МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КРЕМЕНЧУЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ МИХАЙЛА ОСТРОГРАДСЬКОГО

МЕТОДИЧНІ ВКАЗІВКИ
ЩОДО ВИКОНАННЯ РОЗРАХУНКОВО-ГРАФІЧНОЇ РОБОТИ
З НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
«ОСОВИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ»
ДЛЯ СТУДЕНТІВ ДЕННОЇ ФОРМИ НАВЧАННЯ
ЗІ СПЕЦІАЛЬНОСТІ 123 – «КОМП'ЮТЕРНА ІНЖЕНЕРІЯ»

КРЕМЕНЧУК 2018

Методичні вказівки щодо виконання розрахунково-графічної роботи з навчальної дисципліни «Основи інтелектуального аналізу даних» для студентів денної форми навчання зі спеціальності 123 – «Комп’ютерна інженерія».

Укладач к. т. н., доц. В. М. Сидоренко

Рецезент д. т. н., професор М. І. Гученко

Кафедра комп'ютерних та інформаційних систем

Затверджено методичною радою КрНУ імені Михайла Остроградського
Протокол № _____від_____ 2015 р.

Голова методичної ради _____ проф. В. В. Костін

ЗМІСТ

ВСТУП

Дисципліна «Основи інтелектуального аналізу даних» входить до циклу вибіркових дисциплін магістерської програми.

Мета розрахунково-графічної роботи (РГР) – узагальнити знання та навички, набуті під час опрацювання лекційного матеріалу та виконання лабораторних робіт змістового модуля 1 «Базовий. Основи Data Science-проекту. Розвідувальний аналіз даних» і змістового модуля 2 «Моделювання та розгортання Data Science-проекту». Методичні вказівки дозволяють студентам виконати аналіз і синтез типового Data Science-проекту відповідно до міжнародного стандарту CRISP DM. Ураховуючи, що індивідуальне завдання з РГР студент отримує на початку вивчення навчальної дисципліни і відпрацьовує під час виконання лабораторних робіт його окремі частини (наскрізне завдання), завдання РГР полягає в тому, щоб «зібрати докупи» напрацьовані частини за допомогою сучасного інструментарію: мови програмування R, IDE RStudio, RMarkdown, Shiny та інших, що належать до екосистеми R.

Це надає можливість системно підійти до створення доволі складних закінчених проектів, наближених до реальної практики, у стислі терміни в умовах малої кількості аудиторних годин.

Мета і завдання навчальної дисципліни: надати студентам знання і прищепити практичні навички з основ сучасних технологій інтелектуального аналізу даних.

Місце навчальної дисципліни у навчальному процесі: дисципліна базується на знаннях та вміннях, отриманих студентами під час вивчення таких навчальних дисциплін: «Теорія ймовірностей та математична статистика», «Теорія інформації та кодування», «Алгоритми та методи обчислень», «Організація баз даних», «Програмування», «Моделювання», «Організація обчислювальних процесів», «Обробка сигналів та зображень», «Експертні

системи та системи штучного інтелекту», «Комп'ютерні системи», «Комп'ютерні мережі» та ін.

У результаті вивчення дисципліни студент повинен

знати: теоретичні основи технології Data Mining та засоби реалізації Data Science-проектів;

уміти:

розробляти закінчені програмно-аналітичні рішення (так званий Data Science-проект) згідно зі стандартом CRISP DM на базі мови програмування R у середовищі IDE RStudio із застосуванням фреймворків RMarkdown, Shiny та ін.

в рамках Data Science-проекту:

– виконувати імпорт-експорт даних з різних джерел, включаючи реляційні БД;

– маніпулювати даними на рівні інтерфейсу командного рядка на етапах розвідувального аналізу даних та підготовки даних до аналізу;

– виконувати візуалізацію даних і результатів аналізу із застосуванням табличних і графічних візуалізаторів;

– виконувати побудову моделей інтелектуального аналізу: кластеризації, класифікації та регресії на основі статистичного та машинного навчання;

– виконувати розгортання розробленого рішення у вигляді веб-застосунку із використанням клієнт-серверних технологій.

1 ЗАГАЛЬНІ ПОЛОЖЕННЯ ЩОДО ПОСТАНОВКИ ЗАВДАННЯ ТА СТРУКТУРИ РОЗРАХУНКОВО-ГРАФІЧНОЇ РОБОТИ

Мета: набуття навичок розробки повноцінного Data Science-проекту і відповідності до фаз стандарту CRISP DM.

1.1 Загальні положення щодо постановки завдання та структури розрахунково-графічної роботи

1.1.1 Загальна постановка завдання

Кожен варіант розрахунково-графічної роботи (РГР) передбачає розробку Data Science-проекту згідно з методологією CRISP DM (wikipedia 2018). Згідно зі стандартом CRISP-DM 1.0 життєвий цикл проекту Data Mining має складатися із шести фаз (рис. 1):

– розуміння бізнес-процесів (business understanding);

– розуміння даних (data understanding);

– підготовка даних (data preparation);

– моделювання (modeling);

– оцінювання (evaluation);

– розміщення (deployment).

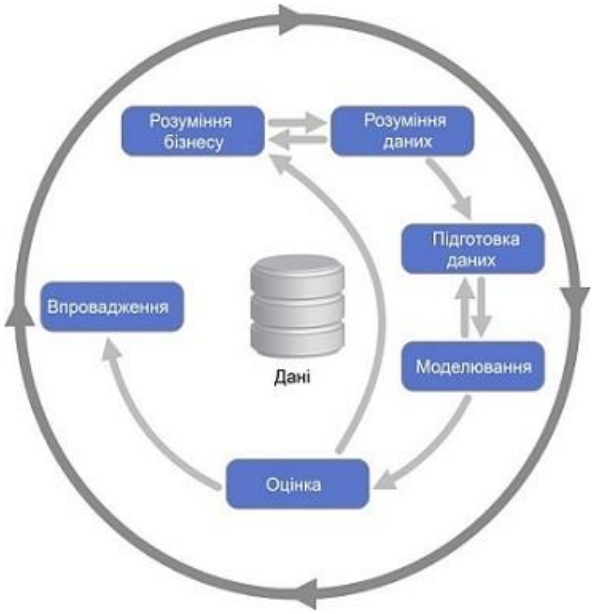

Рис. 1 – Життєвий цикл процесу Data Mining згідно з методологією CRISP

Фаза розуміння бізнес-процесів такі завдання:

– визначення бізнес-цілей;

– визначення ситуації;

– визначення цілей Data Mining;

– створення плану проекта.

Фаза розуміння даних передбачає такі завдання:

– первинне збирання даних;

– опис даних;

– вивчення даних;

– перевірка якості даних.

Фаза підготовки даних включає в себе всі дії, пов’язані з остаточним формуванням набору даних для аналізу. При цьому розв’язуються п’ять завдань:

– вибір даних;

– очищення даних;

– конструювання даних;

– інтеграція даних;

– форматування даних.

Фаза моделювання призначена для вибору оптимального методу побудови моделей і настроювання його параметрів для отримання оптимальних рішень. На цій фазі розв’язуються такі завдання:

– вибір методу моделювання;

– генерація тестового проекту;

– створення моделей;

– оцінювання моделей.

Фаза оцінювання покликана призвана більш ґрунтовно оцінити модель до процесу її остаточного розміщення, щоб упевнитись у досяжності поставлених бізнес-цілей. Ця фаза передбачає такі завдання:

– оцінювання результатів;

– перегляд процесу;

– визначення подальших дій.

Фаза розміщення передбачає розгортання моделі.

Якщо виокремити із цього процесу суто технологічну складову, то типова технологічна основа будь-якого Data Science-проекту має виглядати так [3] (рис. 2).

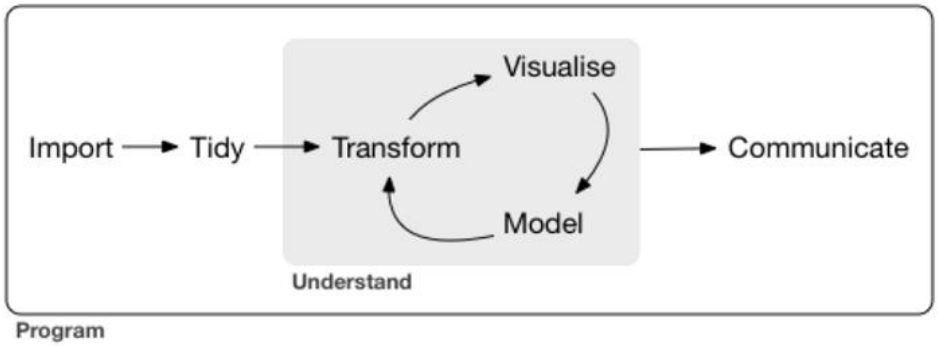

Рис. 2 – Структура типового Data Science-проекту

Перша задача *імпорту даних* (**Import**) полягає у вилученні необхідних «сирих» даних з будь-яких джерел (файли, БД, дані з датчиків у реальному часі та ін.) найрізноманітнішого формату.

Друга задача (**Tidy**) – *приведення даних до* так званого *«охайного» вигляду*, придатного для аналізу. Зазвичай йдеться про приведення даних до табличного вигляду «ключ-значення», або, іншими словами, «об’єкт-ознака». Ця процедура є частиною більш узагальненої задачі підготовки даних до аналізу (**Wrangling, Munging**), яка включає у себе інші процедури, наприклад, заповнення пропущених значень (Missing Value Emputation), вилучення неінформативних даних (data reduction), різного роду трансформації (**Transforming**) та ін. Тобто, **Tiding+Transforming=Wrangling(Munging)**.

Тріада задач *трансформація–візуалізація–моделювання* (**Transform-Visualise-Model**) складають ядро Data Mining-процесу, суть якого полягає в пошуку нетривіальних практично корисних закономірностей у даних, за дпопмогою висування та перевірки гіпотез, побудови різного роду моделей, який обов’язково супроводжується візуалізацією як проміжних, так і кінцевих результатів моделювання, що надається замовнику. Саме ця тріада задач

забезпечує фази розуміння (**Understanding**) даних, моделей і паттернів, що знаходяться в даних і створюють основу «монетизованого» продукту інтелектуального аналізу даних (**Data Product**).

Останній етап – надання проміжних чи остаточних результатів інтелектуального аналізу даних (**Communicate**) іншим членам команди, які задіяні у проекті, у зручному для сприйняття вигляді з використанням прози, таблиць, графіків і коду.

Для реалізації типового Data Science-проекту необхідний відповідний інструментарій, який би був гнучким і відповідав усім необхідним вимогам, яких потребують вищезгадані задачі, і надавав би можливість розв'язати будь-яку задачу без визначних затрат часу, насамперед на маніпулювання даними та підготовку результатів.

Мова і середовище програмування R з великим арсеналом програмних пакетів, IDE RStudio та технологія так званого *грамотного програмування* дозволяє ефективну реалізацію всіх етапів типового Data Science-проекту, об'єднавши всі задачі в єдине ціле.

1.1.2 Структура розрахунково-графічної роботи

На першому етапі виконання РГР студент, згідно з отриманим варіантом, має виконати коротку і чітку постановку завдання за кожною з наведених вище фаз, обґрунтувавши свої інженерні рішення.

На другому етапі створюється скелет Data Science-проекту у середовищі IDE R Studio з використанням мови програмування R, фреймворка RMarkdown, мови розмітки даних LaTex та підключивши необхідні для того бібліотеки. Структура динамічного документа має відповідати назвам фаз перших **п'яти фаз** життєвого циклу процесу Data Mining згідно з методологією CRISP.

На третьому етапі формується звіт у форматах .html.nb, .doc та .pdf.

На четвертому етапі виконується реалізація **п'ятої фази** – розгортання проекту за допомогою фреймворка Shiny у вигляді інтерактивного веб-застосунку.

1.1.3 Структура звіту

На перевірку студент має подати РГР у наступному складі:

1. Data Science-проект, який містить такі файли:

– rmd – RMarkdown-динамічний документ;

– html.nb – html-звіт у форматі R-Notebook;

– pdf – звіт у форматі pdf;

– docx – звіт у форматі Word.

2. Файли з первинними даними (підкаталог `data`), рисунками (підкаталог `image`), додатковими файлами (підкаталог `doc`) та ін.

1.2 Матеріали довідкового характеру

1.2.1 Концепція грамотного програмування

Подання проміжних чи остаточних результатів проекту може бути виконано у тому чи іншому вигляді – звіту, презентації, методичних вказівок, наукової статті тощо, в одному з поширених форматів – .doc, pdf, .html тощо. Протягом багатьох років у наукових та ділових колах стандартом де-факто є застосування так званої парадигми *грамотного програмування* для підготовки електронних документів з використанням у тому числі і потужніх засобів комп'ютерної графіки. З виникненням і розвитком Data Science методологію грамотного програмування було взято на озброєння і реалізоване практично в кожному потужному інструменті Data Science.

Грамотне програмування (Literate Programming) – концепція, методологія програмування і документування, у якій програма складається з прози природною мовою упереміж з макропідстановками та кодом мовами програмування.

В основу технології грамотного програмування покладено поняття *динамічного документа* – текстового документа, який складається з тексту та коду, з використанням необхідних мов програмування, який дозволяє згенерувати власне електронний документ заданого формату. Для цього використовуються як можливості мов розмітки документів (напр. Markdown,

YAML, HTML, LaTeX), так і можливості доступу до потужних програмних бібліотек, призначених для обробки даних та комп'ютерної графіки.

Отже, логічно мати певне програмне середовище, яке дозволить поєднати низку таких технологій разом і створити зручний інтерфейс розробника. Існують різні програмні засоби і середовища, що дозволяють реалізувати технологію грамотного програмування.

У лабораторній роботі пропонується низка наразі актуальних і популярних інструментів для створення динамічних документів (рис. 3), як от:

– IDE RStudio як інтегроване середовище розробки;

– спеціалізовану мову програмування R та арсенал її потужних бібліотек для маніпулювання даними та візуалізації результатів;

– фреймворк RMarkdown – для підготовки динамічних звітів мовою розмітки Markdown [1, 4];

– мову розмітки даних LaTex для високоякісного оформлення наукових документів.

Рис. 3 – Структура процесу підготовки електронного наукового чи ділового документа в парадигмі грамотного програмування (literate programming)

1.2.2 Markdown і RMarkdown

Markdown (МФА: [маркдаун]) – полегшена мова розмітки даних, яку створено з урахуванням прочитності та зручності у публікації з подальшим перетворенням її на structurally valid XHTML або HTML.

Такі сайти, як GitHub, Reddit та Stack Overflow, використовують Markdown для полегшення обговорень між користувачами.

RMarkdown [7] – фреймворк R, який дозволяє створювати динамічні Markdown-документи у середовищі IDE RStudio у стилі грамотного програмування з використанням усіх можливих потужностей мови R та її бібліотек. Дозволяє реалізовувати інтерфейс так званих ноутбуків для створення документів з текстом і кодом разом для виготовлення елегантно відформатованого виводу. Дозволяє використовувати декілька мов, включаючи R, Python, C++, HTML, SQL, Stan. Через конвертор Pandoc дозволяє здійснювати вивід у html-, doc- або pdf-формат у вигляді веб-сторінок, брошур, буклетів, слайдів.

1.2.3 Інсталяція R

Для цього необхідно зайти на CRAN, скачати і встановити актуальну версію R. Цей дистрибутив R має свій GUI, однак його можливості досить обмежені. Тут, у розділі ‘Contributed’, також можна знайти безліч цікавої літератури, написаної різними мовами. Одне з найкоротших і доступних введень у мову R можна знайти на сторінці Дмитра Храмова [8]. Зокрема, ознайомлення з елементами базової графіки.

1.2.4 Інсталяція RStudio

Для зручної роботи і відлагодження програм, зокрема роботи з фреймворком RMarkdown для створення динамічного документу, необхідно встановити IDE RStudio.

1.2.5 Створення RMarkdown-документа

1. Завантажити RStudio.

2. Створити RMarkdown-документ у форматі R Notebook, вибравши відповідний пункт меню.

1.2.6 Генерація електронного документу

Генерація електронного документа здійснюється натисканням комбінації *Ctrl+Alt+K*.

2 КРИТЕРІЇ ОЦІНЮВАННЯ ЯКОСТІ ВИКОНАННЯ РОЗРАХУНКОВО-ГРАФІЧНОЇ РОБОТИ СТУДЕНТАМИ

Розрахунково-графічну роботу студенти виконують у 1-му семестрі. Максимальна кількість балів, яку отримують студенти за виконання РГР, складає 15 балів: 5 – «задовільно», 10 – «добре», 15 – «відмінно».

Шкала оцінювання: національна та ECTS

<table>
<tr><td rowspan="2">Сума балів за всі види навчальної діяльності</td><td rowspan="2">Оцінка ECTS</td><td colspan="2">Оцінка за національною шкалою</td></tr>
<tr><td>Для іспиту, курсового проекту (роботи), практики</td><td>Для заліку</td></tr>
<tr><td>90 – 100</td><td>A</td><td>Відмінно</td><td rowspan="5">Зараховано</td></tr>
<tr><td>82 – 89</td><td>B</td><td rowspan="2">Добре</td></tr>
<tr><td>74 – 81</td><td>C</td></tr>
<tr><td>64 – 73</td><td>D</td><td rowspan="2">Задовільно</td></tr>
<tr><td>60 – 63</td><td>E</td></tr>
<tr><td>35 – 59</td><td>FX</td><td>Незадовільно з можливістю повторного складання</td><td>Не зараховано з можливістю повторного складання</td></tr>
<tr><td>0 – 34</td><td>F</td><td>Незадовільно з обов’язковим повторним вивченням навч. дисципліни</td><td>Не зараховано з обов’язковим повторним вивченням навч. дисципліни</td></tr>
</table>

СПИСОК ЛІТЕРАТУРИ

1. Belousovv.ru. 2013. “Syntaxis Markdown.” Article. http://belousovv.ru/markdown_syntax#fnref:2.

2. Chan, Chung-hong, Geoffrey CH Chan, Thomas J. Leeper, and Jason Becker. 2018. *Rio: A Swiss-Army Knife for Data File I/O*.

3. Garrett Grolemund, Hadley Wickham. 2018. *R for Data Science*. http://r4ds.had.co.nz/index.html.

4. sandino. 2013. “Cheat Sheet of Markdown.” Article. https://github.com/sandino/Markdown-Cheatsheet#emphasis.

5. Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

6. wikipedia. 2018. “Cross-Industry Standard Process for Data Mining.” Article. https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining.

7. Yihui Xie, Garrett Grolemund, J. J. Allaire. 2018. *R Markdown: The Definitive Guide*. https://bookdown.org/yihui/rmarkdown/.

8. Страница Дмитрия Храмова. 2015. *Введение в R.* http://dkhramov.dp.ua/Comp.R.html#.W56ISegzY2z

9. Hadley Wickham, Lionel Henry, Romain Francois. 2018a. “Dplyr Part of Tidyvers. Overview.” Documentation. https://dplyr.tidyverse.org/.

10. ———. 2018b. “Dplyr: A Grammar of Data Manipulation.” Documentation. https://cran.r-project.org/web/packages/dplyr/.

11. ———. 2018c. “Introduction to Dplyr. Translation of Andrey Ogurtsov.” Documentation. http://rpubs.com/aa989190f363e46d/dplyr_intro.

12. RStudio. 2018. “Databases Using Dplyr.” Documentation. https://db.rstudio.com/dplyr/.

13. Wickham, Hadley. 2014. “Tidy Data.” https://www.jstatsoft.org/article/view/v059i10.

14. Casas, Pablo. 2018. Data Science Live Book. https://livebook.datascienceheroes.com/.

15. ———. 2018b. “Exploratory Data Analysis.” Article. https://en.wikipedia.org/wiki/Exploratory_data_analysis.

16. ———. 2018c. “Principal Component Analysis.” Article. https://en.wikipedia.org/wiki/Principal_component_analysis.

Методичні вказівки щодо виконання розрахунково-графічної роботи з навчальної дисципліни «Основи інтелектуального аналізу даних» для студентів денної форми навчання зі спеціальності 123 – «Комп'ютерна інженерія».

Укладач к. т. н., доц. В. М. Сидоренко

Відповідальний за випуск зав. кафедри КІС А. В. Луговой

Підп. до др. _____. Формат 60×84 1/16. Папір тип. Друк ризографія.

Ум. друк. арк. ____. Наклад _____ прим. Зам. №_____. Безкоштовно.

Видавничий відділ
Кременчуцького національного університету
імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600