Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт экономики и менеджмента

УТВЕРЖДАЮ:
Директор Института
экономики и менеджмента
_____ Е.В. Нехода
« 7 » 04 20 22 г.

Рабочая программа дисциплины

Интеллектуальный анализ данных

по направлению подготовки

38.04.01 Экономика

Направленность (профиль) подготовки:
«Экономика»

Форма обучения
Очная

Квалификация
Магистр

Год приема
2022

Код дисциплины в учебном плане: Б1.В.ДВ.01.02.05

СОГЛАСОВАНО:
Руководитель ОП
_____ Н.А. Скрыльникова

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины (модуля)

Целью освоения дисциплины является формирование следующих компетенций:
– ПК-1 – Способен определять направления развития организации.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ПК-1.4. Оценивает бизнес-возможности организации, необходимые для проведения стратегических изменений в организации.

2. Задачи освоения дисциплины

– Освоить математический аппарат, лежащий в основе методов интеллектуального анализа данных.

– Научиться применять аналитические платформы и библиотеки для решения практических бизнес-задач, связанных с анализом данных.

3. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, является обязательной для изучения. Относится к профессиональному модулю «Бизнес аналитика».

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Семестр 2, зачет.

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: «Премодуль (Бизнес-аналитика)», «Премодуль (Вероятностные и статистические методы в бизнес-аналитике)», «Python и R для анализа данных», «Эконометрика», «Системное и критическое мышление».

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:
– лекции: 8 ч.;
– практические занятия: 20 ч.;
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины (модуля), структурированное по темам

Тема 1. Введение в интеллектуальный анализ данных.

Основные понятия и терминология анализа данных. Отраслевой стандарт CRISP-DM. Классификация задач машинного обучения. Примеры бизнес-задач, для решения которых может быть использовано машинное обучение. Базовые операции при работе с данными. ABC, XYZ, FRM, RFM-анализ данных.

Тема 2. Задача регрессии.

Постановка задачи. Методы решения задачи регрессии на основе линейных регрессионных моделей, нейронных сетей и деревьев решений. Оценка качества регрессионных моделей.

Тема 3. Задача классификации.

Постановка задачи классификации. Методы решения задачи классификации на основе логистической регрессии, нейронных сетей, деревьев решений. Методы оценки качества классификаторов.

Тема 4. Задача кластеризации.

Постановка задачи кластеризации. Методы решения задачи кластеризации на основе иерархических и агломеративных алгоритмов. Сети Кохононена.

Тема 5. Методы работы с несбалансированными наборами данных.

Метрики качества работы классификаторов. Методы исправления дисбаланса: undersampling, oversampling, ансамбли моделей, на основе стоимостей ошибок.

Тема 6. Методы отбора признаков.

Методы на основе фильтрации. Прямые и обратные методы включения/исключения признаков. Методы регуляризации и методы на основе деревьев решений.

Тема 7. Метода подготовки признаков.

Цели операции подготовки признаков. Заполнения пропусков. Кодирование категориальных переменных. Преобразования и масштабирование переменных. Дискретизация. Обработка выбросов.

Тема 8. Методы оптимизации гиперпараметров.

Цели и задачи процедуры оптимизации гиперпараметров. Кросс-валидация. Поисковые алгоритмы. Байесовская оптимизация.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости и выполнения домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет с оценкой во втором семестре состоит из двух частей.

Первая часть предполагает письменный ответ на вопрос экзаменационного билета.

Вторая часть предполагает защиту подготовленного проекта по выбранной теме.

Билет содержит один теоретический вопрос. Продолжительность теоретического экзамена 1,5 часа.

Примерный перечень теоретических вопросов, проверяющих ПК-1.4:

1. Сформулируйте задачу регрессии и кратко перечислите подходы к её решению. Опишите алгоритм решения задачи регрессии на основе модели множественной линейной регрессии.

2. Сформулируйте задачу регрессии и кратко перечислите подходы к её решению. Опишите алгоритм решения задачи регрессии k-ближайших соседей. Опишите критерии для выбора оптимального значения параметра k.

3. Сформулируйте задачу регрессии и кратко перечислите подходы к её решению. Опишите алгоритм решения задачи регрессии на основе искусственной нейронной сети. Опишите принципы для выбора количества скрытых слоев в нейронной сети и количества нейронов на каждом слое.

4. Сформулируйте задачу регрессии и кратко перечислите подходы к её решению. Опишите методы учёта нелинейных зависимостей и качественных факторов при использовании модели множественной линейной регрессии.

5. Сформулируйте задачу регрессии и кратко перечислите подходы к её решению. Перечислите основные проблемы (выбросы, гетероскедастичность, мультиколлинеарность), возникающие при использовании модели множественной линейной регрессии, и способы их решения.

6. Сформулируйте задачу регрессии и кратко перечислите подходы к её решению.

7. Сформулируйте задачу классификации и кратко перечислите подходы к её решению. Опишите алгоритм решения задачи классификации на основе линейного дискриминантного анализа.

8. Сформулируйте задачу классификации и кратко перечислите подходы к её решению. Опишите алгоритм решения задачи классификации на основе модели множественной логистической регрессии. Как интерпретируются коэффициенты модели логистической регрессии?

9. Сформулируйте задачу классификации и кратко перечислите подходы к её решению. Опишите алгоритм решения задачи классификации на основе линейного дискриминантного анализа.

10. Сформулируйте задачу классификации и кратко перечислите подходы к её решению. Опишите алгоритм решения задачи классификации на основе квадратичного дискриминантного анализа.

11. Сформулируйте задачу классификации и кратко перечислите подходы к её решению. Перечислите методы оценки качества и сравнения классификаторов.

12. Перечислите подходы к оценке качества решений задач регрессии и классификации на основе методов построения повторных выборок, назовите их достоинства и недостатки.

13. Перечислите методы регуляризации алгоритмов в задачах регрессии и классификации. Опишите применение подходов оценки качества решений на основе методов построения повторных выборок в задаче регуляризации алгоритмов обучения.

14. Опишите подходы к регуляризации алгоритмов в задачах регрессии и классификации на основе методов понижения размерности. Опишите алгоритм регрессии на главные компоненты.

15. Перечислите подходы к учёту нелинейности в задачах классификации и регрессии, основанные на полиномиальных, ступенчатых, сплайнах и обобщённых аддитивных моделях.

16. Опишите алгоритмы построения деревьев решений в задачах классификации и регрессии. Укажите, в каких ситуациях, решения, основанные на деревьях решений, предпочтительнее.

17. Опишите подходы к решению задач классификации и регрессии, на основе алгоритмов Bagging, RandomForest и Boosting.

18. Опишите подходы к решению задач классификации и регрессии на основе алгоритма SVM.

19. Сформулируйте задачу кластеризации и опишите алгоритм k-means.

20. Сформулируйте задачу кластеризации и опишите алгоритм иерархической кластеризации.

21. Сформулируйте задачу кластеризации и опишите её решение на основе сетей Кохонена.

22. Перечислите основные методы отбора признаков. Укажите их достоинства, недостатки и условия применения.

23. Перечислите основные методы построения признаков. Приведите примеры.

24. Опишите подходы к решению задачи поиска оптимальных значений гиперпараметров.

Проект, выносимый на защиту, направлен на проверку ПК-1.4 и предполагает:

1. Самостоятельный выбор темы проекта.
2. Сбор и подготовку исходных данных по теме проекта.
3. Формулировку проверяемых гипотез.
4. Разработку решения.
5. Анализ результатов и их бизнес-интерпретация.

6. Подготовку презентации и доклад с описанием основных элементов решения и полученных результатов.

Результаты зачета определяются оценками «зачтено», «не зачтено».

Оценка «зачтено» выставляется при выполнении следующих условий:

1. Ответ на теоретический ответ является полным, содержит постановку проблемы, описание методов её решения с указанием их достоинств, недостатков и условий применимости.

2. Практический проект направлен на решение задачи определения развития организации, является практически полезным. Проект должен содержать оценку текущего состояния проблемы (формулировку в бизнес-терминах), формальную постановку задачи, описание доступных источников данных (доступность, качество), решение проблемы, описание процедуры внедрения полученного решения, включая оценку бизнес-возможностей организации, необходимых для проведения стратегических изменений, и достигаемого экономического эффекта.

Оценка «не зачтено» выставляется при выполнении любого из следующих условий:

1. Ответ на теоретический ответ не является полным, не содержит постановку проблемы, описания методов её решения с указанием их достоинств, недостатков и условий применимости.

2. Практический проект, направленный на решение задачи определения развития организации, не содержит оценку текущего состояния проблемы (формулировку в бизнес-терминах), формальную постановку задачи, решения проблемы, описание процедуры внедрения полученного решения, включая оценку бизнес-возможностей организации, необходимых для проведения стратегических изменений, и достигаемого экономического эффекта.

11. Учебно-методическое обеспечение

а) электронный учебный курс по дисциплине в электронном университете «Moodle»; https://moodle.tsu.ru/course/view.php?id=16476

б) оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Рассел М. Data Mining : извлечение информации из Facebook, Twitter, Linkedin, Instagram, GitHub / Мэтью Рассел, Михаил Классен ; [пер. с англ. А. Киселев]. - 3-е изд.. - Санкт-Петербург [и др.] : Питер, 2020. - 459, [3] с.

– О'Нил К. Data Science : инсайдерская информация для новичков, включая язык R / Кэти О'Нил, Рэйчел Шатт ; [пер. с англ. И. Пальти и др.]. - Санкт-Петербург [и др.] : Питер, 2019. - 362, [2] с

– Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / Петер Флах ; [пер. с англ. А. А. Слинкина]. - Москва : ДМК Пресс, 2015. - 399 с.

– Foundations and novel approaches in data mining / edited by Tsau Young Lin [a. o.]. - Berlin [a. o.] : Springer, 2006. - x, 376 p.: ill. - (Studies in computational intelligence / ed. by Janusz Kacprzyk ;Vol. 9:)

б) дополнительная литература:

– Келлехер Д. Д. Основы машинного обучения для аналитического прогнозирования : алгоритмы, рабочие примеры и тематические исследования : пер. с англ. / Джон Д. Келлехер, Брайан Мак-Нейми, Аоифе д'Арси. - Санкт-Петербург [и др.] : Диалектика, 2019. - 656 с.

в) ресурсы сети Интернет:
– http://www.machinelearning.ru/wiki/

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:
– Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);
– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

б) информационные справочные системы:
– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system
– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index
– ЭБС Лань – http://e.lanbook.com/
– ЭБС Консультант студента – http://www.studentlibrary.ru/
– Образовательная платформа Юрайт – https://urait.ru/
– ЭБС ZNANIUM.com – https://znanium.com/
– ЭБС IPRbooks – http://www.iprbookshop.ru/

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.
Аудитории для проведения практических занятий, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.
Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

15. Информация о разработчиках

Богданов Александр Леонидович, к.т.н., доцент, ИЭМ ТГУ, доцент кафедры Информационных технологий и бизнес-аналитики.