МІНІСТЕРСТВО НАУКИ І ОСВІТИ УКРАЇНИ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

**ПРАКТИЧНЕ ЗАНЯТТЯ № 1 ЗА ТЕМОЮ:**

**Класифікаційні правила та дерева рішень**

Група \_\_\_\_\_11\_\_\_\_\_\_

Курс \_\_\_\_\_1\_\_\_\_\_\_\_\_\_

Студент (ка) \_\_\_\_Супруненко М. І. \_\_\_\_

Дата оформлення \_\_\_\_\_\_28.02.2025\_\_\_\_\_\_\_

Перевірив \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

МЕТА: Метою практичної роботи є формування професійних вмінь та навичок щодо використання інструментів (методів, бібліотек, пакетів програмного забезпечення R) для побудови дерев рішень (класифікації), вміння застосовувати отримані знання на практиці в практичних задачах аналізу даних.

ТЕОРІЯ: Дерево рішень - це графік (граф), який представляє варіанти та результати рішень (подій) у вигляді дерева, або дозволяє здійснити класифікацію сутностей, які досліджуються. Вузли на графу представляють подію чи вибір, а гілки графа представляють правила чи умови прийняття рішення. В основному використовується у додатках машинного навчання та інтелектуального аналізу даних в задачах прикладної аналітики, які використовують R. Детально теоретичні питання розглянуті в курсі лекцій.

Дерева рішень — це універсальний алгоритм машинного навчання, який може виконувати завдання класифікації, і регресії. Це дуже потужні алгоритми, здатні проаналізувати дуже складні набори даних. Крім того, дерева рішень є фундаментальними компонентами випадкових лісів, які є одними із найпотужніших інструментів машинного навчання.

Детально теоретичні питання розглянуті в курсі лекцій.

Бібліотеки та функція, які використовуються в лабораторній роботі:

library(dplyr)

library(rpart)

library(rpart.plot)

Основний синтаксис для створення дерева рішень у R.

ctree(formula, data)

де formula - це формула, що описує предиктор та змінні відповіді.

Data - ім'я набору даних, що використовується.

В роботі використовується вбудований набір даних R з іменем readingSkills для створення дерева рішень. Він описує оцінку чиєїсь навички читання, якщо ми знаємо змінні «вік», «розмір взуття», «оцінка» і те, чи є людина носієм мови чи ні. Набір не є реальним дослідженням, а лише абстрактний набір даних для легкого засвоєння методики побудови дерев.

ПЛАН ПРОВЕДЕННЯ ПРАКТИЧНОГО ЗАНЯТТЯ:

1. Відповіді на вхідні питання.
2. Проведення досліджень гіперпараметрів функції побудови дерева рішень в середовищі R.
3. Підготовка звіту за результатами дослідження.
4. Підготовка домашнього завдання.
5. Захист домашнього завдання.

ПИТАННЯ ДЛЯ ВХІДНОГО КОНТРОЛЮ СТУДЕНТІВ:

1. Призначення дерев класифікації. Визначення та основна ідея.
2. Термінології дерева рішень.
3. Ідея методу дерев рішень.
4. Методи вибору атрибутів.
5. Приріст інформації та ентропія.
6. Інформаційна ентропія.
7. Приріст інформації.
8. Деякі додаткові функції та характеристики індексу Джіні.

ХІД ПРОВЕДЕННЯ ПРАКТИЧНОГО ЗАНЯТТЯ:

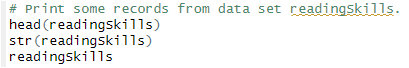
1. Інсталюйте необхідні для роботи з деревами пакети та бібліотеки. Ви також повинні встановити залежні пакети, якщо такі є.

install.packages("party")

library(party)

Пакет «party» має функцію ctree(), яка використовується для створення та аналізу дерева рішень.

1. Завантажте бібліотеку «party». Прочитайте заголовки змінних у наборі даних readingSkills, визначте і опишіть їх формат. За допомогою довідки опишіть набір даних, що означають змінні, скільки досліджень включено до набору. Чи є це реальний набір, чи тільки навчальний.



Опис змінних та їх формат:

…..

1. За допомогою функції (help \_\_\_ або ?\_\_\_) виведіть довідку по всі параметрам функції ctree(). Здійсніть дослідження параметрів та опишіть їх властивості.

Функція ctree() використовується для побудови умовного дерева рішень. Ось ключові параметри:

…..

controls: об'єкт класу TreeControl, створений за допомогою функції ctree\_control. Визначає параметри, що впливають на процес побудови дерева.

Опис функції ctree\_control ():

…

1. Побудуйте дерево за допомогою функції ctree().

model\_1 <- ctree()

plot(model\_1)

1. Змініть параметри функції ctree() в частині опису формули та додаткових параметрів, опишіть отримані зміни.

model\_2 <- ctree( ……., control = ctree\_control( ))

model\_2 <- ctree( ……., control = ctree\_control( ))

1. Зробіть висновки за отриманими результатами моделювання, які фактори є значущими для отримання максимальної кількості балів з мовного тесту. Чи знайшли ви якісь цікаві залежності.

ДОМАШНЄ ЗАВДАННЯ:

Обрати будь-який навчальний набір (самостійний вибір), який можна проаналізувати з використанням дерева рішень. Здійсніть побудову дерева, зробіть висновки. Підготуйтеся до захисту дослідження.

**Практична робота 1**

1. **Ініціалізація середовища**

|  |
| --- |
| # Встановлення необхідних пакетів (виконується один раз)  install.packages("party")  install.packages("rpart")  install.packages("rpart.plot")  install.packages("dplyr")  install.packages("partykit")  install.packages("ggparty")  # Підключення бібліотек  library(party)  library(rpart)  library(rpart.plot)  library(dplyr)  library(partykit)  library(ggparty) |

1. **Завантаження та перегляд набору даних**

|  |
| --- |
| # Завантаження вбудованого набору даних readingSkills  data("readingSkills", package = "party")  # Виведення перших рядків даних  head(readingSkills)  # Перевірка структури набору даних  str(readingSkills)  # Опис змінних у наборі даних  summary(readingSkills) |
|  |

1. **Вивчення довідки по функції ctree()**

|  |
| --- |
| ?ctree  # Вивести документацію про функцію ctree() |
| Recursive partitioning for continuous, censored, ordered, nominal and multivariate response variables in a conditional inference framework.  ctree(formula, data, subset = NULL, weights = NULL, controls = ctree\_control(), xtrafo = ptrafo, ytrafo = ptrafo, scores = NULL)  formula: a symbolic description of the model to be fit. Note that symbols like ':' and '-' will not work and the tree will make use of all variables listed on the rhs of 'formula'. data: a data frame containing the variables in the model. subset: an optional vector specifying a subset of observations to be  ...  ### have fun browseURL(file.path(tmpdir, "party", "inst", "documentation", "html", "index.html")) ## End(Not run) |

1. **Побудова базового дерева рішень**

|  |
| --- |
| # Побудова дерева рішень  model\_1 <- ctree(nativeSpeaker ~ age + shoeSize + score, data = readingSkills)  # Візуалізація дерева  plot(model\_1) |
|  |

1. **Налаштування гіперпараметрів**

|  |
| --- |
| # Налаштування параметрів побудови дерева  control\_params <- ctree\_control(minsplit = 5, minbucket = 3, maxdepth = 3)  # Побудова дерева рішень  model\_2 <- ctree(nativeSpeaker ~ age + shoeSize + score,                   data = readingSkills,                   control = control\_params)  # <-- Виправлено!  # Візуалізація дерева  plot(model\_2) |
|  |

1. **Аналіз отриманих результатів**

|  |
| --- |
| # Перевірка важливості предикторів  print(model\_2)  # Інтерпретація факторів, які найбільше впливають на визначення nativeSpeaker |
|  |

1. **Домашнє завдання**

|  |
| --- |
| # Використання іншого набору даних (наприклад, iris)  data("iris")  # Побудова дерева рішень для набору iris  model\_iris <- ctree(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)  # Візуалізація  plot(model\_iris) |
|  |

1. **Перевірка даних Iris**

|  |
| --- |
| # Виведення перших рядків даних  head(iris)  # Перевірка структури набору даних  str(iris)  # Опис змінних у наборі даних  summary(iris) |
|  |

1. **Спроба іншої бібліотеки візуалізації за допомогою rpart (фаворит на сьогодні)**

|  |
| --- |
| # Побудова дерева за допомогою rpart  model\_rpart <- rpart(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris, method = "class")  # Візуалізація дерева  rpart.plot(model\_rpart, type = 4, extra = 101, under = TRUE, tweak = 1.2) |
|  |

1. **Побудова для Native Speaker**

|  |
| --- |
| # Побудова дерева за допомогою rpart  model\_rpart\_2 <- rpart(nativeSpeaker ~ age + shoeSize + score,                          data = readingSkills,                          method = "class",                          control = rpart.control(minsplit = 5, minbucket = 3, maxdepth = 3))  # Візуалізація дерева  rpart.plot(model\_rpart\_2, type = 4, extra = 101, under = TRUE, tweak = 1.2) |
|  |

1. **Інший метод візуалізації (ggparty)**

|  |
| --- |
| # Створюємо дерево  model\_rpart <- rpart(nativeSpeaker ~ age + shoeSize + score,                       data = readingSkills,                       control = rpart.control(minsplit = 5, minbucket = 3, maxdepth = 3))  # Конвертуємо у partykit  model\_party <- as.party(model\_rpart)  ggparty(model\_party) +    geom\_edge() +    geom\_edge\_label() +    geom\_node\_label(aes(label = splitvar), ids = "inner") +    geom\_node\_label(aes(label = info), ids = "terminal") |
|  |

**Висновок:**

У результаті виконання практичної роботи було здобуто цінний досвід у побудові дерев рішень за допомогою інструментів R. Використання функцій з бібліотеки rpart дозволило глибше ознайомитися з методами класифікації та регресії, що можуть бути застосовані в реальних задачах аналізу даних. Найбільше задоволення принесло застосування функцій для налаштування гіперпараметрів та дослідження їх впливу на результати моделювання.

Завдяки проведеній роботі вдалося краще зрозуміти, які фактори впливають на ефективність дерева рішень, а також як правильно налаштовувати параметри для отримання точних та зрозумілих моделей. Використання дерев рішень для аналізу реальних та навчальних наборів даних дозволяє отримувати не тільки високоточні прогнози, але й інтерпретувати логіку моделі, що є важливим аспектом у галузі машинного навчання.

Таким чином, виконання роботи сприяло не лише розвитку навичок роботи з бібліотеками R, а й поглибленню розуміння застосування методів дерев рішень у практичних задачах аналізу даних.