Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://nv.nltu.edu.ua
https://doi.org/10.36930/40290827
Article received 26.10.2019 p.
Article accepted 31.10.2019 p.
УДК 004.[9+89]

ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

@ Correspondence author
V. R. Verhun
vverhun@gmail.com

В. Р. Вергун
Національний університет "Львівська політехніка", м. Львів, Україна

ЗАСТОСУВАННЯ ДЕРЕВ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ АНАЛІЗУ ВПЛИВУ НЕАКАДЕМІЧНИХ ЧИННИКІВ НА ПОЧАТКОВИЙ РІВЕНЬ ЗНАНЬ СТУДЕНТІВ

Проаналізовано резюме, що подали кандидати на навчальні курси в галузі інформаційних технологій. Проаналізовано потенційні чинники, що можуть бути включені до вибірки для проведення експерименту. З цього списку резюме вибрано незалежні неакадемічні чинники, які беруться до уваги в дослідженні. Ці чинники можуть мати вплив на успішність кандидатів, що розпочинають навчання на освітніх програмах із напряму інженерії програмного забезпечення, та можуть бути розглянуті у вирішенні задачі прогнозування успішності. На основі цієї вибірки чинників розглянуто методи інтелектуального аналізу даних для класифікації кандидатів, беручи за основу результати проходження тесту на виявлення початкового рівня знань. Під час експерименту використано алгоритми генерації дерев прийняття рішень. Алгоритми, які застосовувалися під час дослідження: J48, LMT, Random Forest, Random Tree. Для оцінки точності класифікації застосовували метод перехресної перевірки. Проведено оцінку атрибутів, що враховуються під час експерименту. Згенеровано дерево прийняття рішень для аналізу чинників, що впливають на початковий рівень знань. Здійснено порівняння вибраних алгоритмів за точністю та швидкодією. Експериментальним способом виявлено основний чинник, що має найбільший вплив на якість проходження тесту на початковий рівень знань. Виявлено другорядні чинники, що також мають вплив на проходження тесту.

Ключові слова: інтелектуальний аналіз даних навчальних програм; класифікація; дерева класифікацій; продуктивність; J48; LMT; Random Forest; Random Tree; метод перехресної перевірки; навчальні програми; IT; інженерія програмного забезпечення; прогнозування; порівняння алгоритмів.

Вступ. Якість та постійне оновлення навчальних планів є предметом постійного аналізу та досліджень навчальними установами. Беручи до уваги потребу в висококваліфікованих працівниках, а також постійне збільшення уваги до креативних професій, всі навчальні програми можна оцінити за такими характеристиками: тривалість, ціна, охоплення знань. Якщо коротко – найцінніші є навчальні програми, які найбільш доступні як за критерієм ціни, так і за інструментами та нетривалим періодом навчання (Hug et al., 2019). Такі навчальні програми повинні надавати всі необхідні знання та інструменти для того, щоби оволодіти необхідними знаннями та поведінкою, які критично необхідні для подальшого кар'єрного зростання.

Загалом поєднання навичок, знань та поведінки є найважливішим критерієм для оцінювання прогресу студентів упродовж навчання, а також пізніше у виконанні регулярних робочих обов'язків (Moore et al., 2002).

Індустрія інформаційних технологій відіграє важливу роль у зростанні економіки в Україні загалом. Приблизна кількість спеціалістів, залучених до індустрії, перевищила 90 000 і експерти очікують зростання вдвічі впродовж наступних трьох років (Demchenko, 2018).

Отже, з огляду на популярність і великі можливості, щороку збільшується інтерес до IT індустрії з боку вже працевлаштованого населення в інших галузях. А для того, щоби отримати відповідну кваліфікацію, потрібно приділяти значну увагу освіті, і це є найкращим способом отримати необхідні навички.

Навчальні заклади можна поділити на 3 категорії: університети та інститути, приватні університети, корпоративні університети. І всі вони стикаються з однаковими проблемами під час навчального процесу, якщо брати до уваги IT напрямок (Pasichnyk & Kunanets, 2015; Krasnikov, 2016).

Першою проблемою є процес створення навчального плану та графіків навчання. Через швидку зміну технологій та підходів до розроблення програмного забезпечення навчальні програми не повністю покривають всі потреби індустрії. Індустрія не має можливості чекати довгий час на підготовку та перекваліфікацію працівників. Тому здатність до швидкого та автономного розроблення програм та висока адаптованість є критичними навичками будь-яких навчальних закладів.

Іншою проблемою є початковий рівень кандидатів, які хочуть розпочати навчання на освітніх програмах, та коректне визначення цього рівня. Знання на початковому етапі навчання є критично важливими для того, щоб приймати правильні рішення щодо участі в навчальних програмах певних кандидатів, та для створення більш персоналізованих програм. Такі програми дають

Інформація про авторів:
Вергун Володимир Ростиславович, аспірант, кафедра автоматизованих систем управління. **Email:** vverhun@gmail.com;
https://orcid.org/0000-0003-0683-0841

Цитування за ДСТУ: Вергун В. Р. Застосування дерев прийняття рішень для аналізу впливу неакадемічних чинників на початковий рівень знань студентів. Науковий вісник НЛТУ України. 2019, т. 29, № 8. С. 147–151.
Citation APA: Verhun, V. R. (2019). Application of decision trees for analysis of the influence of non-academic factors on the initial level of students' knowledge. *Scientific Bulletin of UNFU, 29*(8), 147–151. https://doi.org/10.36930/40290827

змогу зменшити час навчання, з більшою ймовірністю завершити весь цикл і отримати необхідні знання та навички.

Наступною проблемою є терміни та графіки навчання. Студенти повинні засвоїти достатньо багато матеріалу в короткі терміни. Відповідно, добрі навики самонавчання є дуже важливими.

Під час цього дослідження ми дискутуємо можливості прогнозування початкового рівня кандидатів, беручи до уваги неакадемічні чинники, та виключаючи попередні оцінки за успішність, середній бал тощо. Всі атрибути були згенеровані з публічних резюме кандидатів. Ми не розглядаємо результати Зовнішнього Незалежного Оцінювання. На наше переконання, ці результати мають занадто очевидний вплив на здатність кандидатів до навчання (Likarchuk et al., 2010). Оцінки за навчання в університеті також не враховуються через непрозору систему оцінювання в різних навчальних закладах (Vasylyeva & Merkle, 2018).

У цьому дослідженні ми не дискутуємо коректність визначення початкового рівня знань кандидатів. Ми беремо до уваги тільки конкретний результат певного кандидата, незважаючи на сам принцип визначення.

Метою дослідження є аналіз впливу неакадемічних чинників на початковий рівень знань студентів шляхом застосування дерев прийняття рішень. Для досягнення зазначеної мети необхідно вирішити такі *основні завдання*:

- створити набір неакадемічних чинників, які можуть впливати на початковий результат оцінювання кандидата;
- проаналізувати вибірку за допомогою алгоритмів побудови дерев прийняття рішень;
- проаналізувати отримані результати на предмет точності та швидкодії;
- визначити чинники, які мають найбільший вплив на початковий результат та можна брати до уваги для розв'язання задач прогнозування.

Аналіз літературних джерел. Інтелектуальний аналіз даних є ітеративним процесом видобутку корисної інформації та шаблонів з організованих чи неорганізованих масивів даних, різноманітних форматів. Цей підхід часто використовують для прогнозування успішності суб'єктів навчальних програм. Алгоритми, які добувають корисні знання з даних, є основним ядром будь-якої системи прийняття рішень (Hien et al., 2007; Livieris et al., 2016, 2017). Окрім різноманітних підходів до прогнозування успішності, важливим є також контекст та предметна галузь, конкретний навчальний заклад і залежні від них характеристики. Все це зумовлює постійну потреба до вдосконалення наявних алгоритмів та їх визначальних особливостей (Zughoul et al., 2018).

Кластеризацію та класифікацію найчастіше використовують у розв'язанні задач в інтелектуальному аналізі даних навчальних програм. Більшість досліджень зосереджені на вирішенні таких запитань:

- які є визначальні чинники, що мають вплив на успішність у навчальному процесі?
- якою є точність вибраних методів та підходів?

У всіх дослідженнях є актуальною проблема вибору даних та атрибутів. Здебільшого це невелика кількість даних та однотипні чинники. Велику кількість даних за фактом можна отримати, аналізуючи інформацію з онлайн курсів, де навчаються одночасно велика кількість слухачів. Проте, під час врахування чинників, у випадку онлайн курсів, неможливо враховувати оцінку успішності вибраного студента. Цей чинник легко порівнювати через його точне значення (Hellas et al., 2018), але слухачі онлайн курсів відрізняються віком, освітою, досвідом, тому прогнозування успішності та вибір точних методів є актуальною науковою задачею. Поточну середню оцінку також беруть до уваги під час розв'язання задачі прогнозування (Sarker et al., 2013), але, у випадку онлайн курсів, поточний результат навчання є неактуальним, оскільки зріз знань здебільшого є добровільним бажанням конкретного слухача.

Персональні характеристики також часто розглядають у багатьох дослідженнях. Такими є, наприклад, вік та стать. Причиною, чому розглядають стать, є факт, що особи жіночої статі мають свій, відмінний стиль навчання. Вони більш дисципліновані та сфокусовані. У них вищий рівень самомотивації і вони приділяють більше уваги самодисципліні (Meit et al., 2007).

Вік також часто використовують у багатьох дослідженнях (Verma et al., 2019). Існують дослідження, які прямо вказують, що молодші студенти краще справляються з матеріалом та засвоєнням знань, ніж старші колеги (Pellizzari et al., 2012). Відповідно до дослідження (Hong, 1984), вік переважає інші чинники, такі як: метод навчання та середовище навчання. Під час цього дослідження також було встановлено, що персональні звички мають доволі невеликий ефект. Чинники, що стосуються персональних якостей, таких як: інтереси, хобі, підтримка сім'ї, не часто розглядаються в алгоритмах прогнозування, оскільки важко отримати вимірювані результати і здійснювати вимірювання чи порівняння.

Розглядають також деякі соціо-економічні чинники, такі як: освіта батьків, загальний прибуток сім'ї, місце народження та місце проживання. Наприклад, було доведено, що освіта матері і загальний прибуток сім'ї високо корелюються з результатами навчання студента (Devasia et al., 2016). Проте під час дослідження (Naqvi, 2006) такий чинник не було доведено. Хоча вік матері виявився доволі важливим чинником, молоді матері можуть легше і швидше впливати та мотивувати студентів, аніж старші батьки.

Іншим цікавим полем для досліджень є вплив психосоціальних чинників на успішність студентів. Деякі дослідження вказують, що дуже важливими чинниками впливу є загальне ставлення та взаємодія під час навчання (Garba et al., 2017). Самооцінка немає відношення до загальної успішності (Hoeschler et al., 2014). Іншими чинниками, які впливають на успішність, є стрес, правильне управління часом, залучення в різноманітні університетські активності та загальне емоційне задоволення від академічного навчання. Комунікація у класі під час заняття не впливає безпосередньо на середню оцінку успішності (Krumrei et al., 2013).

З використанням методів інтелектуального аналізу даних прогностичне моделювання зазвичай використовують для прогнозування успішності студентів. Найчастіше дослідження зосереджені над розв'язанням задач класифікації. Серед алгоритмів, що використовуються, найчастіше трапляються Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor та Support Vector Machine. Проте традиційні алгоритми та підходи інтелектуального аналізу даних не можна застосовувати до вирішення проблем у навчальному про-

цесі, оскільки вони можуть мати специфічну мету та функцію. Це означає, що спочатку потрібно застосувати алгоритм попередньої обробки і тільки тоді можна застосувати деякі специфічні методи аналізу даних. Одним із таких алгоритмів попередньої обробки є кластеризація (Dutt, Ismail & Herawan, 2017).

Загалом кількість досліджень у сфері інтелектуального аналізу даних у освітніх програмах швидко зростає, а також зростає різноманітність методів, що використовуються (Hellas et al., 2018). Дерева прийняття рішень залишаються однією з найпопулярніших технік для розв'язання задач прогнозування (Manjarres et al., 2018).

Під час використання техніки дерев прийняття рішень у своїх дослідженнях автори використовують різноманітний набір алгоритмів та здебільшого вибірки з невеликої кількості елементів (Kumar et al., 2012). Беручи до уваги неоднорідні результати точності вибраних алгоритмів дерев прийняття рішень, можна зауважити, що структура вибраних атрибутів для дослідження відіграє велику роль в остаточних результатах.

Метод дослідження. Дані для цього дослідження щодо впливу чинників на успішність проходження тесту для визначення початкових знань було зібрано у 101 кандидатів, які зареєструвались на онлайн курс з програмування. Результати такого тесту розглядали як початковий рівень кандидата. Напрям підготовки – інженерія програмного забезпечення. У межах процесу початкового оброблення дані студентів переважно вибирали з резюме. Оскільки неможливо достовірно визначити академічні успіхи слухачів під час їх навчання, до уваги брали тільки неакадемічні чинники, такі як: результат початкового тесту (test score), вік (age), стать (gender), освіту (degree), попередній досвід (experience) та наявність додаткових сертифікацій (training). Результат тесту вважали успішним, якщо слухач відповів на 66% поставлених запитань. Запропонували 3 категорії віку слухачів: ≤22 роки, поточний студент, >22 та <29 – випускник, ≥29 – здебільшого це заявники, що хочуть змінити професію. Освіта в слухачів або профільна, оскільки це онлайн курс з програмування, або не профільна. Досвід розглядали будь-який, якщо він існує, сертифікації розглядали тільки профільні, якщо вони існують.

Набір даних, що розглядався, містить 73 % чоловіків і 28 % жінок-заявників. Це співвідношення відображає наявний розподіл у IT-секторі (Sapiton, 2019). 51 % заявників з набору даних пройшли додаткові тренінги, пов'язані з галуззю інженерії програмного забезпечення.

Алгоритми побудови моделі класифікують студентів у 2 категорії залежно від результатів початкового тесту. У нашій вибірці 23 % кандидатів правильно відповіли на більше як 66 % запитань під час тестування. Під час дослідження було використано декілька різних алгоритмів. Для аналізу та візуалізації використовували програмне забезпечення Weka. Під час дослідження застосовували алгоритми дерева прийняття рішень. Ці алгоритми дають змогу отримати модель, яка легко читається та яку легко можна проаналізувати для подальшого використання. Алгоритми, які застосовувалися під час дослідження: J48, LMT, Random Forest, Random Tree. Для оцінки точності класифікації застосовували метод перехресної перевірки. Процес перехресної перевірки повторюється 10 разів для кожного виконання алгоритму.

Для порівняння продуктивності, корисності та точності обраних алгоритмів використовували такі метрики: точність, F-вимір, правильно класифіковані екземпляри, неправильно класифіковані екземпляри.

Для перевірки атрибутів застосували визначення та аналіз надлишкових атрибутів. Оцінку здійснювали на всіх наборах даних навчання. Використовували метод пошуку рейтингу атрибутів. Результат оцінки атрибутів зображено на рис. 1.

```
Attribute Evaluator (supervised, Class (nominal): 1 test score):
        Correlation Ranking Filter
Ranked attributes:
 0.2034  4 degree
 0.1512  5 experience
 0.1127  6 training
 0.0726  3 gender
 0.0666  2 age
Selected attributes: 4,5,6,3,2 : 5
```

Рис. 1. Оцінка атрибутів

Згідно з рис. 1, атрибут із найвищим рангом є Освіта, а найнижчий ранг має атрибут Вік серед 5 вибраних атрибутів (Освіта, Досвід, Курси, Стать, Вік).

Результати дослідження. У табл. 2 виявлено точність, отриману за обраними алгоритмами, з урахуванням точності, F-виміру, правильно класифікованих екземплярів, некоректно класифікованих випадкових метрик. Як видно з цієї таблиці, найкращу середню точність на нашій вибірці отримали під час використання алгоритму Random Tree.

Табл. 2. Порівняння точності алгоритмів

Алгоритм	Точність	F-Вимір	Правильно класифіковані, %	Неправильно класифіковані, %
J48	0,591	0,658	74,2574	25,7426
LMT	0,593	0,663	75,2475	24,7525
Random Forest	0,663	0,684	72,2772	27,7228
Random Tree	0,708	0,711	71,2871	28,7129

Розмір дерева – 47. Матрицю неточності для цього алгоритму подано на рис. 2.

```
  a  b   <-- classified as
  8 15 |  a = YES
 14 64 |  b = NO
```

Рис. 2. Матриця неточності

На рис. 3 подано сформоване дерево рішень алгоритмом Random Tree.

Модель, отриману алгоритмом дерева, легко читаємо і розуміємо. Представлений алгоритм має достатню точність. Алгоритм J48 показав дещо точніший або однаковий результат, ніж у інших дослідженнях (Osmanbegovic & Suljic, 2012; Hussain et al., 2018). У дослідженні (Hussain et al., 2018) алгоритм Random Forest показав 100% точний результат. Але, згідно з результатами (Rao et al., 2016), цей алгоритм вже не є точним, і збігається з точністю цього дослідження, що, в принципі, підтверджує тезу про важливість кількості вибраних даних та структури вибраних атрибутів.

Висновки. Під час цього дослідження проаналізовано базові чинники, які можуть вплинути на результати початкового тесту кандидатів для навчання програми з інженерії програмного забезпечення. Було взято до уваги лише такі чинники, які можна легко зібрати з резюме кандидатів. Для визначення чинників, які найбільше

впливають на результат, були застосовані алгоритми для генерування дерева прийняття рішень. Під час дослідження було встановлено, що алгоритми дерева рішень найбільш застосовні для прийняття рішень щодо успішності студентів. Такі алгоритми є зручними для користувача й отримують максимальну точність серед інших. Для вимірювання продуктивності та ефективності використали чотири алгоритми: J48, LMT, Random Forest, Random Tree.

```
degree = N
|  gender = M
|  |  training = Y
|  |  |  experience = Y
|  |  |  |  age = > 22 and < 29 : YES (2/0)
|  |  |  |  age = <= 22 : YES (0/0)
|  |  |  |  age = >= 29 : NO (8/2)
|  |  |  experience = N
|  |  |  |  age = > 22 and < 29 : NO (4/0)
|  |  |  |  age = <= 22 : YES (3/1)
|  |  |  |  age = >= 29 : NO (2/0)
|  |  training = N
|  |  |  age = > 22 and < 29 : NO (10/0)
|  |  |  age = <= 22
|  |  |  |  experience = Y : YES (2/1)
|  |  |  |  experience = N : NO (4/0)
|  |  |  age = >= 29 : NO (5/1)
|  gender = F : NO (14/0)
degree = Y
|  age = > 22 and < 29
|  |  gender = M
|  |  |  experience = Y
|  |  |  |  training = Y : YES (2/1)
|  |  |  |  training = N : NO (3/1)
|  |  |  experience = N : NO (5/1)
|  |  gender = F
|  |  |  training = Y : YES (2/0)
|  |  |  training = N : NO (2/0)
|  age = <= 22
|  |  gender = M
|  |  |  experience = Y
|  |  |  |  training = Y : YES (3/1)
|  |  |  |  training = N : NO (2/0)
|  |  |  experience = N
|  |  |  |  training = Y : NO (6/0)
|  |  |  |  training = N : YES (8/4)
|  |  gender = F
|  |  |  experience = Y : NO (1/0)
|  |  |  experience = N
|  |  |  |  training = Y : NO (3/1)
|  |  |  |  training = N : NO (5/1)
|  age = >= 29
|  |  experience = Y
|  |  |  training = Y : NO (1/0)
|  |  |  training = N : NO (3/1)
|  |  experience = N : YES (1/0)
```

Рис. 3. Дерево прийняття рішень

Результати проведеного дослідження підтверджують, що кандидати, які не мають відповідного рівня знань та досвіду, реально не мають шансів успішно пройти початковий відбірковий тест для навчання на програмі.

Окрім того, встановлено, що значний вплив має чинник віку. Заявники з останньої вікової категорії, порівняно з іншими чинниками, показують гірші результати. Додаткова освіта та будь-який досвід роботи також мають великий вплив на проходження початкового тесту.

Найвищий вузол рішення у дереві, який відповідає кращому чиннику прогнозування, у нашому дослідженні, є чинник вищої освіти. Відповідно встановлено, що освітній рівень залишається найвагомішим чинником, що впливає на успішність студентів у нашій вибірці.

Як галузь для подальших досліджень, ми пропонуємо розширити список чинників для того, щоб знайти більш приховані чинники впливу на успішність студентів. Також необхідно здійснити додаткові дослідження серед заявників віком від 29 років, оскільки очевидно, що це окремий сегмент, який можна вважати іншим поколінням з іншими впливовими чинниками. Напрям освіти для розроблення програмного забезпечення набуває дедалі більшої популярності, тому такі дослідження претендентів мають перспективу для подальшого аналізу.

Перелік використаних джерел

Demchenko, Dmytro. (2018). Labor market 2018 according to DOU: salaries and number of IT Specialists are increasing. *Ukraine – Startup And Technology News.* Retrieved from: https://ain.ua/en/2018/12/28/labor-market-2018-according-to-dou/

Devasia, Tismy & T P, Vinushree & Hegde, Vinayak. (2016). *Prediction of students performance using Educational Data Mining*. 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access, 5,* 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247

Garba Kolo, Ali & Munira Binti, Wan & Wan Jaafar, Wan & Binti Ahmad, Nobaya. (2017). *Influence of Psychosocial Factors on Students Academic Performance in One of Nigerian Colleges of Education*.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., et al. (2018). Predicting Academic Performance: A Systematic Literature Review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. (pp. 175–199). (ITiCSE 2018 Companion). New York, NY, USA: ACM. https://doi.org/10.1145/3293881.3295783

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, Vangel, V., et al. (2018). Predicting academic performance. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (ITiCSE 2018 Companion). ACM, New York, NY, USA. (pp. 175–199). https://doi.org/10.1145/3293881.3295783

Hien, Thi Ngoc, Nguyen & Haddawy, Peter. (2007). A decision support system for evaluating international student applications. Proc. Frontiers Educ. Conf. F2A-1. https://doi.org/10.1109/FIE.2007.4417958

Hoeschler, Peter & Backes-Gellner, Uschi. (2014). *College dropout and self-esteem*. https://doi.org/10.5167/uzh-102178

Hong, Sun−Mook. (1984). The Age Factor in the Prediction of Tertary Academic Success. *Higher Education Research & Development, 3*(1), 61–70. https://doi.org/10.1080/0729436840030105

Hug, Theo, Lindner, Martin & Bruck, Peter. (2019). *Microlearning: Emerging Concepts, Practices and Technologies after e-Learning*.

Hussain, Sadiq & Abdulaziz Dahan, Neama & Ba-Alwib, Fadl & Najoua, Ribata. (2018). Educational Data Mining and Analysis of Students Academic Performance Using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science, 9,* 447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Krasnikov, By Denys. (2016). Growing IT industry to fuel tech education evolution in Ukraine. *The Kyiv Posts IT coverage is sponsored by Ciklum. The content is independent of the donors.* Retrieved from: https://www.kyivpost.com/article/content/ukraines-it-edge/growing-it-industry-to-fuel-tech-education-evolution-in-ukraine-414137.html

Krumrei, E. J., Newton, F. B., Kim, E., & Wilcox, D. (2013). *Psychosocial factors predicting first-year college student success.* Retrieved from: http://krex.ksu.edu

Kumar, Surjeet & Bharadwaj, Brijesh & Pal, Saurabh. (2012). Mining Education Data to Predict Students Retention: A comparative Study. *International Journal of Computer Science and Information Security, 10,* 113–117.

Likarchuk, I., Rakov, S., Gudzynsky, V., & Kovtunets, V. (2010). Quality of Universities Admission Based on External Independent Assessment in Ukraine. *36th annual conference, International Association for Educational Assessment*.

Livieris, I. E., Drakopoulou, K., Kotsilieris, T., Tampakas, V., & Pintelas, P. (2017). DSS-PSP – A Decision Support Software for Evaluating Students Performance. In G. Boracchi, L. Iliadis, C. Jayne, A. Likas (Eds), *Engineering Applications of Neural Networks*. EANN 2017. Communications in Computer and Information Science, 744. Springer, Cham.

Livieris, I., & Mikropoulos, Tassos & Pintelas, P. (2016). A decision support system for predicting students performance. *Themes in Science & Technology Education, 9,* 43–57.

Manjarres, Andrés Villanueva, Luis Gabriel Moreno Sandoval, & Martha Salinas Suárez. (2018). Data mining techniques applied in educational environments: Literature Review. *Digital Education Review, 33,* 235–266.

Meit, Scott & J. Borges, Nicole & A. Early, Larry. (2007). Personality Profiles of Incoming Male and Female Medical Students: Results of a Multi-Site 9Year Study. *Medical Education Online.* 12. 1 https://doi.org/10.3402/meo.v12 i.4462

Moore, David R., Cheng, Mei-I. & Dainty, Andrew, R. J. (2002). Competence, competency and competencies: performance assessment in organisations. *Work Study, 51*(6), 314–319, https://doi.org/10.1108/00438020210441876

Naqvi, Sayyed. (2006). Factors affecting students performance a Case of Private Colleges. *Bangladesh e-Journal of Sociology, 3*.

Osmanbegovic, Edin, & Suljic, Mirza. (2012). Data Mining Approach for Predicting Student Performance: Economic Review. *Journal of Economics and Business, 10*(*1*), *3–12.* University of Tuzla, Faculty of Economics, Tuzla.

Pasichnyk, V., & Kunanets, N. (2015). IT education and IT business in Ukraine: Responses to the modern challenges. *Computer Sciences and Information Technologies: 10th International Scientific and Technical Conference*, Lviv. (pp. 48–51).

Pellizzari, Michele, & Francesco C. Billari. (2012). The Younger, the Better? Age-Related Differences in Academic Performance at University. *Journal of Population Economics, 25*(2), 697–739. JSTOR, www.jstor.org/stable/41408932.

Rao, Prasad & Chandra, M. V. P. & Ramesh, B. (2016). Predicting Learning Behavior of Students using Classification Techniques. *International Journal of Computer Applications, 139,* 15–19. https://doi.org/10.5120/ijca2016909188

Sapiton, Mike. (2019). How the IT industry of Ukraine and Eastern Europe works: a report. *Ukraine – Startup And Technology News.* Retrieved from: https://ain.ua/en/2019/02/15/it-industry-of-ukraine-and-eastern-europe/

Sarker, F. et al. (2013). Students performance prediction by using institutional internal and external open data sources. *CSEDU13: 5th International Conference on Computer Supported Education*, Aachen, Germany.

Vasylyeva, A., & Merkle, O. (2018). "Combatting corruption in higher education in Ukraine," MERIT Working Papers 021, *United Nations University – Maastricht Economic and Social Research Institute on Innovation and Technology* (MERIT).

Verma, Chaman & Illés, Zoltán & Stoffova, Veronika. (2019). *Age Group Predictive Models for the Real Time Prediction of the University Students using Machine Learning: Preliminary Results*.

Zughoul, O., et al. (2018). *Comprehensive Insights Into the Criteria of Student Performance in Various Educational Domains*. In IEEE Access, 6, 73245–73264. https://doi.org/10.1109/ACCESS.2018.2881282

V. R. Verhun

Lviv Polytechnic National University, Lviv, Ukraine

APPLICATION OF DECISION TREES FOR ANALYSIS OF THE INFLUENCE OF NON-ACADEMIC FACTORS ON THE INITIAL LEVEL OF STUDENTS' KNOWLEDGE

The resumes submitted by candidates for educational program in information technology area are analyzed. Potential factors that may be included in the experiment are studied. Previous studies in this field have been examined. Factors that have mostly often been considered by authors of other studies have been identified. The problems that are mostly often solved by different Data Mining approaches are surveyed. The problems mostly often encountered by the authors of the research were distinguished. The current forecasting algorithms are determined not to consider all the information. Now, it is mostly only one-time tasks, which in most cases do not consider the student's progress in studying. Independent non-academic factors that are considered in the study are selected from the list of resumes. The descriptions of each factor are provided. These factors may have an impact on the success of applicants who begin training in software engineering programs and may be considered in the success prediction task. Based on this sample of factors and taking into account the results of passing the initial level of knowledge test, the different approaches of data mining for candidates' classification were considered. The decision trees algorithms used in the study are as follows: J48, LMT, Random Forest, and Random Tree. The cross-validation method was used to evaluate the classification accuracy. The attributes that were considered during the experiment were evaluated. A decision tree was generated to analyze the factors that have the most affect to the initial level of knowledge. The performance of the selected algorithms is compared. All the necessary results of the study are presented, as well as the result of the generated decision tree, which determines the most important factors. The main factor that has the greatest influence on the quality of passing the test to the initial level of knowledge has been experimentally proved. Other minor factors have been identified that also has influence on the results of the test. Some possibilities for further research are analyzed. Therefore, we should conclude that the education industry in information technology has a great perspective and has a huge potential for research.

Keywords: data mining of training programs; classification; classification trees; productivity; J48; LMT; Random Forest; Random Tree; cross-checking method; educational programms; IT; software engineering; prognostication; comparison of algorithms.