УДК 004.622(623)

Захарова К. В., студентка
Нескородєва Т. В., к.т.н., доцент, завідувач кафедри інформаційних технологій

КЛАСИФІКАЦІЯ НАЯВНОСТІ РАКУ ГРУДЕЙ

Донецький національний університет імені Василя Стуса, м. Вінниця

Мета роботи показати один з векторів аналізу даних, використовуючи мову програмування R. Досліджуючи кожну змінну: її значення, походження, кореляцію та вплив на інші змінні; обираючи одну модель при різних результатах точності, можна дійти до різних висновків. Оскільки тема датасету для аналізу – тема здоров'я, а власне найстрашнішої хвороби XXI століття – раку, необхідно бути дуже обачливим і з скрупульозністю приймати кожне рішення.

У роботі було використано «Wisconsin Breast Cancer Data» (дані про рак грудей у штаті Вісконсин) датасет [1].

Першим кроком іде підготовка датасету для роботи. Спочатку необхідно завантажити .csv файл з даними. Далі видалити ті рядки, де відсутні будь-які дані. Для зручності можна перейменувати цільові змінні.

Після підготовки, перевіривши датасет, можна побачити, що набір даних трохи не збалансований. Це може бути проблемою, оскільки перевага на стороні «Доброякісна» (пухлина), та й більше ніж в 1.5 рази. Така проблема стосується лише даного датасету (та схожих йому за призначенням). Рак – це коли краще перевіритись зайві 100 разів, аніж заспокоїтись одразу, тому для моделі краще невірно передбачити злоякісну пухлину, ніж доброякісну пухлину та змусити пацієнта відчувати ілюзію того, що все добре.

Вивід матриць кореляцій доповість, що по-перше дуже багато кореляцій, по-друге кореляції між типами змінних (mean, se, worse) хоч і залишаються приналежними до одних функцій, але все ж таки різниця візуально помітна. Також, у висновку, можна сказати, що датасет має «багатоколінеарність» між змінними. Додавши до матриць розсіювання цільові змінні, можна побачити, що кластеризування буде не найпростішим – дані змішані, концентруються в одній і тій самій точці майже для усіх змінних.

Занадто велика кількість змінних може викликати такі проблеми, наведені нижче [2]:

- Підвищена пропускна здатність комп'ютера;
- Занадто складні проблеми візуалізації;
- Знижує ефективність, включивши змінні, які не впливають на аналіз;
- Ускладнює інтерпретацію даних;

Рішення: використання одного основного компоненту для розробки моделі, зменшивши кількість змінних з високою кореляцією. При визначенні кількості головних компонентів використовується кумулятивну частку та використовується діаграма отримана вище.

PCA (Principal Component Analysis) – аналіз основних компонентів використовує стандартизовані дані, щоб уникнути спотворення даних, викликаного різницею в масштабі [3]. У результатах PCA, якщо кумулятивна частка становить 85% або вище, то її можна визначити за кількістю основних компонентів. Наприклад, якщо кумулятивна частка PC4 дорівнює 88,7, це означає, що сума часток PC1~PC4 дорівнює 88,7. В обраному датасеті, сумарна частка від PC1 до PC6 становить приблизно 88,7%. (понад 85%). Це означає, що PC1~PC6 може пояснити 88,7% усіх даних. Таким чином відфільтровуються непотрібні змінні, які будуть заторможувати роботу та створювати «шум» в даних.

Далі, у роботі, створюються декілька моделей: C5.0, C5.0 – Tune, rpart, Prune, OneR, JRip, naiveBayes, randomForest, ctree, KNN – Tune, K-Means, GBM, adaBoost, SVM, SVM – Tune.

Візуалізовані результати показують кількість “відповідей” відповідно по квадратам: TruePositive (TP), FalsePositive (FP), FalseNegative (FN), TrueNegative (TN) [4]; та загальний відсоток точності. Найкраще себе показав алгоритм SVM – Tune з точністю у 99%. Далі йдуть SVM та randomForest з 98%.

Інший метод класифікації діагнозу – графік функції щільності ймовірності. Створюються гістограми для кожної змінної, з розділенням по діагнозам, далі вводяться дані пацієнта. Якщо дані не перетинають медіану діагнозу «злоякісна», то існує висока ймовірність, що пухлина «доброякісна».

Отже, у роботі було розглянуто декілька способів класифікації раку грудей. Проаналізовано змінні та використано висновки аналізу для побудови високоточної моделі машинного навчання.

Список літератури

1. Wisconsin Breast Cancer Data. URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) (дата звернення: 04.12.2021)
2. Too much covariates in a multivariable model may cause the problem of overfitting. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4178069/ (дата звернення: 05.12.2021)
3. Principal Component Analysis. URL: https://en.wikipedia.org/wiki/Principal_component_analysis (дата звернення: 05.12.2021)
4. Оценка качества в задачах классификации. URL: http://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B0_%D0%B2_%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%B0%D1%85_%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D0%B8 (дата звернення: 05.12.2021)

5. Probability density function. URL: https://en.wikipedia.org/wiki/Probability_density_function (дата звернення: 05.12.2021)

УДК 004.82

Цінський С.В., студент 1 курсу
магістратури спеціальності 121
«Інженерія програмного забезпечення»
Київська К.І., к.т.н., доцент, доцент
кафедри інформаційних технологій

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В БУДІВНИЦТВІ

Київський національний університет будівництва і архітектури

Сучасна будівельна галузь вже оперує більше ніж десятком технологій, максимально необхідними в будівництві. Розглянемо найпрогресивніші та найнеобхідніші IT-технології та інноваційні матеріали в будівництві, які з кожним роком все більше інтегруються в будівельну сферу, реалізуючи самі сміливі ідеї майбутнього.

Зростання міст і кількості населення, а також новий формат рівня людських комунікацій в епоху BIG DATA, зростання економіки та добробуту людей, активізувало будівельну галузь на більш динамічну інтеграцію інновацій та технологічних рішень. Тому нові технології в будівництві в світі активно просуваються та використовуються.

До того ж, сама швидкість розвитку технологій веде до масштабного оцифрування будівельної галузі. І питання застосування IT-технологій - це вже питання конкурентоспроможності. Інновації в будівництві видозмінюють будівельний майданчик та збільшують прибуток, а також допомагають вигравати проектні тендери.

Оскільки саме інновації приносять економічну вигоду та підвищують конкурентоспроможність конкретної будівельної компанії, а також в кінцевому підсумку реалізують запит клієнта з максимальною ефективністю.

BIM-технології - (від англ. Building information modeling) стають основою сучасного проектування та основною технологією, яку планується застосовувати для будівництва об'єктів. BIM-технологія – це не просто віртуальне моделювання будівлі, це комплексне уявлення в цифровому вигляді фізичних і функціональних характеристик об'єкта [1]. BIM -технологія враховує не просто зведення, а й оснащення, управління, експлуатацію об'єкта, перспективу ремонту або знесення, тобто охоплює весь життєвий цикл об'єкта в комплексі. Всі складові та нюанси в проектуванні, які мають відношення до об'єкта, обов'язково враховуються і розглядаються в єдиному проекті. При видаленні або заміні