевого оформления кода на языке Python использовать правила написания кода по PEP8.

4.2 Пользовательская документация

Для разработанного приложения и API составьте пользовательскую документацию, представляющую собой руководство пользователя. Руководство должно описывать каждую функцию программы, а также шаги, которые нужно выполнить для использования этой функции. Пользовательская документация должна предоставлять инструкции о том, что делать в случае возникновения проблем. Очень важно, чтобы документация не вводила в заблуждение и была актуальной. Руководство должно иметь чёткую структуру.

4.3 Презентация результатов работы

Необходимо создать презентацию, охватывающую все результаты выполнения задания. В ней должно быть указано ёмкое описание результатов работы с обоснованием выбора того или иного решения. Так же в презентации необходимо отразить скриншоты результатов своей работы. Разрабатывать презентацию рекомендуется в Power Point или аналогичной среде. Опишите перспективы улучшения Вашего решения.

4.4 Устный доклад

Подготовить устный доклад по результатам своей работы, включающие основные результаты по каждому модулю и выводы (не более 5 минут). Устные представления докладов – за 40 мин до окончания сессии.

2. Специальные правила компетенции[2]

Участники могут слушать музыку. Наушники и музыка в виде файлов должна быть предварительно сдана Техническому Эксперту для проверки. Принесенная музыка будет хранится на серверах для конкурсантов, к которым они будут иметь доступ.

[2] *Указываются особенности компетенции, которые относятся ко всем возрастным категориям и чемпионатным линейкам без исключения.*

В подготовительный день, конкурсантам разрешается принести карту памяти, содержащую не более 30 песен. Вся музыка будет упорядочена, проверена и распространена между всеми конкурсантами.

Оборудование не должно иметь доступ к внутренним устройствам для хранения информации. Организаторы соревнования проверят, что доступ был заблокирован.

Эксперты обладают правом запретить определенное оборудование в зоне конкурса.

Экспертам и переводчикам разрешается пользоваться личными компьютерами, планшетами или мобильными телефонами, находясь в помещении для экспертов, за исключением случаев, когда документы, относящиеся к соревнованию, находятся в комнате.

Экспертам и переводчикам разрешается пользоваться фото- и видеооборудованием, находясь в помещении для экспертов, за исключением случаев, когда документы, относящиеся к соревнованию, находятся в комнате, по согласованию с Главным экспертом.

Конкурсантам, экспертам и переводчикам разрешается использовать личные устройства для фото- и видеосъемки на рабочей площадке только после завершения конкурса.

2.1. Личный инструмент конкурсанта

Участники могут использовать защиту для ушей;

Участники могут принести с собой свои клавиатуры, мышки и коврики для мышек. Все принесенные клавиатуры, мышки и коврики должны быть предварительно сданы на проверку Техническому эксперту.

Запрещено:

Использование клавиатур и мышек с подключением по беспроводным каналам.

Устройства ввода не должны быть программируемыми.

2.2. Материалы, оборудование и инструменты, запрещенные на площадке

Участники не должны приносить:

Дополнительные программы

Мобильные телефоны

Портативные электронные устройства (планшеты, и т п)

Устройства для хранения информации (флэш-накопители, диски, и т.п)

3. Приложения

Приложение №1 Инструкция по заполнению матрицы конкурсного задания

Приложение №2 Матрица конкурсного задания

Приложение №3 Критерии оценки

Приложение №4 Инструкция по охране труда и технике безопасности по компетенции «Машинное обучение и большие данные».