Інститут проблем реєстрації інформації
Національної академії наук України

О. Г. Додонов, А. І. Кузьмичов

Датамайнінг в Excel

Розвідувальний аналіз даних та прогнозування з використанням надбудови *Analytic Solver Data Mining*

Київ
Видавництво Ліра-К
2023

УДК 004.67
Д60

Рекомендовано до друку Вченою радою
Інституту проблем реєстрації інформації НАН України.
(протокол № 8 від 5 липня 2022 р.)

Додонов О. Г., Кузьмичов А. І.

Д60 Датамайнінг в Excel. Розвідувальний аналіз даних та прогнозування з використанням надбудови Analytic Solver Data Mining. – Київ : Видавництво Ліра-К, 2023. – 240 с.

ISBN 978-617-520-424-5

Розвідувальний аналіз даних та датамайнінг (Exploratory Data Analysis & Data Mining) – міждисциплінарна методологія на тлі «великих даних», новітні інформаційні технології і процедури, зорієнтовані на виявлення вад в наборах табличних даних великих обсягів про об'єкти, згідно поставлених цілей.

Зазвичай ці набори «сирі», якщо отримані із зовнішніх джерел і, скоріше за все, невідомого походження, чи вони відомі і робочі, що регулярно застосовуються, але пошкодженні, випадково чи штучно. З-за їх великих розмірів шукану інформацію «видобувають»/«майнять» із даних досконалими, потужними й дорого вартісними комп'ютерними засобами, аби швидко й із найбільшою достовірністю зрозуміти їх природу і визначити наслідки виявлених негативних впливів на дані.

Видання призначене для студентів та користувачів-початківців з математичною та програмістською підготовкою на рівні повної середньої освіти, що цікавляться проблематикою Data Science, але у яких для практичної реалізації методів єдиним доступним і зрозумілим обчислювачем є табличний процесор MS Excel. Для них, без потреби щось програмувати, використана надбудова Analytic Solver Data Mining у складі ASPE (Analytic Solver Platform for Education, www.solver.com). Її інструменти разом зі стандартними засобами Excel застосовуються для підготовки і розвідки отриманих наборів даних й подальшого розв'язання задач кластеризації, класифікації та передбачення за технікою машинного навчання.

УДК 004.67

ISBN 978-617-520-424-5

ЗМІСТ

ВСТУП

Розвідувальний аналіз стосується потенційно чи реально пошкоджених даних, де неодмінна передобробка і перетворення спеціальними інструментальними засобами з метою подальшого вилучення корисної і цінної інформації про об'єкти, представленими цими даними, в лаконічних формах, як-от візуалізацією. Історично, це напрямок і складова статистичного аналізу, що охоплює різноманітні підходи експериментальної техніки і засоби досліджень, використовується у різних сферах практики, зазвичай, для передбачувальних/прогнозних цілей[1].

Джон Тьюки, відомий американський математик і статистик, ще у 1970-ті рр., із появою складно організованих масивів даних, далекоглядно довів, що зі зростанням обсягів даних за експериментальних спостережень чи при автоматизованої реєстрації, проведених у різних умовах із різним інструментальним і професійним забезпеченням, розвідувальний аналіз отриманих наборів даних неодмінно має передувати базовому статистичному аналізу, тим самим ставши предтечою майбутньої хвилі великих даних, де цей підхід є основним, має суттєво підсилений й розмаїтим.

Великі дані – характерна риса сучасного світу – супроводжуються власними проблемами, одночасно надаючи можливості для розробки досконалих засобів для їх вирішення задля отримання нових корисних результатів для практики. Наприклад, тепер є можливість прямо досліджувати багатовимірні популяції об'єктів, вимушено не обмежуючись вибірками з них і при тому неодмінно щось втрачаючи, перейти із традиційних табличних форматів до обробки неструктурованих даних як-от сигналів, зображень, медіа чи текстів.

Великі дані, зобов’язані успіхам IT у вигляді автоматизованої реєстрації, передачі, зберігання, накопичення та перетворення даних різного формату і походження, висунули специфічну проблематику щодо їх продуктивного використання, розмовною мовою – mining of data: за буквальним значенням mine (копальня) мова йде про видобування корисної інформації, якою зазвичай цікавляться: розвідники, дослідники, геологи, археологи, слідчі тощо, тобто, «копати, якнайглибше, аби докопатися» за визначеним інтересом.

Згодом, природним розвитком методології розвідувального аналізу даних (Exploratory Data Analysis) на тлі поточного рівня комп'ютерних та мережевих

[1] **Tukey J**. Exploratory Data Analysis. Addison-Wesley, 1977. – 711 p.
Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. Пер. с англ. М.: Мир, 1981. – 694 с.
Berthold M., Hand D. Intelligent Data Analysis, 2-ed. Springer, 2007. – 514 p.
Брюс П., Брюс Э. Разведочный анализ данных. Практическая статистика для специалистов Data Science. Пер с англ. – СПб: БХВ-Петербург, 2018. – 303 с.

інформаційних технологій визначився вдосконалений напрямок аналізу наборів даних великого розміру – Data Mining, тепер це синоніми[2].

Теперішний датамайнінг[3] цілком логічно продовжує продуктивну практику класичного і прикладного статистичного аналізу експериментальних даних, але де отримані за певними спостереженнями/реєстрацією дані потребують попередню підготовку і розвідку, а великі розміри їх масивів – застосовувати новітні підходи і засоби, зокрема, технологію машинного навчання, потужну комп'ютерну техніку, розвинені математичні моделі і методи, швидкі і продуктивні алгоритми.

Завдання датамайнінгу: практична техніка напів- або автоматичного розвідувального аналізу наборів даних: великого обсягу та різного типу/формату для отримання корисної інформації, виявлення раніше невідомих, цікавих і корисних закономірностей «входи-виходи», структур типу виокремлених груп записів (кластерів), аномалій і залежностей (асоціацій) тощо у проблематиці прогнозного моделювання.

Найчастіше, зокрема, в діловій аналітичній практиці, розвідують/«майнять» набори даних у форматі табличних баз даних, відповідно, для цього застосовують середовища табличних процесорів, стандартні і довантажувані інструментальні засоби у них.

За давньою традицією інструментарій обробки даних обмежувався засобами прикладної статистики, що застосовується в експериментальній практиці для невеликих і локальних баз даних, де дані мають відомі, принаймні, призначення, тип і формат. Математична статистика заснована на концепції усереднення за малою вибіркою з популяції великого розміру, що призводить до результатів з дещо фіктивними величинами типу середніх: тривалості життя, житлової площі, вартості товарів/послуг, витрат, доходів населення тощо, які корисні, головним чином, для перевірки заздалегідь сформульованих гіпотез. Апарат статистики продуктивно діє й в датамайнінгу як ланка порівняльного аналізу задля вибору шляхів покращення результату.

Суттєве вдосконалення технологій автоматичної реєстрації, передачі та зберігання даних призвело до появи колосальних потоків даних у різних областях. Тепер діяльність будь-якої організації (комерція, виробництво, медицина, безпека і оборона, соціологія, наука тощо) супроводжується генеруванням, реєстрацією та тривалим зберіганням багатьох подробиць, звідси й знадобилася продуктивна переробка потоків сирих даних з метою отримання раніше невідомих, нетривіальних, практично корисних та доступних інтерпретації відомостей (знань), необхідних для прийняття рішень у різних сферах людської діяльності.

Сучасні властивості і вимоги до такої переробки: дані мають практично необмежений обсяг, вони різнорідні (кількісно, якісно, за типом/форматом, похо-

---

[2] **Myatt G., Johnson W**. Making Sense of Data. A Practical Guide to Exploratory Data Analysis and Data Mining, 2-ed. – Wiley, 2014. – 248 p.

[3] калька з Data Mining з-за відсутності стійкого й однозначного перекладу цього терміну, надбудова Data Mining в тексті позначена DM

дженням), результати мають бути конкретними та зрозумілими, інструментальні засоби мають бути досконалими, з високим рівнем автоматизації аналітичних розрахунків, зі зрозумілим інтерфейсом і простими у використанні.

Порівняння цілей прикладної статистики та датамайнінгу

Статистика	Датамайнінг
Середні показники травматизму для курців та некурців?	Які фактори найкраще передбачають нещасні випадки?
Середні розміри телефонних рахунків існуючих клієнтів у порівнянні з рахунками колишніх клієнтів (відмовилися від послуг компанії)?	Які характеристики відрізняють клієнтів, які, ймовірно, збираються відмовитися від послуг компанії?
Середня величина щоденних покупок за вкраденою/діючою кредитною карткою?	Які схеми покупок притаманні шахрайству з кредитними картками?
Середня сума неповернених кредитів банку? Середній % боржників? Архівування звітів про проведені транзакції.	Характерні властивості боржників? Аналізуючи минулі транзакції, які згодом виявилися шахрайськими, банк виявляє деякі стереотипи такого шахрайства; сегментація клієнтів

Інформаційно-аналітична складова управлінської діяльності

Ключовою перевагою датамайнінгу є можливість здійснювати всебічний поглиблений аналіз повних наборів даних, користуючись спеціальними аналітичними інструментами, де реалізовано кращі моделі, методі і алгоритми, деякі з них, як-от Аналіз головних компонент (PCA), лише зараз увійшли в реальну практику, без обмежень на розмір відповідних задач передбачення.

Для цього у виданні використовується trial-версія професійної надбудови Analytic Solver Data Mining від Frontline Systems Inc. (www.solver.com), спеціально розроблена для здійснення освітньої обчислювальної практики в середовищі Excel (ключове слово Analytic).

Згідно ABOK[4] інтегрований напрям Analytics, у найширшому його розумінні, це процес науково обґрунтованого перетворення даних в інформацію задля її подальшого застосування в прийнятті якнайкращих організаційних рішень (згідно поставлених цілей):

дані → інформація (знання) → рішення.

За досягнутим світовим рівнем наукових досліджень та накопиченого в інституті INFORMS наукового/практичного досвіду Analytics – композиція двох специфічних складових:

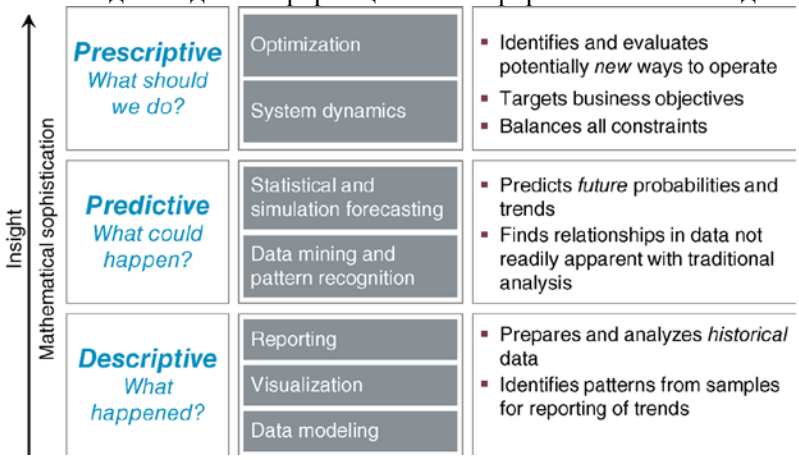
- Data-Centric Analytics: описова і розвідувальна (descriptive, explanatory) та передбачувальна (predictive) аналітика

[4] **INFORMS**. Analytics Body of Knowledge (ed. J. Cochran)/Wiley, 2019. – 386 p.

- Decision-Centric Analytics: приписна, зорієнтована на пошук рішень оптимізаційна та імітаційна (prescriptive, optimization, simulation) аналітика,

що утворюють фундаментальні напрямки Data Analytics та Decision Analysis в сучасній комплексній дисципліні Business Analytics[5].

Якщо напрямок Decision Analysis має порівняно тривалу історію розвитку і досягнень, зокрема, утворення інституту INFORMS (Institute for Operations Research and Management Science), то Data Analytics – новітній «продукт» 1990-их, результат інтенсивного розвитку комп'ютерних та мережевих технологій, обидві, на сьогодні – єдина інформаційна платформа аналітичних досліджень:



Аналітика та інформаційно-аналітична діяльність в цілому – це процеси використання обчислювальних методів для виявлення та практичного застосування впливових закономірностей в наявних даних.

Дані є мірою накопиченої історичного досвіду, тож за визначенням, Аналітика використовує і досліджує історичні дані. Сам термін Analytics виник і набув популярності із 2005-их рр., значною мірою завдяки масштабному запровадженню платформи Google Analytics, хоча ідеї аналітики як виду дослідницької/розвідувальної діяльності цілком природні, у різні часи представлені по різному, як-от: статистика, кібернетика, аналіз даних, нейронні мережі, розпізнавання образів, датамайнінг тощо, тепер узагальнені теперішньою наукою про дані (Data Science).

Зростання ролі аналітики та аналітичної діяльності в останні роки є суто прагматичним: оскільки організації реєструють все більше даних, їх зберігають й

[5] **Ragsdale C**. Spreadsheet Modeling and Decision Analysis. A Practical Introduction to Business Analytics, 9-ed. Cengage, 2022. – 908 p.
Camm J., Cochran J. Business Analytics, 4-ed. Cengage, 2021. – 882 p.
Evans J. Business Analytics, 3-ed. Pearson, 2021. – 705 p.
Albright C., Winston W. Business Analytics. Data Analysis and Decision Making, 7-ed. Cengage, 2020. – 914 p.
Powell S., Baker K. Business Analytics. The Art of Modeling with Spreadsheets, 5-ed. Wiley, 2017. – 555 p.

узагальнюють, то цілком природно цей цінний ресурс, на отримання якого витрачено немалі кошти, треба продуктивно використовувати задля покращення оцінок, прогнозів, рішень і, зрештою, ефективності у будь-чому.

Датамайнінг якнайкраще використовує прогнозну аналітику та прогнозне моделювання на її основі задля виявлення цікавих та значущих закономірностей, прихованих/присутніх в неозорих масивах даних[6]. Результатом отриманого передбачення є зважена оцінка значення певної цільової змінної, наприклад, наскільки достовірним є клас, до якого в наборі даних віднесено пацієнта чи клієнта банку.

Чим прогнозна аналітика відрізняється від інших видів аналітики?

По-перше, прогнозна аналітика data-driven, керована даними – алгоритми отримують ключову характеристику шуканих моделей із самих даних, а не з теоретичних припущень, ці сформовані алгоритми формують в залежності від наявних даних щодо конкретних об'єктів: параметри моделі, вагові коефіцієнти змінних тощо, нарешті – саме тут визначається складність моделі, звідки висувається непроста проблема – як її реалізувати.

По-друге, її алгоритми автоматизують процес пошуку бажаного знову саме із даних, розраховуючи не лише на обчислення, скажімо, коефіцієнтів змінних у складі моделей чи статистичних оцінок, а й визначаючи форму моделей. Так, зокрема, із усіх змінних/властивостей визначається група «кращих» змінних чи властивостей, найсильніше впливаючих на цільову змінну, чим визначаючи шукану модель і результат.

І, нарешті, у цьому класі алгоритмів і моделей є ще одне суто допоміжне, але нове і важливе завдання – автоматизувати процес перетворення великих обсягів «сирих»/«брудних»/пошкоджених значень вхідних даних для їх подальшого коректного використання відповідними і вимогливих до входів інструментальними засобами.

Алгоритми прогнозного моделювання, базуючись на принципах машинного навчання, поділяють на дві групи згідно поставлених задач і властивостей наборів даних, коротко, це навчатися будувати модель:

- «з учителем», на прикладах з відповідями, коли в наборі даних є значення цільової змінної-«учителя», та
- «без учителя», без заданих значень цільової змінної, розраховуючи на виявленні «схожості»/«близькості» наче однакових на вигляд записів у наборі даних.

Типові задачі прогнозного моделювання «з учителем»: кластеризація, класифікація із числовою/категоріальною цільовою змінною та множинна регресія для цілих/неперервних цільових змінних.

Навчити алгоритм визначати клас чи показник без такої змінної, «без учителя», набагато важче і непросто досягти пристойну точність результату, бо ж його навіть нема з чим порівняти для оцінки. Іноді, як не дивно, виручають великі розміри масивів даних, де можна відстежити типову поведінку «маси» (як-от

[6] **Abbott D**. Applied Predictive Analytics. Principles and techniques for the professional data analyst, 2-ed. – Wiley, 2014. – 453 p.
Larose D., Larose C. Data Mining and Predictive Analytics. Wiley, 2015. – 827 p.

правила асоціації для аналізу кошика покупця) чи отримати її керованим перебором комбінацій за допомогою нейронній мережі і надалі формалізувати колективні рішення у клас сценаріїв «Що-якщо?».

Світовий досвід і практика

Прогнозне моделювання – жива практика, адже люди усюди і завжди думають про завтрашній день, звідси перед ними вічна проблема – передбачення (prediction, forecasting) майбутнього.

Мовою сучасної дата-аналітики сформований прогноз – конкретна відповідь на запит «Що-якщо?» у вигляді корисної інформації для прийняття подальших рішень, знайденої власними можливостями і зусиллями та/чи певними спеціальними засобами, зокрема, досконалим аналізом даних – датамайнінгом.

Процес і методологія Data Mining – сучасний етап неперервного розвитку наукового передбачення, прогнозної аналітики і моделювання у дослідницькій практиці, що базується на накопиченому досвіді математичного, оптимізаційного і статистичного моделювання та новітній IT-платформі.

Свіжий факт: наприкінці 2020 р. INFORMS повідомив про випуск нового журналу Journal on Data Science в області комп'ютерних наук та практик, щоби заповнити критичну прогалину для дослідників, викладачів та менеджерів шляхом публікації ефективних аналітичних методологій науки про дані, щоби розвиток цієї науки вдосконалював процеси формування організаційних рішень у бізнесі, менеджменті та промисловості засобами бізнес-аналітики.

Головний редактор – Galit Shmueli (Distinguished Professor of Business Analytics, National Tsing Hua University, Taiwan), авторка перших публікацій про датамайнінг в Excel з використанням надбудови XLMiner[7].

Складові журналу:

[7] **Shmueli G**. Data Mining in Excel: Lecture Notes and Cases/ 2005. – 270 p.
Shmueli G. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, 2-ed. Wiley, 2010. – 726 p.

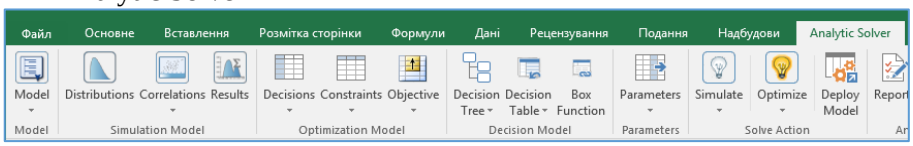
- Data: real-world or simulated
- Models/algorithms: innovative data science methodology (model/algorithm/ approach)
- Managerial/industrial relevance: decision-making motivation and potential/ actual impact
- Implications: consideration of relevant practical implications.

Орієнтація напряму Data Analytics на прийняття управлінських рішень відповідає міждисциплінарному характеру науки про дані, фундаментом якої є математичне моделювання, статистика, машинне навчання, дослідження операцій, системна інженерія, економетрія та інформаційні технології.

Реалізація прогнозних моделей

Програмна платформа ASP (Analytic Solver Platform, www.solver.com) – узгоджений комплект аналітичних інструментів-розв'язувачів (solvers), з-за їх досить великої кількості розділений і розташований у двох вкладках вікна Excel:

- Analytic Solver[8] та



- Data Mining[9]

, які логічно відповідають двом напрямкам, Decision Analysis та Data Analytics.

Історія

Analytic Solver – флагманська розробка лінії надбудов Excel компанії Frontline Systems, розпочатої у 1991 р., на зараз її складові: Analytic Solver Optimization, Analytic Solver Simulation та Analytic Solver Data Mining[10].

Найвідоміша розробка у світі – стандартна надбудова Excel Solver (укр. Розв'язувач), розроблена для щойно створеної ОС MS Windows на замовлення Microsoft для освітніх цілей, завдяки якій у світі суттєво зріс практичний рівень оптимізаційного моделювання, за що у 2010 р. компанія отримала премію INFORMS Impact Prize (за впливовість на дослідження і освіту), тим самим забезпечивши технологічну ланку Decision Analysis.

Лідер компанії Daniel Filstra – піонер комп'ютерної історії, відомий потужною пропагандою ефективності і активним впровадженням перших електронних таблиць (spreadsheets), президент компанії Frontline Systems, заснованої ним у 1988 р.

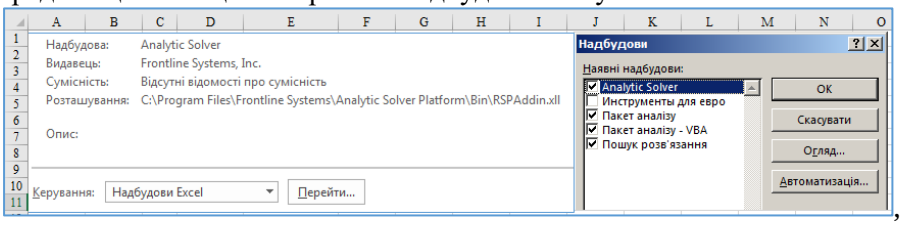
[8] **Frontline Solvers**. For Use With Excel 2013-2019. Analytic Solver Optimization and Simulation User Guide, v. 2023. – 667 p.

[9] **Frontline Solvers**. For Use With Excel 2013-2019. Analytic Solver Data Mining User Guide, v. 2023. – 273 p.

[10] **Кузьмичов А. І.** Оптимізаційне моделювання в Excel. ІПРІ НАНУ, 2017. – 438 с.

У 2016 р. задля реалізації недостатньої ланки Data Analytics в триєдиній структурі Аналітики компанія придбала права на використання надбудови XLMiner для датамайнінга в Excel, орієнтованим на суто освітні цілі. Після повної її переробки і наповнення новими інструментами, підтриманими власними алгоритмами оптимізації та симуляції, із 2019 р. діє потужна надбудова Analytic Solver Data Mining у складі платформи ASP та в її освітній версії ASPE.

В середовищі Excel це інтегрована надбудова Analytic Solver:

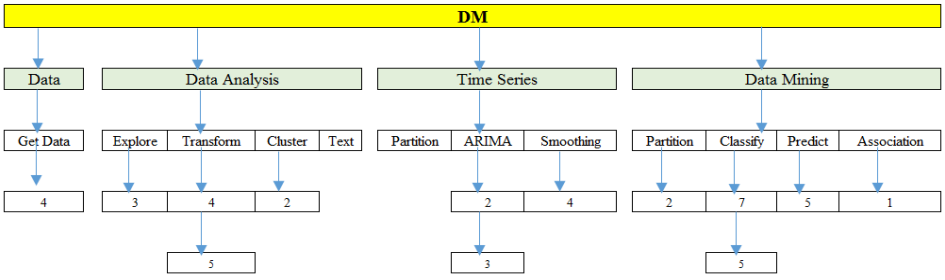


, яка після встановлення робить Excel потужним і досконалим аналітичним засобом для ділової та освітньої практики електронно-табличного моделювання та прийняття рішень.

ASP регулярно вдосконалюється у вигляді випуску оновлених версій, якнайкраще реалізуючи, зокрема, в надбудові DM базові процедури підготовки та перетворення значень наборів даних, кластеризації, класифікації і прогнозування, аналізу часових рядів та асоціації. Оперативно враховуються й реалізуються сучасні тенденції в дата- та бізнес-аналітиці й інформатиці: багатоядерні процесори, розподілені і хмарні обчислення, бази даних, «великі дані», мови програмування тощо.

Надбудова DM у її складі містить декілька десятків різних інструментів – програмних модулів, де кожен якнайшвидше й якісно реалізує відповідний метод/алгоритм аналізу набору даних, а також регулярно доповнюваний до стандартних список функцій для спрощення і прискорення відповідних процедур.

Інструменти надбудови DM об'єднані в 4 групи та проблемні секції за типовими класами задач (в них кількість основних інструментів, без версій):



Безкоштовна education-версія (ASPE, програмний продукт Analytic Solver Basic) платформи відрізняється від повної комерційної («промислової») версії ASP (вартістю $5000 ÷ $6250 за різних ліцензій) лише обмеженнями на розміри наборів даних та параметри отриманої інформації, як-от кількості кластерів чи класів (див. Додаток 1), після інсталяції прив’язана до е-адреси замовника і синхронно діє на його кількох пристроях, розрядність ЦП 32/64, Web/Win/Mac.

НАБОРИ ДАНІ: ОТРИМАННЯ, ПІДГОТОВКА І АНАЛІЗ

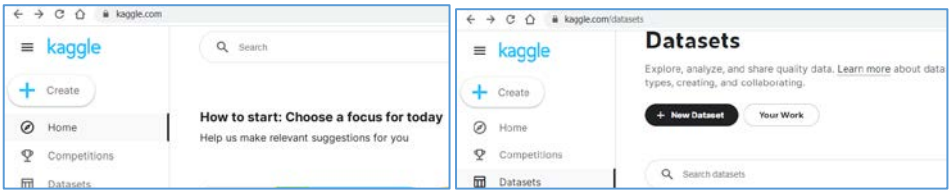
Видобуток з даних корисної інформації, цінних відомостей про об'єкти, представленими цими даними, вимагає навичок, які за традиційною освітою і практикою не отримували в економіці, статистиці чи в інженерній справі, де зазвичай проводять певні експерименти на моделях.
Йдеться про допоміжний, але необхідний етап датамайнінгу – попередню підготовку набору даних, отриманого звідкись, від розуміння із застосуванням стандартних засобів Excel аж до наступних основних кроків, де використовуються інструментальні аналітичні засоби отримання інформації та її узагальнення для формування знань. Адже у традиційній дослідницькій практиці цей етап немає сенсу – експериментатор, як-от агроном, фізик чи конструктор космічного апарату, завчасно знають про формат, значення та смисл даних, які очікується отримати для майбутньої роботи.
Зазвичай зрозуміти набір даних просто, якщо він видимий (на папері чи екрані, ними супроводжують навчальну літературу) і зразу видно його вади (пропуски, некоректні значення чи формат), але для великого за обсягом набору даних із тисяч записів і десятків полів/змінних це не так, навіть швидка прокрутка таблиці не допоможе.
Тому процедура підготовки, відсутня у літературі із класичної та бізнес-статистики, що має справу з «чистими» і невеликими за розмірами масивами даних, неодмінна складова в сучасних книгах із розвідувального аналізу даних (Intelligent Data Analysis, Exploratory Data Analysis) у вигляді процедур: Data Preprocessing, Cleaning, Exploration, Preparation тощо.
Апарат та процедури датамайнінгу поєднують суворі норми і правила, строга дисципліна та кваліфікація персоналу, вміле володіння різними інструментами та технологіями: математико-статистичним аналізом, машинним навчанням, розподіленими базами даних та хмарними сховищами. Для цього видання ще: впевнене користування засобами Excel, у його україномовній версії – оригінальним (англомовним) представленням функцій.
Безпідставно вважається, що необхідні, відкриті і якісно організовані набори даних є будь-де, варто лише знайти і отримати, а їх майнити будуть досконалі інструментальні засоби. Але за практикою виявляється, що все не так просто, «хороших» (сформованих за нормами табличних баз даних, як-от приклади на сайті ООН, www.un.org) наборів мало або вони недоступні. Зате ті, що можна нашвидкуруч скачати, вимагають здійснити саме цю – непросту і нешвидку – передобробку отриманих «сирих»/«брудних» даних.

Отримання наборів даних

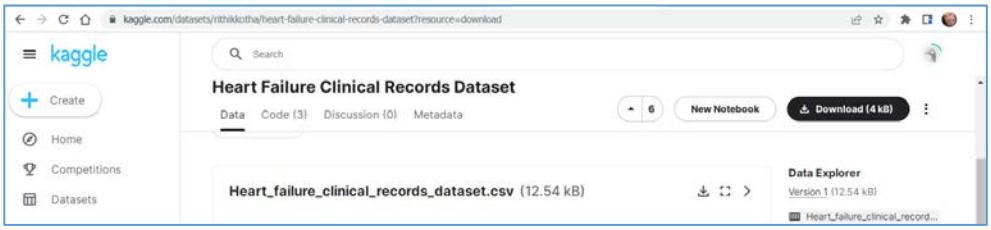
Першочергова задача дата-аналітика – знайти достойне джерело, вибрати набір даних порівняно великого розміру, отримати відповідні зовнішні характеристики і доступ, для подальшої роботи скачати і зберегти оригінал.
З-за використання середовища Excel цей набір має табличну форму, у зовнішніх джерелах зазвичай представлений у компактному текстовому форматі csv.

Приклад.
Отримання набору даних з kaggle.com[11]

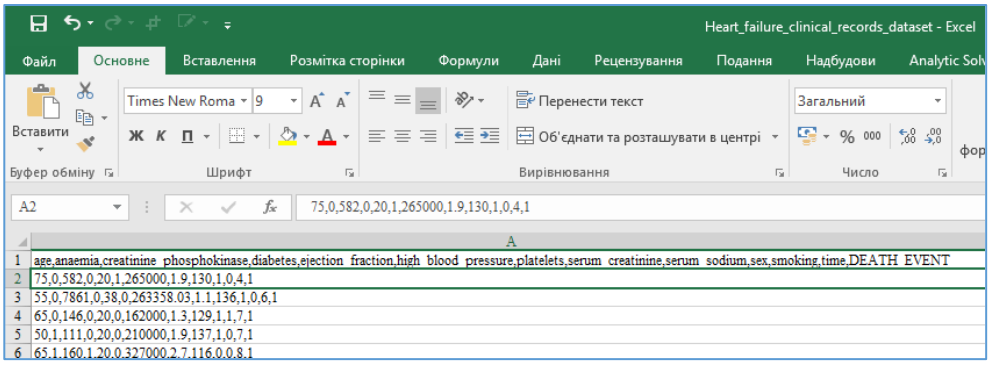
Крок 1: вхід



Крок 2: завантаження



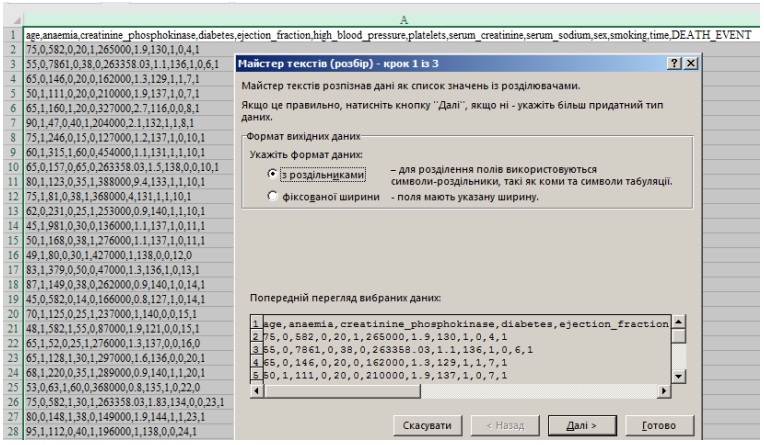
Після реєстрації на сайті потрібний набір даних можна скачати як архів (4 кВ) чи csv-файл (12,54 кВ), після відкриття записи компактно розташовані у стовпці A, в клітинці A1 показані імена змінних, їх значення розділені комою згідно табличного формату csv:



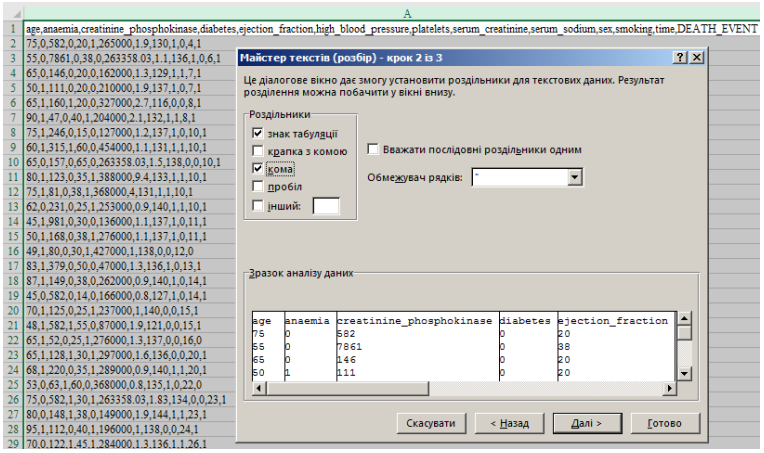
Перетворення форматів csv → xlsx
Команда: Дані → Текст за стовпцями

[11] Kaggle – платформа з багатьма можливостями: конкурси з датамайнінгу та машинного навчання; готові набори даних (datasets) для самостійного вивчення та побудови моделей; різноманітні професійні дискусії з різної тематики та рівнів учасників; курси для початківців (за основами науки про дані, мов програмування Python та R).

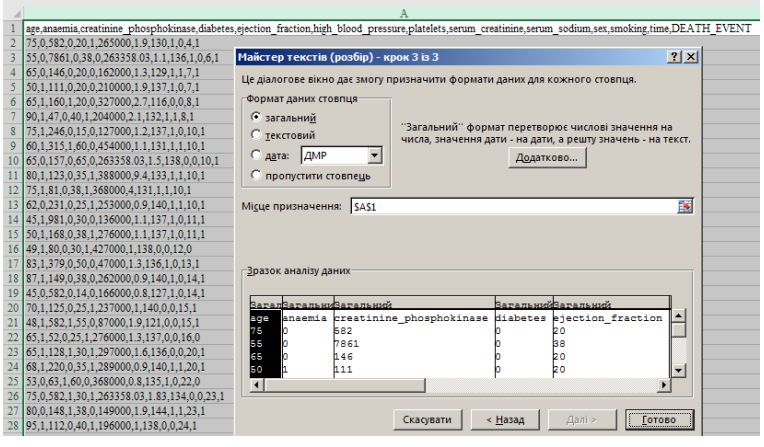
Крок 1



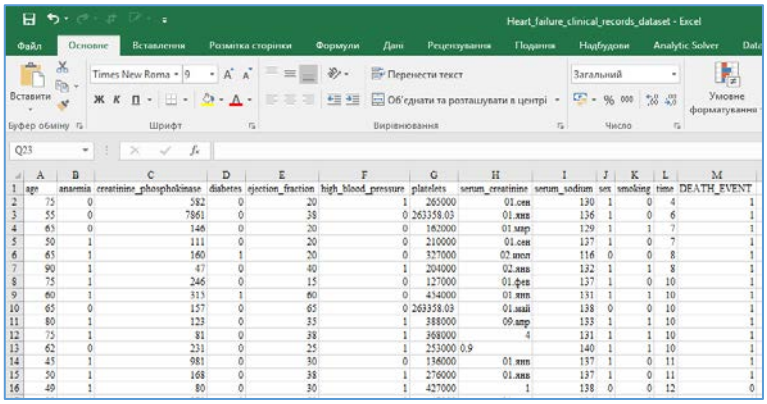
Крок 2



Крок 3



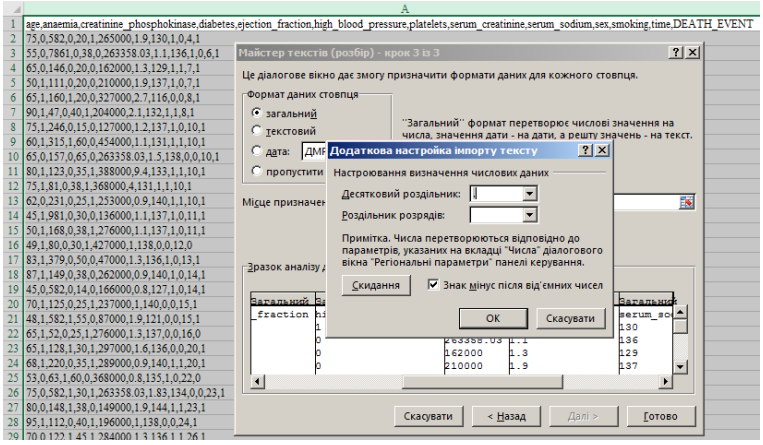
Результат



age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	01.сен	130	1	0	4	1
55	0	7861	0	38	0	263358.03	01.янв	136	1	0	6	1
65	0	146	0	20	0	162000	01.мар	129	1	1	7	1
50	1	111	0	20	0	210000	01.сен	137	1	0	7	1
65	1	160	1	20	0	327000	02.июл	116	0	0	8	1
90	1	47	0	40	1	204000	02.янв	132	1	1	8	1
75	1	246	0	15	0	127000	01.фев	137	1	0	10	1
60	1	315	1	60	0	454000	01.янв	131	1	1	10	1
65	0	157	0	65	0	263358.03	01.май	138	0	0	10	1
80	1	123	0	35	1	388000	09.апр	133	1	1	10	1
75	1	81	0	38	1	368000	4	131	1	1	10	1
62	0	231	0	25	1	253000	0.9	140	1	1	10	1
45	1	981	0	30	0	136000	01.янв	137	1	0	11	1
50	1	168	0	38	1	276000	01.янв	137	1	0	11	1
49	1	80	0	30	1	427000	1	138	0	0	12	0

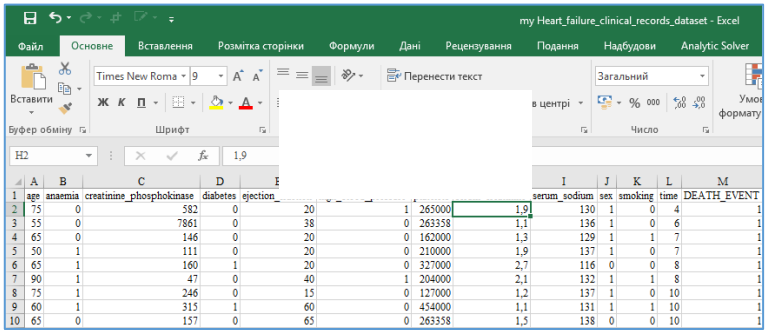
Увага! В стовпці H числа з десятковою точкою (що превалює в англомовних документах) подані помилково як дата, стандартним засобом Формат клітинок … виправити непросто.

Тож треба повернутися на Крок 3 і для десяткових чисел в оригіналі змінити роздільник цілої і дробової частин з точки на кому:



Результат

Файл (буде використаний нижче, див. інструмент Find Best, класифікація)



age	anaemia	creatinine_phosphokinase	diabetes	ejection_			serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1,9	130	1	0	4	1
55	0	7861	0	38	0	263358	1,1	136	1	0	6	1
65	0	146	0	20	0	162000	1,3	129	1	1	7	1
50	1	111	0	20	0	210000	1,9	137	1	0	7	1
65	1	160	1	20	0	327000	2,7	116	0	0	8	1
90	1	47	0	40	1	204000	2,1	132	1	1	8	1
75	1	246	0	15	0	127000	1,2	137	1	0	10	1
60	1	315	1	60	0	454000	1,1	131	1	1	10	1
65	0	157	0	65	0	263358	1,5	138	0	0	10	1