Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
факультет соціології
кафедра методології та методів соціологічних досліджень

КВАЛІФІКАЦІЙНА РОБОТА

на тему:

«ЗАСТОСУВАННЯ МЕТОДУ ВИПАДКОВОГО ЛІСУ ДЛЯ КЛАСИФІКАЦІЇ ВСТУПНИКІВ ЗВО НА ОСНОВІ АНАЛІЗУ ВСТУПНИХ ЗАЯВ»

Галузь знань: 054 «Соціологія»
Освітня програма: «Соціологія»
Освітній ступінь: бакалавр
Кваліфікація: бакалавр соціології

Виконавець:
Жлуктенко Юлія Русланівна,
студентка 4 курсу

Науковий керівник:
Сидоров Микола Володимир-Станіславович,
кандидат фізико-математичних наук, доцент

Бакалаврська робота допущена до захисту
рішенням кафедри *методології та методів соціологічних досліджень*

Протокол №_____ від «___» _____ 20__ р.

Зав. кафедри _____ канд. ф. – м. наук Сидоров М.В.– С.
підпис

Київ 2020

Реєстрація

номер | дата | підпис лаборанта кафедри

Рекомендовано до захисту

підпис наукового керівника | ініціали, прізвище наукового керівника

Результат захисту

оцінка | дата захисту

Голова ЕК

підпис | ініціали, прізвище

Члени ЕК

підпис | ініціали, прізвище

підпис | ініціали, прізвище

підпис | ініціали, прізвище

підпис | ініціали, прізвище

Секретар ЕК

підпис | ініціали, прізвище

Зміст

Вступ

Актуальність. Завдання класифікації об’єктів дослідження - є одним з найбільш поширених на сьогодні. Чимало замовників, серед яких як і комерційні, так і некомерційні (громадські або ж державні) організації бажають краще дізнатися об’єкти, вивченням яких вони зацікавлені. Для здійснення такої вимоги невід’ємною частиною аналізу і є процес класифікації, завдяки якому надалі можна виробляти стратегії проведення піар кампаній, більш точно виокремлювати цільову аудиторію чи просто більш поглиблено розумітися на об’єктах, що піддаються вивченню.

Відповідно до цього у дослідників постає чимало питань, наприклад, таких як: який метод є найбільш доречним; за якими критеріями правильно здійснювати класифікацію; чи дають можливість виокремлені класи робити подальші прогнози, тощо.

Повертаючись до першого з них - зрозумілим є те, що вимоги сучасного суспільства (особливо зважаючи на його технологічний розвиток та наявні можливості) полягають у більш швидкому та автоматизованому процесі, який не буде потребувати для власного втілення багатьох додаткових зусиль від дослідника. Перевага надається таким методам, що завдяки чітко налагодженому, автоматизованому процесу, за доволі невеликий проміжок часу, надають якісний результат для подальшого аналізу.

Так, одним із доволі популярних методів класифікації серед іноземних (здебільшого західних) дослідників і науковців, окрім класичних методів кластерного аналізу, є метод з використанням дерев прийняття рішень. Один з різновидів зазначеного методу - метод випадкового лісу, який повністю задовольняє потребу у надійності наданих результатів. До того ж, побудований він на актуальних алгоритмах машинного навчання, які враховують специфіку наявних даних. Якщо ж брати до уваги саме українських дослідників - даний метод не є широковідомим серед них,

оскільки багато хто залишається прихильниками класичних та перевірених часом методів.

Виходячи з цього, актуальність даної роботи пояснюється необхідністю постійного покращення та спрощення наявних методів аналізу не тільки з метою надання більш точних та якісних результатів за менший проміжок часу, але й також з метою розвитку вітчизняної науки, наближення її до сучасності. До того ж, американські та європейські дослідники доволі часто користуються авторитетом в українському науковому просторі.

Теоретичним об'єктом роботи є методи здійснення класифікації об'єктів/даних;

Предмет: застосування методу випадкового лісу для класифікації вступників до закладів вищої освіти.

Метою роботи є розкриття можливостей та потенціалу методу випадкового лісу для завдань класифікації об'єктів дослідження (власне, на прикладі класифікації вступників до закладів вищої освіти за допомогою аналізу їхніх вступних заяв).

Для виконання мети роботи, були виокремлені наступні **завдання:**

1. Розкриття змісту поняття класифікації як методу аналізу даних;
2. Виокремлення особливостей, переваг та недоліків методу випадкового лісу;
3. Практичне застосування методу випадкового лісу для здійснення класифікації об'єктів;

Розділ 1. Класифікація об’єктів. Методи машинного навчання

Для більш детального розуміння поняття класифікації (перед тим як переходити до безпосередньо до математичних та статистичних методів її створення) необхідно деяку увагу приділити концептуалізації та інтерпретації безпосередньо даного поняття.

Розділ 1.1. Підходи до визначення поняття класифікації

Класифікація - одна з тих задач, що зустрічається доволі часто у процесі аналізу. Вона має вирішальне значення для розуміння об’єктивної реальності та передбачає групування об’єктів у групи або класи на основі їхньої подібності. А вже таке групування вже надає можливості подальшого розуміння наявної структури [Bailey, 1994], [Bailey, 2005].

На сьогодні визнаним є той факт, що класифікація - це невід’ємний дослідницький етап для покращення розуміння дослідницького простору, однак незважаючи на це, протягом історичного періоду (та й досі) велися й ведуться численні дебати з приводу найкращого методу для здійснення класифікації, критеріїв, на основі яких потрібно виокремлювати об’єкти. Також іноді під питання виноситься й загальна мета, з якою взагалі може здійснюватися класифікація [Lambert, 2015].

Ще з давніх часів, щойно наука стала активно розвиватися, а продукти людського існування активно досліджуватися, історики намагалися «здійснити упорядкування навколишнього хаосу» та протягом останніх століть біологи все частіше усвідомлювали нагальну потребу класифікації об’єктів у відповідності до загальної, поширеної та прийнятної у науковому середовищі схеми класифікації, яка полегшує процедуру назви об’єктів і робить їх визначеними та єдиними для всіх. Так, вивчення різнорідних, не схожих один на одного об’єктів, дало поштовх для здійснення таксономічних досліджень, що стало основою для подальших біологічних досліджень [Huxley, 2007].

Однак не лише біологія визнає цінність загальних класифікаційних схем (як теоретично виокремлених, так і емпірично). Так, дослідники в області організаційних, математичних, інформаційних, комп'ютерних, поведінкових та соціальних наук активно використовують їх для своїх сфер дослідження.

Без наявності певного консенсусу стосовно класифікації об'єктів у відповідній галузі є неможливим процес накопичення знань та поєднання аналізу статистичними методами результатів окремих досліджень, а розвиток теорії відбувається виключно на великих масштабах. Класифікації дозволяють вивчити та зробити узагальнення щодо дискретних однорідних груп об'єктів і, зрештою, дають можливість перейти до теорій середнього радіусі дії, що, у свою чергу, застосовуються лише до тих дискретних груп об'єктів [Rich, 1992].

Під час дослідження об'єктів широко використовуються дві теорії класифікації: есенціалізм та емпіризм. Застосування тієї чи іншої теорії залежить від мети класифікації.

Есенціалізм випливає з арістотелівської точки зору, що існує кілька істотних характеристик, які визначають сутність організму, і що, ідентифікуючи ці характеристики, можна створити класи організмів. Класи, засновані на невеликій кількості характеристик, які вважаються істотними для визначення сутності групи, називаються монотетичними (гомогенними) групами. Об'єкти, аби бути умовними «членами» виокремленої групи, повинні володіти характеристиками, які використовуються для визначення власне цієї групи. Володіння такою характеристикою є і достатнім, і необхідним для членства в групі [Bailey, 1994], [McKelvey, 1982]. Класифікації, що є продуктом есенціалістичної філософії, називають типологіями. Типології можуть мати форму традиційної (загальної) або теоретичної класифікації [Rich, 1992]. Традиційні класифікації залежать від неявного розпізнавання зазначених категорій, оскільки немає чітких

класифікаційних критеріїв. Ґрунтуються вони на широких подібностях та відмінностях, що є очевидними для дослідників [Warriner, 1984].

Традиційні класифікації корисні для виявлення та надання назв процесам та предметам, що існують у реальному світі. Типологізації, з іншого боку, розробляються на основі попередньо існуючої теорії. Дослідник концептуалізує та виокремлює «типи», що мають відношення до дослідження, і визначає декілька характеристик, які представляють сутність об’єкта. Це у свою чергу має пряме відношення до мети класифікації. Типології здебільшого породжуються за допомогою якісної класифікації, а не кількісного аналізу, хоча вони можуть бути сформовані шляхом концептуалізації типів, а потім аналізу результатів за допомогою статистичних методів. Теоретичні класифікації у свою чергу можуть взагалі не мати емпіричних еквівалентів і можуть бути ідеальними типами або ж взагалі повністю гіпотетичними [Baden-Fuller, 2013]. Забезпечення невеликої кількості визначальних характеристик узгоджується з есенціалістичною філософією про те, що існує лише кілька характеристик, які фіксують та визначають сутність об'єкта. У тому випадку, якщо ж дослідники мають справу з великою кількістю визначальних характеристик, як правило есенціалістичний підхід не є придатним та правильним для цього.

На відміну від есенціалізму, емпіризм базується на принципах адансонівської теорії класифікацій, у якій кожній властивості елемента присвоюється рівне значення, і, відповідно, утворюються поліетинчні групи. Такі групи мають найбільшу кількість спільних символьних станів, і жоден єдиний стан не є істотним для членства в групі, або достатнім для того, щоб зробити об'єкт членом групи [Sneath, 1973]. Класифікації, що є продуктом емпіричної філософії, називаються таксономіями. Таксономія - це емпірично виведена класифікація об'єктів на основі сукупності їх спостережуваних характеристик. Термін таксономія також

використовується для позначення теоретичного дослідження класифікації, включаючи її основи, принципи, процедури та правила [Simpson, 1961]. Під час розробки класифікаційних схем дослідники здійснюють таксономічну діяльність, але їх результат, тобто фактичні класифікаційні схеми, можуть бути як типологіями (специфічні класифікації), так і таксономіями (загальні класифікації). Емпірично отримана класифікація стала відомою як числова таксономія [Sneath, 1973]. Числові таксономії оцінюють спорідненість між об’єктами, яка потім виражається за допомогою числових метрик (для чого існує чимало методів). Таксони (категорії) створюються на основі великої кількості характеристик, які зазвичай називають змінними. Згодом об'єкти впорядковуються відповідно до їх ступеня спорідненості [McKelvey, 1982], [Sneath, 1973]. Апріорі всі характеристики мають однакову значимість, а схожість між предметами - це функція подібності між кожним на основі безлічі індивідуальних характеристик.

Таксономія може слугувати загальною класифікацією об'єктів, з якої можна зробити узагальнення, запропонувати гіпотези та, врешті-решт, розробити теорію середнього радіусу дії, оскільки саме зв'язок із емпіричною реальністю дозволяє розробляти перевірену, релевантну та дійсну теорію [Eisenhardt, 1989]. За допомогою великої кількості змінних потенційно зменшується упередженість дослідника, чого не можна сказати про типології. Тим не менш все одно є необхідність прийняття багатьох суб’єктивних рішень. Під час проведення пошукових досліджень, коли ще доступно доволі мало інформації про об’єкт, дослідник повинен аналізувати дані, використовуючи стільки змінних, скільки отримає на практиці. Небезпека такого підходу полягає в тому, що найбільш вагомі змінні можуть залишитися не поміченими, а ті, що не мають жодної цінності, можуть домінувати. У результаті, отримана класифікація може бути статистично достовірною, але не може бути інтуїтивно зрозумілою або корисною.

Отже, есенціалістичні та емпіричні теорії класифікації мають деякі відмінності в систематичних підходах. Практична значимість класів, що виокремлюється на основі вище зазначених підходів також відрізняється. Типологія, розроблена з певним призначенням, базується лише на кількох характеристиках і тому має обмежену корисність. Таксономії ж навпаки, є результатом групування об'єктів на основі сукупності багатьох спостережуваних характеристик. Хоча багато дослідників використовують терміни взаємозамінно, вони не є рівнозначними: типології та таксономії мають свої переваги та недоліки [Lambert, 2015].

Розділ 1.2. Класифікація об’єктів як метод аналізу даних

Тож, зрозумівши, що класифікація об’єктів є дійсно невід’ємним елементом під час аналізу у багатьох випадках, дещо зосередимо увагу на її особливості саме під час практичного втілення. Перш за все, для цього виділимо те визначення, яке найбільше підходить до теми роботи, а також до її емпіричної частини: класифікація – внутрішньо системний розподіл об’єктів чи предметів, що вивчаються, за істотними ознаками для зручності їхньої подальшої інтерпретації; певне впорядкування початкових понять, що вказує на ступінь їхньої схожості. Ключовим моментом є те, що класи, на які розподіляються об’єкти, є заздалегідь відомими, а отже процес класифікації є керованим методом машинного навчання, можливі варіанти якого будуть описані у наступному підрозділі [Чубукова, 2008].

Для здійснення класифікації необхідним є дотримання певних правил, які забезпечать щонайменше правильність цього процесу [Чубукова, 2008]:

- Кожний новий поділ вимагає застосування відповідно однієї основи (ознаки);

- Має бути забезпечена пропорційність розподілу: тобто узагальнений обсяг видових ознак мусить бути еквівалентним обсягу розподіленої ознаки;
- Одиниці розподілу повинні бути взаємовиключними: одна одиниця може бути віднесеною виключно до одного класу;
- Необхідно забезпечити послідовність розподілу.

До завдань класифікації часто відносять прогнозування номінальної або порядкової залежної змінної на основі вибірки безперервних і/або категоріальних незалежних змінних. У зв'язку з цим виділяють бінарну класифікацію та множинну. Для першого варіанту характерним є те, що змінна може приймати виключно два значення (наприклад приналежності: вона або належить, або ні). Натомість множинна класифікація допускає те, що змінна може мати своє значення з безлічі визначених класів. У такому випадку розглядається безліч класів для залежної змінної. Класифікуватися об'єкти можуть як за однією ознакою (одновимірна класифікація, есенціалістичний підхід), так і за багатьма (багатовимірна класифікація, емпіричний підхід) [Чубукова, 2008].

Однією з цілей процесу класифікації є побудова такої модель, яка б використовувала прогнозуючі параметри у якості вхідних параметрів і на основі цього отримувала б значення відповідного залежного атрибута [Sukumaran, 2013].

Основними етапами, що характерні для здійснення процесу класифікації є наступні: конструювання моделі та її застосування.

1. Конструювання моделі: характеристика множини створених класів.

- Кожен набір даних відноситься до одного класу;
- Для здійснення цього етапу використовується тестувальна вибірка, на якій і відбувається власне побудова моделі;

- Підсумкова модель представляється у вигляді математичної формули, дерева рішень або ж пояснюється класифікаційними правилами.

2. Практичне застосування моделі: класифікація нових або попередньо невідомих значень.

Щонайменше виділяють два способи перевірки класифікаційного методу на точність:

1. Перший варіант – розподіл масиву на тренувальну та тестову підвибірки (у співвідношенні 70 на 30 відсотків. Після тренування методу (на тренувальному масиві) відбувається порівняння із тим, наскільки точним було передбачення на тестовій вибірковій сукупності.
2. Перевірка точності класифікації та її оцінка проводитися за допомогою крос-перевірки (крос-валідації) на даних з тестової множини. Згодом результати перевірки тестової множини порівнюються із результатами перевірки навчальної. Якщо зазначені класифікації мають приблизно схожі результати вважається, що дана модель пройшла перевірку на точність. Поділ на навчальну і тестову множини здійснюється шляхом пропорційного та взаємовиключного ділення вибірки. Однак цей спосіб є доречним для використання для достатньо великих вибірок. Якщо ж розмір вибірки не задовольняє цієї умови, рекомендованими до застосування є поділ масиву на вибірки, що не є рівнозначними за об’ємом.

Точною вважається класифікація, що відповідає нижче зазначеним вимогам:

- Значення тестової множини, що є відомими, зіставляються з результатами використання отриманої моделі;

- Рівень точності відповідає відсотку вірно класифікованих прикладів в тестовій множині;
- Не допускається існування залежності між тестовими та навчальними множинами (для великих масивів).

Якщо точність моделі оцінюється як позитивна (кількість правильно спрогнозованих кейсів значно вища від половини) модель може бути використаною для класифікації нових об’єктів, класова приналежність яких не відома [Чубукова, 2008].

Розділ 1.3. Некеровані та керовані методи машинного навчання

Перед тим, як переходити до розгляду власне можливостей створення класифікацій за допомогою статистичних методів, варто зазначити, що будь-який з них базується на заздалегідь визначених алгоритмах машинного навчання, що й забезпечують можливість його використання.

Зокрема ці методи машинного навчання (або ж методу штучного розпізнавання нейронних зв’язків) можна розподілити на два види: контрольовані та неконтрольовані.

Можливість контролювати та задавати процес «тренування» методу визначається, як не дивно, наявною в дослідника інформацією. Так, якщо дослідження, що проводилося, було експлораторним, то, ймовірніше за все, дослідник лише матиме на меті дізнатися структуру отриманих даних. Простіше кажучи, він не матиме достатньої кількості необхідної інформації для виокремлення окремих структурних елементів у своєму масиві [Sathya, 2013]. Тож, зрозумілим виходячи з цього є те, що тут у нагоді стануть некеровані методи машинного навчання.

Нейронні мережі, що самоорганізуються під час навчання, використовують некеровані алгоритми аби ідентифікувати та більш детально повідомити дослідника про структуру заданої ознаки. Така «некерованість» полягає у відсутності необхідності заздалегідь визначати

можливу помилку для оцінки потенційного рішення. Відсутність чіткого спрямування такого методу має перевагу у тому сенсі, що модель може сама знайти такі взаємозв’язки, про які дослідник може навіть й не здогадуватися [Kohonen, 1996].

Основними характеристиками некерованого підходу є наступні особливості:

1. Він перетворює схему початкових даних довільного розміру в одну або двомірну карту і виконує це адаптивно (підлаштовуючись під заданий масив);
2. Мережа відображає структуру, що складається з впорядкованих нейронів, що розташовані у стовпцях та рядках;
3. Нейрони, що мають спільні риси - розташовані відповідно близько один до одного та поєднані між собою синаптичними зв’язком.

Даний вид машинного навчання є характерним для використання у багатьох випадках, таких як от, наприклад, кластеризація чи розпізнавання мови.

Керовані методи машинного навчання базуються на попередньо визначеній тренувальній вибірці, яка виокремлюється з існуючого масиву даних, де певна класифікація вже розроблена. Такі методи мають наступні особливості:

1. Дозволяють виокремити один або більше шарів прихованих нейронних зв’язків, які не є очевидними у «сирому» масиві;
2. Отримана модель демонструє великий ступінь пов’язаності елементів структури.

Застосовуються такі методи машинного навчання зазвичай тоді, коли ми маємо фрагмент даних, який ми хочемо пояснити або передбачити [Awodele, 2009], [Rao, 2007].

Однак є деякі нюанси, на які потрібно перевіряти дані перед тим, як застосовувати керовані методи машинного навчання. До них відносяться наступні [Hodge & Austin, 2004]:

- Значення, що введені випадково під час створення масиву, так званий шум, який «збиває» алгоритм та дещо викривлює отриманий результат. Найкраще - не допускати таких значень у масиві (контролюється під час створення), але якщо вже вони є - необхідно видалити їх, виконавши процедуру «пропущених значень».
- Відсутні значення. Від даної проблеми неможливо застерігтися, оскільки ідеальних даних (як і нічого ідеального) у реальному світі не існує. Єдине, що може допомогти - розуміння самої причини, через яку ці дані відсутні. Так, наприклад, може бути таке, що для зазначеної характеристики такої інформації не існує взагалі або ж значення просто було загублено або не введено на етапі створення масиву. Відповідно до цього, дослідник має можливість з численних варіантів обрати метод, який допоможе замінити відсутні дані.

Тож, визначення потрібного методу насамперед залежить від першочергових особливостей наявного масиву даних. Так, якщо дослідження було пошуковим і основною метою його був якраз-таки пошук певних взаємопов’язаних елементів - то, скоріше, дослідник обере некеровані методи навчання. Якщо ж ми маємо справу з пояснювальним дослідженням - тут доречнішим буде використання контрольованих методів, які допоможуть пояснити та передбачити ту структуру, яка вже відома досліднику, однак у той самий час визначить ще й таку, що непомітна з першого погляду.

Тепер, розуміючи необхідні базові поняття класифікації та орієнтуючись у видах методів машинного навчання, можна безпосередньо переходити до методів аналізу, які дозволяють досліднику здійснювати процес класифікації заданих об’єктів. Перед тим зауважимо, що враховуючи специфіку практичної частини даної роботи, яку буде розглянуто далі, мова йтиме саме про керовані методи.

Розділ 1.4. Керовані методи класифікації

Як вже зазначалося раніше, класифікація є доволі поширеним методом аналізу даних серед дослідників, тож не дивно, що наразі існує чимала кількість алгоритмів на основі штучного інтелекту (логічні та символічні, такі як, наприклад, дерева рішень) та тих, що базуються на статистиці (баєсівські мережі) або сприйнятті (штучні нейронні мережі) для їхнього виконання. Розглянемо деякі з цих методів дещо детальніше.

Метод побудови дерев рішень

Одним з базових методів є метод прийняття дерев рішень, на основі яких робиться класифікація об’єктів. Дерева рішень - це дерева, які класифікують об’єкти, сортуючи їх на основі значень функції (Додаток 1) [Kotsiantis, 2007]. У деревоподібних структурах умовне листя являє собою класифікації (також називаються мітками), проміжні вузли, є ознаками, а гілки являють собою поєднання ознак, які власне і призводять до класифікації [Tan, 2015]. Дерева прийняття рішень є найбільш популярним індуктивним методом у сучасному використанні. Найчастіше такі дерева будуються у два етапи. При зростанні алгоритм знаходить на кожному вузлі (підмножині даних) характерні властивості об’єктів, тим самим розподіляючи їх на класи. Другий етап – розбиття даних на два нові вузли на основі попереднього кроку [Nisbet, et al., 2009].

Загалом, до переваг дерев рішень можна віднести наступне:

- Легка інтерпретація. Ця перевага робить модель легшою для пояснення. Незважаючи на те, що інші алгоритми можуть продукувати більш точну модель в окремій ситуації, дерево рішень може бути підготовлене для здійснення прогнозів стосовно результатів цього методу;

- Можна використовувати як категоріальні, так і числові незалежні змінні;
- Здатність моделювання високого ступеня нелінійності для пояснення взаємозв'язку між залежними і незалежними змінними;
- Висока швидкість об’єднання;
- З деяких реалізацій алгоритмів дерев рішень індуктивним шляхом можна робити висновки стосовно певних правил, починаючи з кінця об’єднаного дерева.

Натомість до недоліків методу належать:

- Схильність до перенаповнення (хоча різні способи скорочення можуть зменшити цю проблему). Перенаповнення – властивість моделі описувати та прогнозувати тестувальні дані гірше, ніж ті, на яких проводилось її навчання (тренувальні дані);
- Мають труднощі з класифікацією тоді, коли вихідними даними є декілька класів [Nisbet, et al., 2018];
- На додаток до двох попередніх пунктів – нестабільність до змін основного масиву.

Хоча дерева рішень є найменш точними з основних методів (за умови окремого використання), вони, як було зазначено, мають багато інших сильних сторін. До таких також можна віднести ефективність навіть з високою часткою відсутніх даних [Kotsiantis, 2007].

Штучні нейронні мережі. Одношарові перцептрони

Це така математична комп’ютерна модель, яка маючи н-ну кількість об’єктів та відповідно для них вагові значення, обчислює суму усіх зважених об’єктів на вході. У результаті, якщо вона виходить вище за визначену порогову - маємо призначення до класу «1». Якщо нижче - «0».

Найпоширеніший спосіб використання такої моделі - почергово запускати на тренувальній вибірці, доки не буде знайдено вектор прогнозування, який задовільнить усі тренувальні набори. Обране методом правило потім використовується для розподілення та прогнозування елементів тестової вибірки. Перевагою перцептронів є те, що завдяки спеціальному алгоритму (експоненціальний алгоритм оновлення), який самостійно підбирає ваги для змінних у відповідності до об’єкта (в залежності від їхньої значущості) - він може швидко адаптуватися до змін масиву. Однак до головного недоліку методу одношарового перцептрона можна віднести те, що він вміє розпізнавати лише лінійний зв’язок.

Так, аби якимось чином вирішити цю проблему, було створено **метод багатошарового перцептрону**. Такі багатошарові перцептрони (або їх ще називають штучними нейронними зв’язками) – це спеціальні обчислювальні системи, можуть навчатися задачам завдяки аналізу заданих прикладів (без попереднього програмування під специфічні задачу) і поступово покращувати свої результати [Yonamoris, 2017]. Вони складаються з численної кількості юнітів (нейронів), які пов’язані один з одним моделями зв’язку. Одиниці в мережі зазвичай поділяються на три класи: одиниці введення, які отримують інформацію, що безпосередньо підлягає обробці; вихідні одиниці, де продемонстровані результати обробки; і приховані одиниці. Номери передачі штучних нейронних мереж дозволяють рухатися сигналам лише по траєкторії вхід-вихід.

Спершу, мережа навчається набору парних даних для визначення відображення вводу-виводу. Потім ваги зв’язків між нейронами фіксуються, а мережа використовується для визначення класифікацій нового набору даних. Під час класифікації сигнал на вхідних блоках поширюється на весь шлях через мережу для визначення значень активації на всіх вихідних одиницях. Кожен вхідний блок має значення активації, яке

представляє деяку особливість зовнішньої мережі. Потім кожен вхідний блок надсилає своє значення активації кожному з прихованих блоків, до яких він підключений. Кожен з цих прихованих блоків обчислює власне значення активації, і цей сигнал передається на вихідні одиниці. Значення активації для кожної одиниці прийому обчислюється відповідно до простої функції активації. Функція підсумовує внески всіх одиниць, що відправляються , які визначаються як вага зв'язку між одиницями, що надсилають і приймають, помножена на значення активації відправника.

Проблемою у даному випадку є правильне визначення розміру прихованого шару, оскільки недооцінка кількості нейронів може призвести до поганого наближення та узагальнення можливостей, тоді як переповнені вузли можуть призвести до надмірного пристосування, а з часом ускладнити пошук глобальної сукупності найсприятливіших умов [Kon & Plaskota, 2000].

Штучні нейронні мережі є переважно залежними від трьох основних аспектів: функцій вводу та активації блоку, будови мережі та ваг кожного вхідного з’єднання. Якщо дві перші залежності - фіксовані, то третя може постійно змінюватися й відповідно змінювати умовну «поведінку» алгоритму. Так, першочергово ваги встановлюються випадковими значеннями, а потім елементи тренувального набору неодноразово піддаються впливу мережі. Значення змінної для введення розміщуються на вхідних одиницях, а вихідна мережа порівнюється з бажаним виходом для цієї змінної. Потім ваги відповідно регулюються або вхідні значення наблизилися до бажаних значень виходу. Оскільки процедура зміни ваг може бути доволі тривалою (через постійних підбір найбільш доречних) - існують спеціальні стоп-правила:

- зупиняється після визначеної кількості повторів процедури підбору;

- зупиняється, коли значення помилки досягає порового значення;
- зупиняється, коли протягом декількох останніх повторів не було виявлено жодних покращень;
- зупиняється, коли рівень похибки на тренувальній вибірці перевищує встановлений рівень похибки у тестувальній.

Мережі радіально-базисних функцій

Мережі радіальної основи також широко застосовуються у багатьох наукових та технічних областях для розробки класифікацій. Це тришарова мережа зворотного зв’язку, в якій кожен прихований блок реалізує функцію радіальної активації (значення якої залежить від відстані до початку координат), і кожен блок виведення - зважену суму прихованих одиниць виводу. Процедура його навчання зазвичай поділяється на два етапи:

1. Спочатку центр та ширин прихованого шару визначаються алгоритмами кластеризації.
2. Потім, ваги, що з'єднують прихований шар із вихідним шаром, визначаються методом сингулярного значення декомпозиції або найменшим середнім квадратом.

Проблема вибору відповідної кількості базових функцій залишається основною проблемою для мереж РБФ. Кількість базових функцій контролює загальну складність та узагальнюючу здатність даного алгоритму. Мережі РБФ з надто малою базовою функцією не можуть адекватно відповідати тренувальним даним через обмежену гнучкість. З іншого боку, ті, що мають занадто багато базових функцій, характеризуються слабкою здібністю до узагальнення, оскільки вони занадто гнучкі та вразливі, а відповідно, помилково визначають результати тестової підмножини даних [Kotsiantis, 2007].

Статистичні методи. Баєсівські мережі

На відміну від штучних нейронних мереж, методи, що засновані на статистиці, характеризуються наявністю певної таблиці ймовірностей. Тобто, у результаті вони не просто відносять об'єкт до певного класу, а визначають ймовірність приналежності до нього. Найбільш відомим серед статистичних методів класифікації є **метод баєсівських мереж** [Kotsiantis, 2007].

Байєсівська мережа (БМ) - це графічна модель відношення ймовірностей між набором змінних (Додаток 2). Структура мережі Баєса - це ациклічний граф (де відсутні орієнтовані шляхи, що починаються і закінчуються в одній і тій самій вершині). Вузли, що є незалежними один від одного, в такій структурі находяться в одному співвідношенні з особливостями змінних. Зазвичай завдання вивчення баєсівської мережі можна розділити на дві підзадачі:

- Спочатку вивчення структури мережі;
- Потім визначення її параметрів.

Імовірнісні параметри кодуються у набір таблиць, по одній для кожної змінної, у вигляді локальних умовних розподілів змінної. З огляду на незалежність, закодовану в мережу, спільний розподіл для об'єкта можна реконструювати шляхом простого множення цих таблиць. У загальних рамках індукції баєсівських мереж існує два сценарії: структура відома або ж не відома.

У першому сценарії структура мережі задається дослідником і у результаті може бути скорегованою. Як тільки чітко визначено мережеву структуру, вивчення параметрів у таблицях умовних ймовірностей зазвичай вирішується шляхом оцінки локальної експоненціальної кількості параметрів із наданих даних [Jensen, 1996]. Кожен вузол у мережі має

пов'язану таблицю ймовірностей, яка описує умовний розподіл ймовірності цього вузла з урахуванням різних значень попередніх вузлів.

Незважаючи на неабияку поширеність баєсівських мереж, для них можливе певне обмеження. Це обчислювальна складність дослідження невідомої раніше мережі. Враховуючи проблему, що описується певною кількістю ознак, кількість можливих структурних гіпотез більше ніж кількість ознак. Якщо структура невідома, одним із підходів є введення функції балів, який оцінює «придатність» на тренувальних даних, а потім шукає найкращу мережу відповідно до цього результату.

До переваг методу, порівняно із деревам рішень чи штучними нейронними мережами відноситься можливість врахування структурних особливостей взаємозв’язку об’єктів на основі попередньої інформації про них. Таким чином, ще на початку аналізу можна зрозуміти, чи є даний вузол початковим, чи має він продовження, чи є цей вузол прямим наслідком попереднього вузла або ж він взагалі немає жодного відношення до нього. Також можна забезпечити й порядок вузлів [Kotsiantis, 2007]. Основний недолік даного методу – неможливість застосування на масивах з багатьма ознаками (оскільки побудова великої мережі неможлива з точки зору часу та простору). Інший недолік полягає у тому, що кількісні дані повинні бути дискретними аби мати змогу побудувати баєсівську мережу.

Ще один метод, який можна віднести до статистичних – **алгоритм навчання на прикладах** [Kotsiantis, 2007]. Особливістю даних алгоритмів є те, що вони потребують менше часу на навчання (порівняно із описаними вище), але трохи повільніше роблять класифікацію. Один з найбільш відомих з таких методів – метод найближчих сусідів. Він заснований на тому принципі, що об’єкти в наборах даних існують на безпосередній близькості від інших об’єктів, що мають схожі властивості (ознаки).

Загалом, такі «приклади» можуть розглядатися як точки у просторі, де кожна з них відповідає одній ознаці об’єкта. Абсолютна позиція об’єкта у цьому просторі не така значна, як відносна відстань між ними Ця відносна відстань визначається за допомогою метрики відстані. В ідеалі вона повинна мінімізувати відстань між двома аналогічно класифікованими об’єктами, максимізуючи відстань між об’єктами, що належать до різних класів.

Висновки до Розділу 1

Розробка класифікацій, як і було продемонстровано у даному розділі – таки доволі популярний запит для дослідників, оскільки в дійсності такі класифікації допомагають не лише поверхово розумітися на структурі даних, але й бути в курсі особливостей об’єктів, які складають цю структуру. Було зазначено, що наразі у науці поширеними є два підходи до здійснення класифікацій, які здебільшого залежать від необхідного рівня узагальнення, а також від вже відомої інформації про об’єкт (есенціалізм та емпіризм, неконтрольовані та контрольовані методи машинного навчання, тощо).

Також бачимо, що дійсно дослідники не обмежені у своєму виборі методів, за допомогою яких вони можуть класифікувати об’єкти. В кожного з описаних методів є свої переваги та недоліки, в залежності від яких і відбувається обрання того чи іншого алгоритму.

Однак машинне навчання та аналіз даних – сфери, які розвиваються з шаленою швидкістю у наш час, тож на сьогодні існують вже і альтернативні методи, які поєднують у собі й характеристики попередньо описаних (ансамблеві методи). Одним з таких і є, власне, предметом дослідження даної роботи і до його безпосереднього розгляду ми перейдемо у наступному розділі.

Розділ 2. Ансамблеві методи. Метод випадкового лісу

Перш ніж переходити до методу випадкового лісу, зупинимось трохи на особливостях тих алгоритмів, на яких він побудований.

Розділ 2.1. Ансамблеві методи

Так, метод випадкового лісу належить до ансамблевих алгоритмів, які поєднують у особі декілька принципів алгоритмів класифікацій для підвищення загальної точності та ефективності. Основні причини помилок у моделях навчання, як зазначалося у першому розділі, обумовлені шумом, упередженістю та розбіжністю (невисокою точністю).

Ансамблеві методи допомагають мінімізувати ці фактори. Ці методи розроблені для підвищення стабільності та точності алгоритмів машинного навчання. Під час своєї роботи вони використовують групу моделей, комбінований результат з яких майже завжди кращий (з точки зору точності прогнозування) порівняно з використанням якогось одного алгоритму.

Навчання ансамблів складається з двох етапів. На першому кроці тренуються базові моделі, що складають ансамбль. На другому кроці з'ясовується, як поєднуються ці моделі (або їх прогнози) в єдину цілісну модель (або передбачення).

Під час застосування ансамблевого підходу, має сенс намагатися використовувати якнайбільше варіантів поєднання, а потім обирати вже найкращий варіант із запропонованих, оскільки якщо вже дослідник й вдається до використання ансамблю – то очевидно-зрозумілим є той факт, що він хоче покращити точність прогнозування. А зробити можна це, як щойно було зазначено, можна завдяки порівнянню багатьох наявних рішень.

Так, у ансамблевих методів є два можливих варіанти розвитку подій: у першому випадку, різноманітність моделей досягається модифікацією навчальних даних, тоді як у другому випадку різноманітні моделі засвоюються шляхом зміни алгоритму навчання.

Більшість ансамблевих досліджень було зосереджено на методах першої групи, тобто методах, які використовують різні набори навчальних даних. Такі набори даних можуть бути отримані за допомогою методів перекомпонування, таких як бутстреп [DŽEROSKI, 2009], коли навчальні набори складаються випадковим чином із заміною початкового навчального набору. Саме такий підхід застосовується застосовують у беггінгу [Breiman, 1996] та випадкових лісах [Breiman, 2001]. Власне останній метод, натомість, основним чином є поєднанням беггінгу та дерев прийняття рішень.

До переваг ансамблевих методів можна віднести наступні положення [Juhi, 2019]:

1. Більш точні результати прогнозування. Ми можемо порівняти роботу методів ансамблю з диверсифікацією: ансамбль моделей дасть кращу ефективність за тестовими сценаріями (небачені дані) порівняно з окремими моделями у більшості випадків;
2. Стабільніша модель. Сукупний результат декількох моделей завжди менш вразливий до похибок та шумів, ніж окремі моделі. Це призводить до стабільності та надійності моделі.

Недоліки ансамблевих методів можна виокремити такі [Juhi, 2019]:

1. Зменшення можливостей інтерпретації моделей. Використання методів ансамблю знижує здатність модельної інтерпретації через підвищену складність;
2. Час на обчислення та розробку високий;
3. Важко визначитися з алгоритмами, якими наповнювати ансамбль.

Розділ 2.2. Алгоритм бутстреп-агрегування

Бутстреп-агрегування або беггінг - це метод машинного навчання, призначений для поліпшення стабільності і точності алгоритмів машинного навчання, які використовуються в статистичній класифікації і регресії.

Алгоритм зменшує дисперсію і допомагає уникнути переповнення [Strobl, 2009].

Особливістю його є те, що для свого тренування він щоразу випадковим чином підбирає нові об’єкти з навчальної підмножини, шо дає можливість методу краще навчитись (оскільки охоплює більшу частину даних) та в подальшому, відповідно, призводить до більш точних результатів [Reinstein, 2017]. Для утворення підвибірок використовує процедуру повторного відбору об’єктів, тобто у різних підвибірках можуть знаходитись однакові об’єкти. Алгоритм тим самим зменшує дисперсію і допомагає уникнути перенаповнення.

За допомогою цієї вибіркової навчальної підмножини «навчається» колекція базових моделей. Потім їх прогнози поєднуються простим «голосуванням більшості», що також допомагає зменшити дисперсію. (Голосування - це метод поєднання базових моделей, який не стосується їхньої генерації, на відміну від інших методів). Воно поєднує прогнози базових моделей за статичною схемою голосування, яка не залежить від навчальних чи від базових моделей. Найпростіший тип голосування - це множинне голосування (його також називають голосуванням більшості), де кожна базова модель надає голос за своє передбачення. Прогноз, який набере більшість голосів - це остаточний результат ансамблю. Якщо ми прогнозуємо числове значення, передбачувальний прогноз - це середнє значення прогнозів базових моделей) [DŽEROSKI, 2009].

Такий ансамбль часто дає кращі результати, ніж його окремі базові моделі, оскільки поєднує в собі переваги окремих моделей. Беггінг слід використовувати разом з нестабільними алгоритмами навчання (наприклад, з такими як дерева прийняття рішень або нейронними мережами), де невеликі зміни в навчальному наборі призводять до значно різних класифікаторів [Brownlee, 2019].

Розділ 2.3. Метод випадкового лісу. Переваги та недоліки

Метод випадкового лісу (Random Forest) – гнучкий і простий у застосуванні алгоритм машинного навчання, який у більшості випадків призводить до хороших результатів навіть без налаштування багатьох складних параметрів, оскільки має лише два: кількість побудованих дерев та кількість ознак, що використовуються для створення вузлів. Більш того, він також належить до найбільш уживаних алгоритмів, оскільки є універсальним і може вирішувати задачі як класифікації, так і кластеризації чи регресії, а відповідно може бути застосованим як на метричних ознаках, так і на номінальних чи порядкових [Breiman, 2001].

Як вже зазначалося, метод належить до ансамблевих підходів, основний принцип якого полягає у поєднанні декількох наборів дерев прийняття рішень для кращого виконання прогностичної та класифікаційної можливості. Працює метод також на основі беггінгу, характеристика якого була описана у попередньому підрозділі.

Відмінність між ними полягає у тому, що випадковий ліс вводить більше рандомізації і різноманітності (а саме змінні випадково обираються на кожному кроці створення вузла дерева), застосовуючи беггінг до простору ознак. Тобто, замість того, щоб шукати найкращі незалежні змінні, що будуть використовуватися для створення гілок, він випадковим чином відбирає елементи з простору цих незалежних змінних [Brownlee, 2019]. Інакше кажучи, метод випадкового лісу може ще більше покращити показники прогнозування порівняно з бутстреп агрегуванням, тому що окремі дерева, які беруть участь в усередненні є ще різноманітнішими (через участь ще більшої кількості ознак у їх створенні) [Strobl, 2009].

Відмінність від методу дерев прийняття рішень наступна [Donges, 2018]:

- Якщо штучно вводити навчальний набір з певними характеристиками та мітками в дерево прийнятих рішень, буде сформовано певний набір правил, які будуть використовуватися для подальшого

прогнозування. Алгоритм випадкового лісу ж випадковим чином вибирає характеристики для побудови декількох дерев, а потім усереднює результати.

- Якщо даних забагато, а, відповідно, багато і побудованих дерев – вони можуть бути перенаповненими. Метод випадкового лісу натомість не дозволяє цьому трапитись в більшості випадках, створюючи випадкові підмножини функції і будуючи невеликі дерева, використовуючи ці підмножини. З таких піддерев у результаті утворюються поєднані дерева. Однак, це не спрацьовує кожного разу, а також це впливає на час обчислення: чим більша кількість дерев – тим більше часу знадобиться.

Побудова дерев методом випадкового лісу

Для побудови дерев необхідно розподілити масив на навчальні (тренувальні) набори, на яких алгоритм буде навчатися – N, а також маючи уявлення про зв'язок між змінними – виокремити з них ті, які є цікавими безпосередньо для нашого аналізу – m. Далі метод працює за наступним алгоритмом: [Reinstein, 2017], [Benyamin, 2012]:

1. З усіх можливих N, n-кількість разів із повторенням обираються навчальні набори (підмножини даних). Підмножина не повинна становити більше 70% від загальної кількості навчальних вибірок;
2. Для кожного вузла дерева випадковим чином обирається кількість ознак (m), яка повинна бути значно меншою за загальне M. На них базується рішення цього вузла. Ознака, що забезпечує найкраще розщеплення відповідно до мети, використовується для створення розбиття на цьому вузлі. На наступному вузлі обираються інші змінні (m) навмання з усіх M. Така процедура повторюється до отримання кінцевого результату.

3. Кожне дерево поступово «зростає», але не обрізається (на відміну від того, як це відбувається у класичному методі дерев прийняття рішень).
4. Оцінюється точність моделі та відсоток даних, які не брали участі у створенні тренувального масиву для навчання методу. Зазначені спостереження називаються такими, що знаходяться поза вибіркою (OOB, out-of-bag). Спостереження OOB можуть бути використані, наприклад, для оцінки похибки прогнозування за допомогою методу випадкового лісу [Janitza, 2018].

Коли в систему вводяться нові ознаки, пошук виконується знову по всіх деревах [Benyamin, 2012].

Також важливим та невід'ємним етапом у застосуванні методу випадкового лісу є визначення значущості незалежної змінної. Так, окремі класифікаційні дерева інтерпретуються доволі легко (навіть інтуїтивно): як на перший погляд, так і описово, при детальному огляді структури дерева. На відміну від цього, дерева, що побудовані ансамблевими методами (зокрема випадковим лісом) взагалі непросто інтерпретувати, оскільки окремі дерева в них жодним чином не зафіксовані: кожна змінна може з'являтися в різних місцях в дереві (або ж взагалі в різних деревах). Тим не менш така особливість ансамблевого підходу та безпосередньо випадкового лісу надає їм перевагу в тому, що таким чином, вони можуть краще відображати її потенційно складний вплив на залежну змінну.

В принципі, найпростішим способом оцінки важливості змінної – був би шлях підрахунку тієї кількості разів, коли кожна змінна буде обрана усіма окремими деревами в ансамблі. Дещо важчий принцип визначення значимості змінної – зважене значення покращення окремих дерев під час розщеплення, виробленого кожною змінною [Strobl, 2009]. Прикладом такої міри є показник «Середнє зниження Джині». Це середнє значення

загального зменшення «забруднення» вузла, зважене на частку тих змінних, що досягають цього вузла, у кожному окремому дереві рішення у випадковому лісі. Чим більше його значення – тим більший вплив він вносить до розробки класифікації.

Переваги та недоліки методу випадкового лісу

До переваг методу випадкового лісу відносять наступне [Kumar, 2019]:

1. Випадковий ліс ґрунтується на алгоритмі беггінгу та використовує техніку ансамблевого навчання. Він створює стільки ж дерев на підмножині даних і поєднує вихід усіх дерев шляхом усереднення або більшості голосів. Таким чином, це зменшує проблему «перенасичення» в деревах рішень, а також зменшує дисперсію. Таким чином підвищується точність. Ще один можливий варіант попередити перенаповнення – дотримання наступної умови:

$$e^m < O,$$

де m – кількість ознак, які виокремлені як незалежні; o – кількість об’єктів у вибірці (Pires de Oliveira, 2017).

2. Випадковий ліс може використовуватися для вирішення як класифікаційних, так і проблем регресії.
3. Випадковий ліс добре працює як з категоричними, так і з безперервними змінними.
4. Випадковий ліс може автоматично обробляти відсутні значення.
5. Випадковий ліс може автоматично обробляти пропущені значення.
6. Не потрібно робити стандартизацію та нормалізацію розподілу ознак у випадку використання методу випадкового лісу, оскільки він використовує підхід на основі правил замість обчислення відстані.

7. Ефективно обробляє нелінійні параметри: нелінійні параметри не впливають на продуктивність випадкового лісу на відміну від алгоритмів на основі кривих.
8. Випадковий ліс, як правило, є стійким до викидів і може впоратися з ними автоматично.
9. Алгоритм випадкових лісів дуже стійкий. Навіть якщо в набір даних введено нову змінну, загальний алгоритм не піддасться сильній зміні, оскільки нові дані можуть впливати на одне дерево, але не можуть зачепити усі дерева.
10. Внесок аналітика (дослідника) мінімальний, оскільки для використання методу використовується лише два параметри, які непогано налаштовані автоматично (але за потребою їх можна змінювати).

Основне обмеження методу полягає в тому, що велика кількість дерев може зробити алгоритм повільним і неефективним для прогнозів у реальному часі. Загалом, ці алгоритми швидкі для навчальних вибірок, але досить повільні для створення прогнозів після того, як вони пройшли це навчання. Більш точний прогноз вимагає зростання кількості дерев, що призводить до більш повільної моделі. Тим не менш, у багатьох реальних додатках алгоритм випадкового лісу є досить швидким. Але, безумовно, можуть виникнути і ситуації, коли продуктивність під час виконання важлива, а інші підходи будуть кращими [Eulogio, 2017].

З цього виплаває і другий недолік, який полягає у тому, що методу випадкового лісу потрібно більше часу на навчання, оскільки він генерує доволі багато дерев рішень, а потім ще й проводить процедуру обрання більшістю голосів [Kumar, 2019].

Висновки до Розділу 2

З цього розділу можемо побачити, що на сьогодні більш застосованими є ансамблеві підходи, які завдяки використанню декількох алгоритмів одночасно, покращують загальну продуктивність та результат моделі. До таких методів належить і метод випадкового лісу, який у свою чергу є ансамблем беггінгу (бутстреп агрегування) та дерев прийняття рішень.

Завдяки введенню рандомізації та можливості для ознак повторюватися у різних тренувальних підмножинах, метод дозволяє отримати вищу точність передбачення і, відповідно, є більш ефективним, ніж використання окремих методів для класифікації даних. Він доволі швидко дозволяє визначити міру подібності між двома ознаками, підраховуючи кількість разів їхнього розміщення в одному і тому ж вузлі дерев. Більш того, оскільки метод має усього два параметри він не є важким у використанні, а завдяки й сучасному програмному забезпеченню – взагалі потрібно не потребує втручання аналітика/дослідника у свою роботу.

Розділ 3. Класифікація вступників до закладів вищої освіти за допомогою практичного використання методу випадкового лісу

Актуальність. Вибір майбутньої спеціальності та університету – один з найважливіших кроків, з яким стикаються учні після завершення навчання у школі та складання іспитів. Аби дещо полегшити його, наразі існує чимала кількість спеціальних організацій та заходів, які інформують школярів про можливі варіанти для вступу, різноманітні професії та сфери діяльності. Нерідко таким «піаром» займаються й безпосередньо самі факультети (або ж університети) за власною ініціативою.

Так, не винятком є і наш факультет соціології Київського національного університету імені Тараса Шевченка. Протягом останніх двох років студентськими організаціями, із залученням допомоги викладачів було докладено доволі багато зусиль, аби охопити якомога більше абітурієнтів та зацікавити їх у вступі до факультету. Однак під час розробки інформаційних стратегій виникали певні труднощі. Одна з таких – відсутність чіткого уявлення про те, як саме абітурієнти обирають спеціальність/факультет (зокрема, соціологію) аби мати можливість вказати на ті основні моменти, які абітурієнти враховують під час подання вступних заяв до вищих навчальних закладів. У зв’язку з цим, дослідження таких заяв вступників та їхня подальша класифікація – є доволі актуальним запитом, який допоможе краще дізнатися нам про «цільову аудиторію», особливості її вибору та преференції. У результаті, така інформація допоможе виявити найбільш значимі для вступників характеристики, на яких потрібно робити наголос під час реклами факультету, і, у ідеальному випадку, надасть нам змогу залучити дедалі більше абітурієнтів та у майбутньому поповнити нішу соціологічної науки в України висококваліфікованими фахівцями, які будуть розвивати не тільки її, а й допомагати українському суспільству ставати кращим.

Теоретичний об'єкт дослідження: стратегії вибору спеціальностей «Соціологія» та «Політологія» вступниками на основі аналізу їхніх заяв до закладів вищої освіти.

Емпіричний об'єкт дослідження: вступники до закладів вищої освіти України, серед пріоритетів яких були вказані спеціальності «Соціологія» або «Політологія».

Мета: виявлення особливостей вибору вступниками до спеціальностей («Соціологія» та «Політологія») під час подання заяв до закладів вищої освіти. Аби дещо полегшити інтерпретацію, на основі аналізу заяв буде розроблено їхню класифікацію. Для здійснення мети дослідження, виокремлено наступні **завдання:**

1. За допомогою методу випадкового лісу виокремити змінні, які найбільш наочно відображають вибір вступників;
2. Виокремити особливості, які характерні для вибору вступниками на спеціальності («Соціологія» чи «Політологія»);
3. Розробити класифікацію вступників до закладів вищої освіти України, що обрали спеціальність «Соціологія» та «Політологія».

Гіпотези:

1. Вибір спеціалізацій «Соціологія» та «Політологія» здійснюється за однаковим принципом, оскільки їхні предметні сторони дослідження дещо перекликаються між собою.
2. Для обох спеціальностей («Соціологія» та «Політологія») безпосередньо сам університет – буде основною (найбільш значущою) змінною, під час здійснення вибору.
3. Середній бал атестату та зовнішнього незалежного оцінювання не є важливими під час обрання спеціальностей «Соціологія» та «Політологія».

Метод аналізу: випадковий ліс.

Аналіз здійснюватиметься за допомогою програмного забезпечення RStudio із використанням додаткових бібліотек: «readxl» для відкриття масиву, «randomForest», «cowplot», «caret» для безпосереднього застосування методу випадкового лісу та «ggplot2» для часткової візуалізації результатів.

Масив: 6 532 заяви вступників до закладів вищої освіти України за спеціальностями: соціологія та політологія.

- Змінні (ознаки) масиву, за якими здійснюється аналіз:
- Пріоритет, який було вказано абітурієнтами – незалежна;
- Середній бал атестату – незалежна;
- Середній бал зовнішнього незалежного оцінювання – незалежна;
- Університет – незалежна;
- Спеціальність – залежна.

Перш ніж переходити до безпосереднього аналізу даних, необхідно відокремити об’єктів нашого дослідження із загальної кількості вступних заяв (яка дещо перевищувала 500 тисяч), а також створити нову (перекодовану) змінну «університет», оскільки просто назва самого закладу вищої освіти (першочерговий вигляд зазначеної змінної) не додає нам можливостей для кращої інтерпретації результатів. Тож, було прийнято рішення перекодувати її у відповідності до рейтингу університетів України, який було презентовано центром міжнародних проектів «Євроосвіта» у співпраці з міжнародною групою експертів IREG Observatory on Academic Ranking and Excellence [Рейтинг закладів вищої освіти України, 2019].

Для отримання змістовних результатів використовуємо наступний алгоритм:

1. data <- read_excel ("vstup2019.xlsx")
   data1<- data.frame (data)

На даному етапі ми імпортуємо наш масив у середовище RStudio, робимо з даних фрейм аби створити таку структуру, яка могла б із легкістю зчитуватися під час наступних дій.

2. summary(data1)

   data2 <- data1

   data2$spec <- ifelse(data2$speciality == 'Соціологія', 'Так', 'Ні')

   data2$spec <- as.factor(data2$spec)

   data2$speciality <- NULL

Дивимось на завантажений масив, перевіряємо чи все правильно імпортувалося та дізнаємося типи шкал наших даних. Далі створюємо копію нашого масиву (з якою надалі і будемо працювати) та змінну, яка містить значення приналежності до тих, хто обрав спеціальність «Соціологія» – «так» або «ні». Перетворюємо її у фактор та змінну, на основі якої ми створювали нову – видаляємо (оскільки це необхідно для здійснення прогнозу у майбутньому).

3. chart.Correlation(data2[-5], col=data2$spec)

Перевіряємо кореляцію між незалежними змінними. У нашому випадку – бачимо високий зв’язок між середнім балом ЗНО та середнім балом атестату (див. Додаток 3). Оскільки вимога до незалежних змінних – низька кореляція одне з одним (оскільки висока – означає мультиколінеарність, яка у свою чергу характеризує те, що змінні вимірюють одне і те ж саме, і у результаті знижує працездатність та якість моделі). – виключаємо з подальшого аналізу змінну із середнім балом атестату. Після виключення змінної маємо прийнятні коефіцієнти кореляції (див. Додаток 4), що дозволяє продовжити роботу із запропонованою формулою факторів (залишилися – середній бал зовнішнього незалежного оцінювання, пріоритет та рейтинг) і залежної змінної (спеціальність «Соціологія»).

4. model <- train (spec~., data2,

   method = 'rf', TuneLength=5,

trControl=trainControl(

method = 'cv', number = 12,

classProbs = TRUE))

model$results

Створюємо навчальний набір даних, вказуючи залежну та незалежні змінні (у даному випадку – незалежні змінні – всі, окрім залежної, оскільки саме там масив був попередньо відредагований, зокрема аби зменшити кількість часу, що необхідна для його опрацювання). Параметр tuneLength вказує алгоритму спробувати різні значення за замовчуванням для залежної змінної. Метод «CV» означає, що наша навчальна підвибірка буде випадковим чином розподілена на 12 частин, кожна з яких потім буде використовуватися задля навчання моделі. Далі спостерігаємо результати побудованої моделі та оцінюємо точність. У результатах побудованої моделі бачимо, що точність її складає 75% (що є доволі високим значенням для реальних емпіричних розподілів), а кількість змінних характеристик об'єктів), які обиралися випадковим чином (звідки й походить назва методу) для її побудови – 2 (див. Додаток 5). Якщо використовується 3 змінних – точність лише на 1% нижча. Така точність може означати й те, що в масиві існує й певна структура даних, яку було виявлено методом (і власне на основі якої може бути здійснене прогнозування поданих абітурієнтами заяв у поточному чи наступному році). Тож, аби виконати завдання даного дослідження й досягти мети, розберемо дещо детальніше цю структуру.

5. predicteddata <- predict(model, data, "prob")

   head(predicteddata)

На цьому етапі будуємо прогноз для перевірки реальної практичної значущості моделі. Виводимо передбачувані результати для перших шістьох об'єктів нашого масиву. Перевіряємо із тим, що вже маємо на оригінальному масиві (саме для цього на першому етапі створювали його

копію) – моделлю було правильно передбачено шість значень із шести (див. Додаток 6).

6. data$specc <- ifelse (predicteddata$Hi > predicteddata$Так, 0, 1)
   test <- subset (data, specc == 1)
   test2 <- subset (test, speciality == "Соціологія")

Ще додатковий спосіб перевірити на скільки точно працює модель: створюємо нову змінну в оригінальному масиві на основі спрогнозованих результатів із значеннями 0 (якщо «Ні») та 1 (якщо «Так»). Далі у новий масив з усіма змінними виводимо лише тих, в кого спрогнозована змінна дорівнює значенню «Так». Дивимось кількість об’єктів (у нашому випадку – 2 164). Далі у ще один новий тестувальний масив з попередньо створеного відбираємо тих, в кого реально обрана спеціальність «Соціологія». Порівнюємо кількість об’єктів у створених масивах: 2 164 на основі передбачених даних та 2 085 на реальних даних. Бачимо, що результат спрогнозований не співпадає з реальним лише на 79 об’єктів, що дорівнює 96% точності, але також потрібно враховувати й ті об’єкти, значення яких неправильно було передбачено моделлю і які просто не потрапили у дане тестування.

Для цього побудуємо частотну таблицю із змінною з обраними спеціальностям (див. Додаток 7) та побачимо, що з загальної кількості об’єктів (2 318) було правильно спрогнозовано 2 085, що дорівнює 89.9% правильно спрогнозованих значень, що навіть більше, ніж зазначалося самою моделлю початково.

Після побудови загальної моделі та здійснення прогнозів, необхідно перевірити власне яка ж змінна найбільше вносить значущості у здійсненню класифікації (або ж у нашому випадку – який параметр найбільше пов’язаний із вибором абітурієнтом спеціальності). Забезпечується це завдяки наступним крокам:

1. output.forest <- randomForest(spec ~ priority + meanZNO + rating, data = data2)
   print(output.forest)
   plot(output.forest)

Створюємо нову змінну, де за допомогою введення формули визначаємо нашу залежну та незалежні змінні. Далі виводимо наші результати, які у конкретно нашому випадку виявилися наступними:

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 1

OOB estimate of error rate: 24.2% (згадуємо, що точність моделі була оцінена методом у 76%, що відповідає зазначеному рівню можливої помилки). А також спостерігаємо за тим, як при збільшенні кількості побудованих дерев (від 0 до 50) зменшується рівень похибки (див. Додаток 8).

2. randomForest::importance(output.forest)

Безпосередньо дізнаємося значення впливу кожної змінної: найбільший вплив у здійснення класифікації (а відповідно й вибору) вносить рейтинг університету, далі йде середній бал за зовнішнє незалежне оцінювання. Вказаний пріоритет не є таким важливим у порівнянні із двома попередніми змінними, але все одно робить свій внесок, оскільки значення його важливості більше за 5 (див. Додаток 9). На цьому побудова класифікаційної моделі методом випадкового лісу завершена.

За аналогічною послідовністю робимо все те саме й для обраної спеціалізації «Політологія». Точність моделі майже така сама (лише на кілька сотих вище). З 4 214 передбачено моделлю було 4 126, що складає 98%. На етапі перевірки важливості змінних бачимо відмінність: на відміну від спеціальності «Соціологія», на обрання «Політології» мають практично рівнозначний вплив як рейтинг закладу вищої освіти, так і середній бал,

який абітурієнт отримав за складання зовнішнього незалежного оцінювання (див. Додаток 10).

Після отриманих результатів аналізу, спираючись на теоретичні засади, які були описані у першому розділі та базуючись на емпіричному підході, вступників до закладів вищої освіти на спеціальність «Соціологія» та «Політологія» можна класифікувати наступним чином:

1. Вступники, що обирають спеціальність відштовхуючись рівномірно від результатів зовнішнього незалежного оцінювання (і середнього балу атестату, оскільки кореляція між цими змінними висока, пряма) та від університету (що характерне для абітурієнтів, які обрали спеціальність «Політологія»).
2. Вступники, які обирають заклад вищої освіти (така модель вибору була притаманна тим студентам, які подавали свої заяви до спеціальності «Соціологія»). Визначальною (найбільш значущою (у два рази) змінною у такому випадку був університет.

Для схематичної візуалізації отриманого результату особливостей стратегії вибору вступниками було також проаналізовано також статистичні описові розподіли тих змінних, що були обрані як найбільш важливі (за основу взято медіанне значення (аби нівелювати вплив викидів), мінімальне та максимальне значення – для змінної середнього балу ЗНО (тільки для тих, хто подавав заяву на політологію), а також частотний розподіл змінної рейтингу університетів для обох спеціальностей. Також для зручності метричну шкалу перетворюємо в інтервальну.

У результаті маємо наступні схеми (див. Додаток 11 та Додаток 12) для вступників, що обирають соціологію та тих, які подавали заяви до політології.

Тож, як бачимо, гіпотеза №1 не підтвердилася, оскільки вибір абітурієнтами спеціальностей відрізняється.

Гіпотеза №2 підтвердилася, особливо для тих, хто обирає соціологію (загалом, обирається університет). Для політології змінна з рейтингом університетів увійшла у топ і значення її важливості було дещо вищим за середній бал зовнішнього незалежного оцінювання.

Гіпотеза №3 головним чином підтвердилась для тих, хто обирає спеціальність «Політологія». Для тих, хто обирає ж соціологію, середній бал не був настільки значущим (порівняно із рейтингом університету).

Висновки до Розділу 3

У підсумку можна сказати, що метод випадкового лісу є доволі зручним рішенням, коли перед дослідником постає завдання класифікації об'єктів та визначення вагового значення змінних, які безпосередньо здійснюють свій внесок у залежну змінну. Здебільшого однією з його переваг та особливостей є те, що він не потребує значного втручання у процес свого втілення, оскільки усі необхідні параметри шляхом навчання методу на тренувальному наборі даних підбираються автоматично.

З його допомогою нам вдалося визначати певні особливості вибору вступниками до закладів вищої освіти спеціальності. На прикладі аналізу масиву тих абітурієнтів, що подавали заяви на спеціальності «Соціологія» та «Політологія» за 2019 рік було покроково продемонстровано практичне застосування методу. У результаті завдяки використанню вбудованих бібліотек та функціональних можливостей середовища RStudio були визначені найбільш вагомі змінні, на основі яких і в кінцевому підсумку було розроблено класифікацію вступників до ЗВО та схематично продемонстровано стратегії їхнього вибору спеціальності.

Так, ті хто обирають «Соціологію» - здебільшого звертають увагу лише на рейтинг університету. Натомість для вступників до політології характерною особливістю є рівнозначний внесок як середнього балу зовнішнього незалежного оцінювання (та атестату), так і рейтингу університету.

На мою думку, пов'язано це з тим, що соціологія досі не є доволі поширеною наукою серед населення, відповідно доволі мало інформації про дану спеціальність у абітурієнтів. А зважаючи на те, що найважливішою змінною виявився рейтинг університету – можна припустити, що «Соціологія» стає ніби як запасним варіантом, аби все ж таки потрапити у бажаний заклад вищої освіти.

Висновки

Отже, у даній роботі було обґрунтовано необхідність розробки класифікацій, визначені теоретичні підходи, на основі яких вони розробляються під час емпіричних досліджень. Також було описано методи, які можуть бути застосованими та виокремлено особливості методу випадкового лісу і здійснено його практичне використання для розробки класифікації вступників до закладів вищої освіти.

У результаті цього було досягнуто мети роботи: розкрито можливості та продемонстровано потенціал методу випадкового лісу для завдань класифікації об’єктів дослідження (власне, на прикладі класифікації вступників до ЗВО за допомогою аналізу їхніх вступних заяв) та виконано поставлені завдання:

1. Розкриття змісту поняття класифікації як методу аналізу даних.

Так, у роботі було проаналізовано два підходи до визначення класифікацій, основними критеріями для виокремлення яких слугує попереднє знання дослідником про структуру об’єктів; мета класифікації; дані, з яких вона розробляється (теоретичні чи емпіричні); кількість ознак, за якими будуть обиратися схожі об’єкти. Такими підходами є есенціалізм та емпіризм. Відповідно до аналізу теоретичних джерел, найбільш доречним визначенням поняття класифікації (безпосередньо для цієї роботи) було обрано наступне: класифікація – внутрішньо системний розподіл об’єктів чи предметів, що вивчаються, за істотними ознаками для зручності їхньої подальшої інтерпретації; певне впорядкування початкових понять, що вказує на ступінь їхньої схожості. Продемонстровано декілька можливих керованих алгоритмів машинного навчання, які використовуються для розробки класифікацій – метод дерев прийняття рішень, штучні нейронні мережі (що базується на перцепторах), Баєсівські мережі та методи, які використовують відстань між об’єктами для знаходження подібності в них.

2. Виокремлення особливостей, переваг та недоліків методу випадкового лісу.

До особливостей відноситься те, що зазначений метод відноситься до ансамблевих методів та заснований на принципах алгоритму дерев прийняття рішень та беггінгу, відрізняється від усіх поширених «класичних» методів підвищеною точністю (що і було продемонстровано, коли 96% та 98% даних було передбачено правильно): оскільки для побудови моделі використовує достатньо велику кількість дерев, сконструйованих на основі обраних випадковим чином значень змінних, що постійно з кожним створенням нового дерева повторюються. Окрім цього, він стійкий до використання шкал різного рангу вимірювання: поєднання категоріальних та числових змінних не призводить до зміщення результатів. До недоліку відносяться порівняно високі витрати часу на те, щоб спочатку алгоритм «навчився», а потім розробив класифікацію.

3. Практичне застосування методу випадкового лісу для здійснення класифікації об’єктів.

На прикладі аналізу вступник заяв абітурієнтів 2019 року, було продемонстровано практичне застосування методу, що довело доцільність його використання, оскільки було отримано високу точність та результативність. У результаті було розроблено класифікацію вступників до закладів вищої освіти.

Отже, як бачимо, метод випадкового лісу, застосування якого і було предметом даного дослідження, є доволі зручним алгоритмом для розробки класифікацій, які у свою чергу є невід’ємними складовими досліджень, де необхідно краще дізнатися структуру та особливості об’єктів. А класифікація, що була розроблена саме у цьому дослідженні, та надані стратегії вибору спеціальностей стануть у нагоді для вироблення рекламної стратегії для поширення інформації про факультет соціології, і допоможе залучити ще більше абітурієнтів до нього у цьому та наступних роках.

Список використаних джерел

1. Чубукова, И. А., 2008. Data Mining. Москва: Интернет-Университет Информационных Технологий (ИНТУИТ).
2. Yonamoris, 2017. Нейронні мережі - шлях до глибинного навчання. [Електронний ресурс] Режим доступу: https://codeguida.com/post/739
3. Awodele O. & Olawale, J., 2009. Neural Networks and Its Application in Engineering. Proceedings of Informing Science & IT Education Conference (InSITE), pp. 83-95.
4. Baden-Fuller, C. & Haefliger, S., 2013. Business models and technological innovation. Long Range Planning, №46(6), pp. 419-426.
5. Bailey, K. D., 1994. Typologies and Taxonomies: An Introduction to Classification Techniques. Los Angeles: Sage Publications Inc.
6. Bailey, K. D., 2005. Typology Construction, Methods and issues. Encyclopedia of Social Measurement, №3, pp. 889-89.
7. Benyamin, D., 2012. A Gentle Introducdon to Random Forests, Ensembles, and Performance Metrics in a Commercial System. [Електронний ресурс] Режим доступу: https://bit.ly/2MuZ6dG
8. Breiman, L., 1996. Bagging predictors. Mach Learn, №24, pp. 123-140.
9. Breiman, L., 2001. Random forests. Mach Learn, №45, pp. 5-32.
10. Brownlee, J., 2019. Bagging and Random Forest Ensemble Algorithms for Machine Learning. Master Machine Learning Algorithms.
11. Donges, N., 2018. The Random Forest Algorithm. Towards Data Science.
12. DŽEROSKI, S., PANOV, P. & ŽENKO, B., 2009. Machine Learning, Ensemble Methods in. Ljubljana, Slovenia: Jožef Stefan Institute.
13. Eisenhardt, K. M., 1989. Building Theories from Case Study Research. Academy of Management Review, №14(4), pp. 532-550.
14. Eulogio, R., 2017. Introducdon to Random Forests. [Електронний ресурс] Режим доступу: https://bit.ly/3cyU4Ya

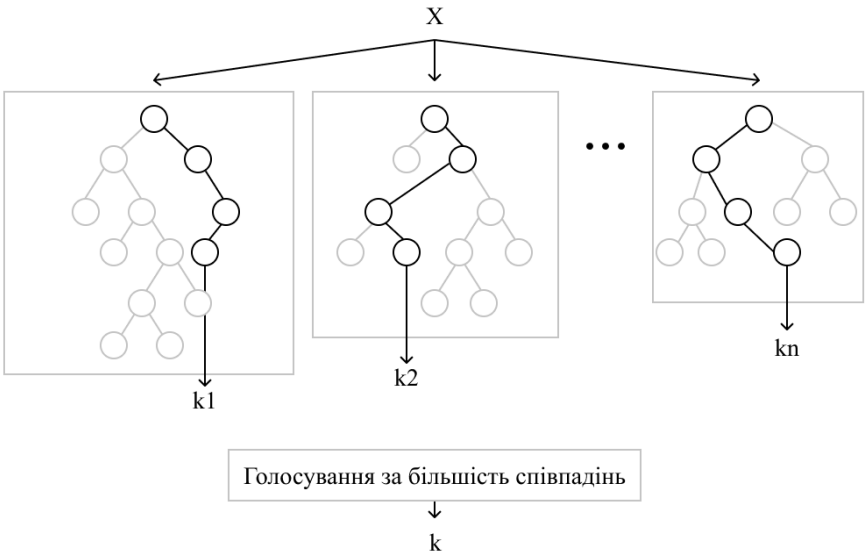
15. Hodge, V. & Austin, J., 2004. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, №22(2), pp. 85-126.
16. Huxley, R., 2007. The Great Naturalists. London: Thames & Hudson.
17. Janitza, S. & Hornung, R., 2018. On the overestimation of random forest's out-of-bag error. [Електронний ресурс] Режим доступу: https://bit.ly/2Y6pxvI
18. Jensen, F., 1996. An Introduction to Bayesian Networks. Springer.
19. Juhi, 2019. Simple guide for ensemble learning methods. [Електронний ресурс] Режим доступу: https://bit.ly/370fOLe
20. Kohonen, T. & Simula, O., 1996. Engineering Applications of the SelfOrganizing Map. Proceeding of the IEEE, №84(10), p. 1354 – 1384.
21. Kon, M. & Plaskota, L., 2000. Information complexity of neural networks. Neural Networks, №13, pp. 365-375.
22. Kotsiantis, S. B., 2007. Supervised Machine Learning: A Review of Classification Techniques, pp. 249-268.
23. Kumar, N., 2019. Advantages and Disadvantages of Random Forest Algorithm in Machine Learning. [Електронний ресурс]. Режим доступу: https://bit.ly/2Y49xKm
24. Lambert, D. S. C., 2015. The Importance of Classification to Business Model Research. Journal of Business Models, №3, pp. 49-61.
25. McKelvey, B., 1982. Organizational Systematics: Taxonomy, Evolution, Classification. University of California Press Berkeley.
26. Nisbet, R., Elder, J. & Miner, G., 2009. Model Evaluadon and Enhancement. Handbook of Stadsdcal Analysis and Data Mining Applicadons, pp. 285–312.
27. Nisbet, R., Miner, G. & & Yale, K., 2018. Classificadon. Handbook of Stadsdcal Analysis and Data Mining Applicadons, pp. 169–186.
28. Pires de Oliveira, S., 2017. A very basic introduction to Random Forests using R. [Електронний ресурс] Режим доступу: https://bit.ly/2MtFjLA

29. Rao, Z. & Alvarruiz, F., 2007. Use of an Artificial Neural Network to Capture the Domain Knowledge of a Conventional Hydraulic Simulation Model. Journal of HydroInformatics, pp. 15-24.
30. Reinstein, I., 2017. Random Forests(r), Explained. KDnuggets.
31. Rich, P., 1992. The Organizational Taxonomy: Definition and Design. Academy of Managment Review, pp. 758-781.
32. Sathya, R. & Abraham, A., 2013. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. International Journal of Advanced Research in Artificial Intelligence, №2(2), pp. 34-38.
33. Simpson, G. G., 1961. Principles of Animal Taxonomy.Columbia University Press New York.
34. Sneath, P. H. & Sokal, R. R., 1973. Numerical Taxonomy: The Principles and Practice of Numerical Taxonomy. San Francisco: W.H. Freeman and Company.
35. Strobl, C., Malley, J. & Tutz, G., 2009. An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychol Methods, №14(4), pp. 323-348.
36. Sukumaran, S. & Kesavaraj, G., 2013. A Study on Classificadon Techniques in Data Mining.
37. Warriner, C. K., 1984. Organizations and Their Environments: Essays in the Sociology of Organizations. JAI Press London.
38. Tan, L., 2015. Code Comment Analysis for Improving Soâware Quality. The art and Science of Analyzing Soâware Data, pp. 493-517.

Додатки

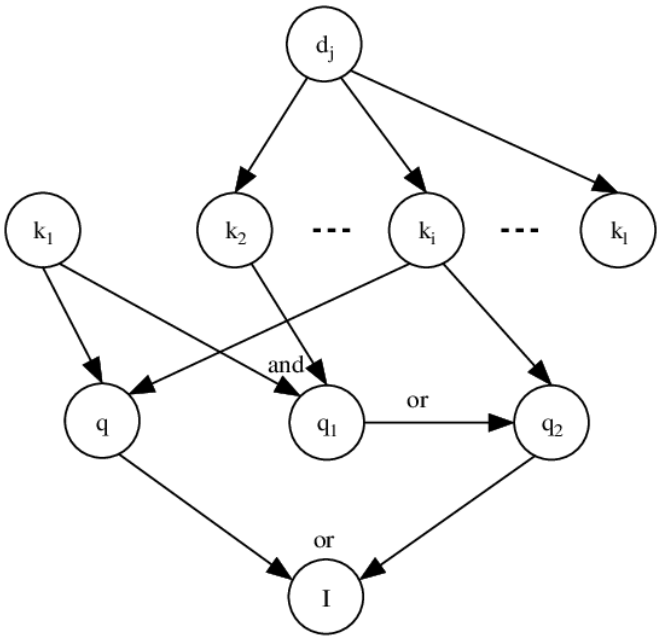
Додаток 1.

Приклад дерева рішень



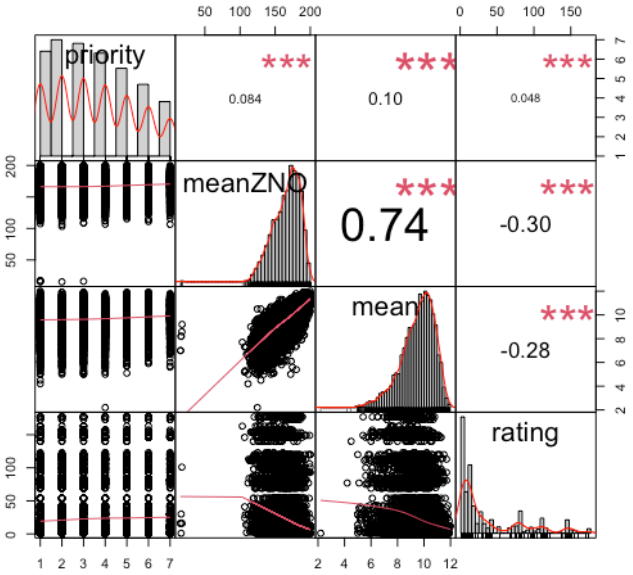
Додаток 2.

Приклад Баєсівської мережі



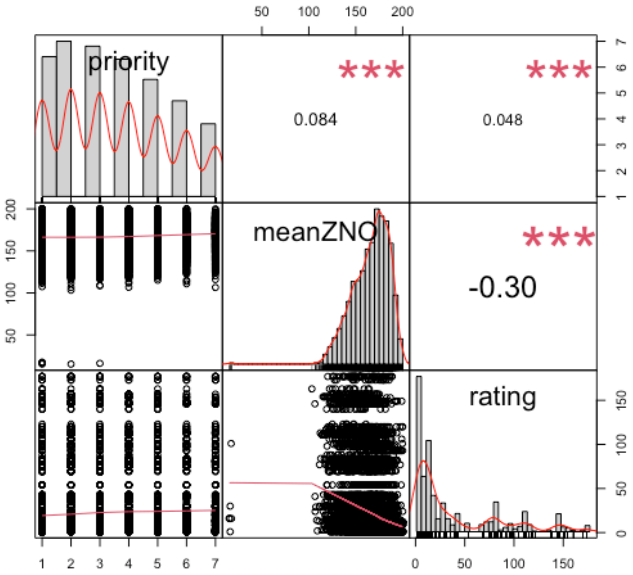
Додаток 3.

Таблиця коефіцієнтів кореляції із високою кореляцією між незалежними змінними



Додаток 4.

Таблиця коефіцієнтів кореляції із задовільною кореляцією між незалежними змінними



Додаток 5.

Результати побудови моделі для класифікації

> model$results

mtry	Accuracy
2	0.7529077
3	0.7418808

Додаток 6.

Порівняння результатів, що були спрогнозовані моделлю та тих даних, що вже реально маємо

> predicteddata <- predict(model, data, "prob")

> head(predicteddata)

Ймовірність того, що обрано спеціальність «Соціологія»		**Чи дійсно обрано спеціальність «Соціологія»**
Ні	**Так**	
0.868	0.132	Ні
0.798	0.202	Ні
0.990	0.010	Ні
0.002	0.998	Так
0.218	0.782	Так
0.954	0.046	Ні

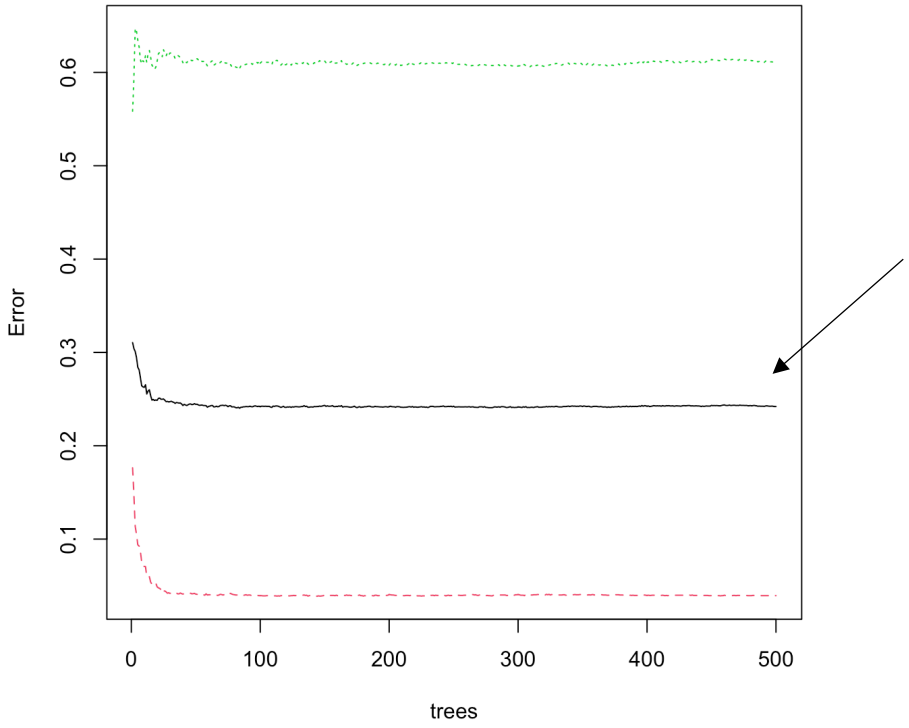
Додаток 7.

Таблиця частот змінної «Спеціалізація»

	Var1	Freq
1	Політологія	4214
2	Соціологія	2318

Додаток 8.

Оцінка похибки методу в залежності від кількості побудованих дерев



Додаток 9.

Значення «важливості» змінних у моделі

	MeanDecreaseGini
Priority	70.17891
meanZNO	383.43702
rating	705.95169

Додаток 10.

Інформація про модель для спеціальності «Політологія»

> model$results

mtry	Accuracy
2	0. 7565853
3	0. 7432671

> predicteddata <- predict(model, data, "prob")

> head(predicteddata)

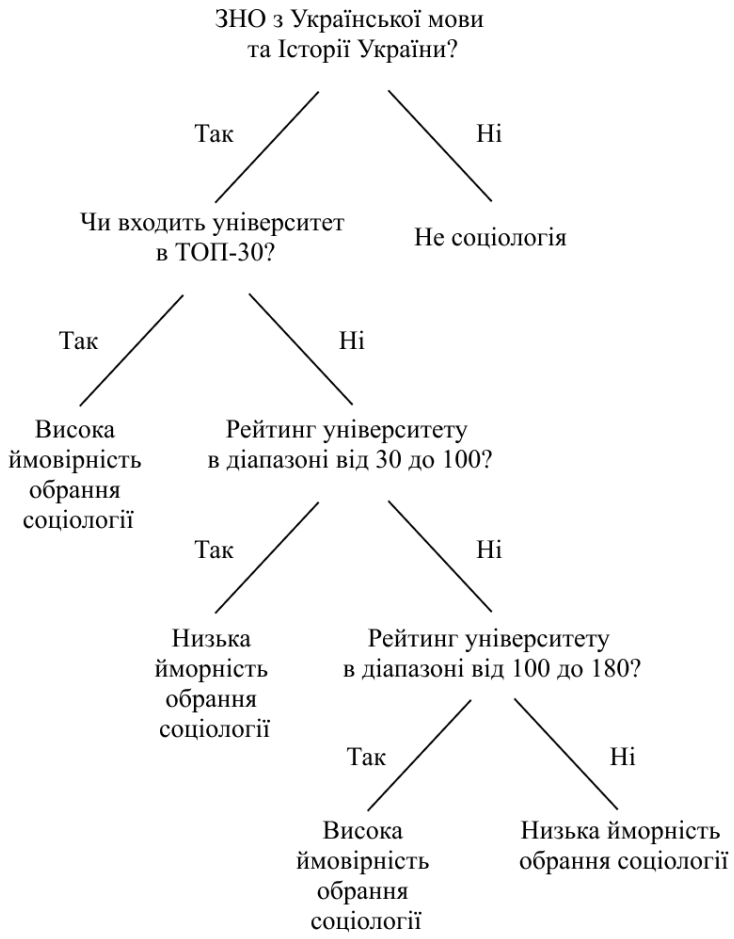
Ймовірність того, що обрано спеціальність «Політологія»		Чи дійсно обрано спеціальність «Політологія»
Ні	Так	
0.122	0.878	Так
0.230	0.770	Так
0.006	0.994	Так
1.000	0.000	Ні
0.792	0.208	Ні
0.042	0.958	Так

> randomForest::importance(output.forest)

	MeanDecreaseGini
Priority	65.81209
meanZNO	470.67593
rating	530.23402

Додаток 11.

Стратегія вибору спеціальності «Соціологія»



Додаток 12.

Стратегія вибору спеціальності «Політологія»