МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Одеська національна академія зв'язку ім.. О. С. Попова
Кафедра «Інформаційних технологій»

«Затверджую»
Ректор ОНАЗ ім. О. С. Попова
_____ П. П. Воробієнко
«__»_____ 20___ р.

Засоби Data Mining в інфокоммунікаціях

ПРОГРАМА

навчальної дисципліни
підготовки бакалаврів
(назва освітньо-кваліфікаційного рівня)

спеціальності 122 Комп'ютерні науки
(шифр і назва спеціальності)

Одеса
2019 рік

РОЗРОБЛЕНО ТА ВНЕСЕНО: _Одеська національна академія зв'язку ім.. О. С. Попова

(повне найменування вищого навчального закладу)

Розробники програми: __*д.т.н., проф. Романюк В.В., к.т.н., доц. Єгошина Г.А.*

Програму розглянуто і схвалено на засіданні кафедри ІТ

Протокол № від "__" _____ 20___ р. Зав. кафедрою ІТ _____

Програму погоджено з кафедрами:

_____ _____ /_____/

_____ _____ /_____/

Програму розглянуто і схвалено Радою навчально-наукового інституту інфокомунікацій та програмної інженерії

Протокол № __ від "__"_____ 20___ р. Директор ННІ ІКПІ _____ проф. І.В. Стрелковська

Програму розглянуто і схвалено методичною радою Одеської національної академії зв'язку ім. О.С.Попова

Протокол №___ від "___"_____ 20___ р. Голова ради _____

/_____/

1. ВСТУП

Програма вивчення навчальної дисципліни “Засоби Data Mining в інфокомунікаціях” складена відповідно до освітньо-професійної програми підготовки бакалаврів спеціальності *122 Комп’ютерні науки*

Предметом дисципліни є вивчення сучасних технологій, пакетів та бібліотек для вирішення задач аналізу даних – Data Mining. У курсі розглядаються задачі аналізу даних, проводиться огляд обчислюваних та статичних методів, вивчаються сучасні технології та бібліотеки мови програмування Python для дослідження Big Data, приділяється увага алгоритмам ефективного зберігання, обробки та візуалізації Big Data.

Програма навчальної дисципліни складається з таких змістових модулів:

1. Вступ до Data Mining.
2. Бібліотека NumPy.
3. Маніпуляції над даними за допомогою бібліотеки Pandas.
4. Візуалізація засобами бібліотеки MatPlotLib.

3. Мета та завдання навчальної дисципліни

Метою викладання навчальної дисципліни є навчання та засвоєння студентами основних теоретичних відомостей та практичних вмінь з курсу. Підготувати студента до ефективного використання як класичних, так і сучасних методів інтелектуального аналізу даних та обробки інформації.

Завдання дисципліни – оволодіння студентами науковими основами, сучасною методологією та особливостями застосування інтелектуальної обробки даних; засвоєння майбутніми фахівцями теоретичних основ інформаційних систем, орієнтованих на застосування стандартів Data Mining; набуття уміння програмувати окремі елементи систем Data Mining різного призначення і різної проблемної орієнтації на всіх стадіях життєвого циклу інформаційної системи; отримання практичних навичок використання і адаптації деяких найбільш відомих систем та бібліотек Data Mining.

Цілі курсу:

- опанувати базові принципи використання IPython, дистрибутиву Anaconda та блокноту Jupiter;
- опанувати роботу з типами даних в мові Python та ознайомитися з особливостями списків, кортежів та словників;
- ознайомитися з атрибутами, особливостями індексації та зрізів масивів бібліотеки NumPy;
- навчитися ефективно проводити операції злиття та розбиття масивів;
- навчитися ефективно використовувати масиви бібліотеки NumPy;
- навчитися ефективно використовувати універсальні функції для роботи з масивами бібліотеки NumPy;
- навчитися ефективно проводити операції транслювання, порівняння, масок (в тому числі використовувати булеві масиви в якості масок);

- навчитися ефективно проводити сортування масивів бібліотеки NumPy;
- ознайомитися з основними об’єктами бібліотеки Pandas;
- ознайомитися з особливостями індексації та зрізів бібліотеки Pandas;
- навчитися ефективно проводити обробку відсутніх даних;
- ознайомитися з особливостями ієрархічної індексації;
- ознайомитися з особливостями проведення операцій об’єднання наборів даних засобами бібліотеки Pandas;
- навчитися ефективно проводити операції агрегування та групування та працювати з засобами бібліотеки Pandas для зведених таблиць;
- ознайомитися з особливостями роботи з бібліотекою MatPlotLib;
- навчитися ефективно проводити візуалізацію даних засобами бібліотеки MatPlotLib.

В результаті успішного засвоєння навчальної дисципліни студент матиме змогу продемонструвати такі результати навчання:

знати:

- базові принципи використання IPython, дистрибутиву Anaconda та блокноту Jupiter;
- атрибути, особливості індексації та зрізів масивів бібліотеки NumPy;
- методи NumPy для операцій злиття та розбиття масивів;
- універсальні функції для роботи з масивами бібліотеки NumPy;
- методи NumPy для операцій транслювання та порівняння;
- методи NumPy для сортування масивів;
- основні об’єкти бібліотеки Pandas;
- особливості індексації та зрізів бібліотеки Pandas;
- методи бібліотеки Pandas для проведення обробки відсутніх даних;
- особливості ієрархічної індексації бібліотеки Pandas;
- особливостями роботи з бібліотекою MatPlotLib та її основні для візуалізації даних.

вміти:

- ефективно проводити операції злиття та розбиття масивів;
- ефективно використовувати масиви бібліотеки NumPy;
- ефективно використовувати універсальні функції для роботи з масивами бібліотеки NumPy;
- ефективно проводити операції транслювання, порівняння, масок (в тому числі використовувати булеві масиви в якості масок);
- ефективно проводити сортування масивів бібліотеки NumPy;
- ефективно проводити обробку відсутніх даних засобами бібліотеки Pandas;
- ефективно проводити операції агрегування та групування та працювати з засобами бібліотеки Pandas для зведених таблиць;
- використовувати засоби бібліотеки MatPlotLib в задачах візуалізації даних.

Вивчення навчальної дисципліни передбачає формування та розвиток у студентів **компетентностей**:

***загальних*:**

- здатність до абстрактного мислення, аналізу та синтезу
- здатність застосовувати знання у практичних ситуаціях;
- знання та розуміння предметної області та розуміння професійної діяльності;
- здатність вчитися й оволодівати сучасними знаннями;
- здатність до пошуку, оброблення та аналізу інформації з різних джерел;
- здатність генерувати нові ідеї (креативність);
- здатність працювати в команді;
- здатність приймати обґрунтовані рішення;
- здатність оцінювати та забезпечувати якість виконуваних робіт.

фахових:

- здатність до виявлення статистичних закономірностей недетермінованих явищ, застосування методів обчислювального інтелекту, зокрема статистичної, нейромережевої та нечіткої обробки даних, методів машинного навчання та генетичного програмування тощо;
- здатність до логічного мислення, побудови логічних висновків, використання формальних мов і моделей алгоритмічних обчислень, проектування, розроблення й аналізу алгоритмів, оцінювання їх ефективності та складності, розв’язності та нерозв’язності алгоритмічних проблем для адекватного моделювання предметних областей і створення програмних та інформаційних систем;
- здатність використовувати сучасні методи математичного моделювання об’єктів, процесів і явищ, розробляти моделі й алгоритми чисельного розв’язування задач математичного моделювання, враховувати похибки наближеного чисельного розв’язування професійних задач;
- здатність до системного мислення, застосування методології системного аналізу для дослідження складних проблем різної природи, методів формалізації та розв’язування системних задач, що мають суперечливі цілі, невизначеності та ризики;
- здатність застосовувати теоретичні та практичні основи методології та технології моделювання для дослідження характеристик і поведінки складних об'єктів і систем, проводити обчислювальні експерименти з обробкою й аналізом результатів;
- здатність проектувати та розробляти програмне забезпечення із застосуванням різних парадигм програмування: узагальненого, об’єктно-орієнтованого, функціонального, логічного, з відповідними моделями, методами й алгоритмами обчислень, структурами даних і механізмами управління;
- здатність до інтелектуального аналізу даних на основі методів обчислювального інтелекту включно з великими та погано структурованими даними, їх-

ньої оперативної обробки та візуалізації результатів аналізу в процесі розв'язування прикладних задач;

- здатність використовувати методи Data Mining для аналізу та структуризації багатовимірних даних і подальшого їх використання при розв'язанні прикладних задач.

Результати навчання даної дисципліни деталізують такі **програмні результати навчання**:

- застосовувати знання основних форм і законів абстрактно-логічного мислення, основ методології наукового пізнання, форм і методів вилучення, аналізу, обробки та синтезу інформації в предметній області комп'ютерних наук;
- використовувати сучасний математичний апарат неперервного та дискретного аналізу, лінійної алгебри, аналітичної геометрії, в професійній діяльності для розв'язання задач теоретичного та прикладного характеру в процесі проектування та реалізації об'єктів інформатизації;
- використовувати знання закономірностей випадкових явищ, їх властивостей та операцій над ними, моделей випадкових процесів та сучасних програмних середовищ для розв'язування задач статистичної обробки даних і побудови прогнозних моделей;
- використовувати методи обчислювального інтелекту, машинного навчання, нейромережевої та нечіткої обробки даних, генетичного та еволюційного програмування для розв'язання задач розпізнавання, прогнозування, класифікації, ідентифікації об'єктів керування тощо;
- проектувати, розробляти та аналізувати алгоритми розв'язання обчислювальних та логічних задач, оцінювати ефективність та складність алгоритмів на основі застосування формальних моделей алгоритмів та обчислюваних функцій;
- розробляти програмні моделі предметних середовищ, вибирати парадигму програмування з позицій зручності та якості застосування для реалізації методів та алгоритмів розв'язання задач в галузі комп'ютерних наук;

застосовувати методи та алгоритми обчислювального інтелекту та інтелектуального аналізу даних в задачах класифікації, прогнозування, кластерного аналізу, пошуку асоціативних правил з використанням програмних інструментів підтримки багатовимірного аналізу даних на основі технологій DataMining, TextMining, WebMining.

На вивчення навчальної дисципліни відводиться 180 годин / 6 кредитів ECTS.

2. Інформаційний обсяг навчальної дисципліни

Змістовий модуль 1. Вступ до Data Mining.

Тема 1. Поняття Data Mining. Порівняння статистики, машинного навчання та Data Mining. Data Mining як частина ринку інформаційних технологій. Сфери використання Data Mining. Дані: набір даних та їх атрибутів. Виміри. Формати зберігання даних. Класифікація видів даних. Метаданні

Тема 2. Методи та стадії Data Mining. Задачі Data Mining. Знання. Співставлення інформації, даних та знань. характеристика методів, які використовуються для вирішення задачі класифікації.

Змістовий модуль 2. Бібліотека NumPy.

Тема 3. Стандартні типи даних бібліотеки NumPy.

Тема 4. Масиви бібліотеки NumPy, створення, індексація та зрізи.

Тема 5. Зміна форми масивів. Злиття та розбиття.

Тема 6. Обчислення засобами NumPy, універсальні функції, агрегування.

Тема 7. Транслювання масивів. Порівняння, булеві маски, комбінована індексація.

Тема 8. Сортування масивів засобами NumPy.

Змістовий модуль 3. Маніпуляції над даними за допомогою бібліотеки Pandas.

Тема 9. Основні об’єкти бібліотеки Pandas: Series, DataFrame.

Тема 10. Індексація та вибірка даних для Series та DataFrame.

Тема 11. Об’єднання датасетів: проста конкатенація, злиття, розбиття, використання ключа злиття.

Тема 12. Агрегування та группування. GroupBy: розбиття, використання, об єднання.

Тема 13. Pivot Tables у Pandas: синтаксис, створення «вручну».

Змістовий модуль 4. Візуалізація засобами бібліотеки MatPlotLib.

Тема 14. Імпорт бібліотеки MatPlotLib, налаштування стилів.

Тема 15. Прості лінійні графіки засобами MatPlotLib.

*Тема 16.*Діаграми розсіювання у MatPlotLib.

Тема 17. Візуалізація погрішностей засобами MatPlotLib. Гістограми, розбиття по інтервалам.

Тема 18. Налаштування користувача: легенди, відображення осей, карти кольорів.

Тема 19. Субграфіки та трьохмірні графіки засобами MatPlotLib.

Тема 20. Бібліотека Seaborn. Порівняння з MatPlotLib.

Тема 21. Візуалізація засобами Seaborn

3. Рекомендована література

1. Дюк В. Data Mining: учебный курс / Дюк В., Самойленко А. — СПб.: Изд. Питер, 2001. — 368 с.

2. Барсегян А. А. Методы и модели анализа данных: OLAP и Data Mining / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко та ін. – 2-е изд., перераб. и доп. – СПб. : БХВ-Петербург, 2004. – 336 с.

3. Сегаран Т. Программируем коллективный разум / Т. Сегаран. – СПб.: Символ-Плюс, 2008. – 368 с.

4. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с.: ил.

5. Дэви С. Основы Data Science и BigData. Python и наука о данных // С. Дэви, М. Арно, А. Мохамед. — СПб.: Питер, 2017. — 336 е.: ил.

6. Інтелектуальний аналіз даних: Підручник / Черняк О.І., Захарченко П.В. / К.: Знання, 2014р. – 599 с.

7. Рашка С. Python и машинное обучение / С. Рашка. – М.: ДМК-Пресс, 2017. – 418 с.

8. Han J. Data Mining: Concepts and Techniques (Second Edition) / J. Han, M. Kamber. – Morgan Kaufmann Publishers, 2006. – 800 p.

9. Witten, I. H. Data mining : practical machine learning tools and techniques / Ian H. Witten, Frank Eibe, Mark A. Hall. – 3rd ed. – Morgan Kaufmann Publishers, 2011. – 630 p.

Інформаційні ресурси

- http://docs.scipy.org/doc/
- www.anaconda.com
- numpy.org
- pandas.pydata.org
- colab.research.google.com
- www.jetbrains.com\pycharm

4. Форма підсумкового контролю успішності навчання

_____залік,іспит_____

5. Засоби діагностики успішності навчання

1. Поточний контроль знань з лекційного матеріалу;
2. Залік, іспит.