https://www.sworldjournal.com/index.php/swj/article/view/swj11-02-045
DOI: 10.30888/2663-5712.2022-11-02-045

УДК 004.9

COMPARISON OF MACHINE LEARNING ALGORITHMS FOR PREDICTING MORTALITY FROM COVID-19 VIRUS
ПОРІВНЯННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ СМЕРТНОСТІ ВІД ВІРУСУ COVID-19

Doroshenko I.V. / Дорошенко І.В.
c. p.-m.s., as.prof. / к. ф.-м.н., доц.
ORCID: 0000-0001-8729-1768
Knihnitska T.V. / Кнігніцька Т.В.
postgraduate / аспірантка
ORCID: 0000-0003-4614-5945
Deretorska T.I/ / Дереторська Т.І.
magistr / магістр
Chernivtsi National University, Chernivtsi, Kotsyubynskoho 2, 58012
Чернівецький національний університет, Чернівці, вул.Коцюбинського 2, 58012

Анотація. *У статті проведено порівняння ефективності методів машинного навчання на прикладі даних поширення Covid-19. Розглянуто наступні алгоритми машинного навчання: логістична регресія (Logistic Regression), метод опорних векторів (Support Vector Machine), випадковий ліс (Random Forest), метод K-середніх (K-Means). Для порівняння моделей використано акуратності побудованих моделей. Всі статистичні дослідження у роботі проведено мовою R Programming.*

Ключові слова: *методи машинного навчання, випадковий ліс, метод опорних векторів, логістична регресія, метод K-середніх.*

Вступ.

Криза в сфері охорони здоров'я набула величезного масштабу внаслідок світової пандемії Covid-19. Ситуація в світі змушує людство поміркувати над тим, що можна зробити, щоб зробити людей більш підготовленими та захищеними від аналогічних ситуацій в майбутньому. Протягом 2 останніх років пандемії однією з основних проблем, з якою стикаються медичні працівники, є дефіцит медичних ресурсів та їх ефективний розподіл. Відповіді на багато питань можна одержати шляхом залучення методів машинного навчання до прогнозування поведінки кривої інфікування вірусом Covid-19. Таким чином стало можливим допомогти лікарям та керівництву держав у організації правильної стратегії боротьби проти поширення пандемії Covid-19, тобто стало можливим врятувати більше людських життів. Основною метою даного дослідження є порівняння методів машинного навчання з та без учителя на прикладі конкретних даних. Дані про поширення Covid-19 вибрано виключно із цікавості до даного напрямку досліджень у наш час.

1. Опис вхідних даних

Дані, використані для побудови методів машинного навчання у даній роботі, отримано з платформи Kaggle (https://www.kaggle.com/). Kaggle – система організації конкурсів з дослідження даних, а також соціальна мережа фахівців з обробки даних та машинного навчання, яка належить корпорації Google з березня 2017 року. Дані https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset?select=covid.csv являють собою інформацію про

пацієнта, історію різних захворювань та звички пацієнта. База даних містить інформацію про 563201 пацієнта та 23 атрибути (стовпці) для кожного з них. У даному дослідженні використано 20000 записів та 10 наступних атрибутів:

Таблиця 1 - Дані

Sex	Стать пацієнта	0 – чоловік, 1 – жінка
Died	Помер/не помер пацієнт	0 – NO, 1 – YES
Pneumonia	Пневмонія	0 – NO, 1 – YES
Age	Вік	
Diabetes	Діабет	0 – NO, 1 – YES
Asthma	Астма	0 – NO, 1 – YES
hypertension	Гіпертонія	0 – NO, 1 – YES
other disease	Інші захворювання	0 – NO, 1 – YES
cardiovascular	Серцево-судинні захворювання	0 – NO, 1 – YES
Tobacco	Куріння	0 – NO, 1 – YES

2. Попередня обробка даних

Для проведення статистичного аналізу та побудови ML моделей для початку необхідно провести попередню обробку даних. Тобто майже всі змінні є бінарними. Необхідно їх відповідно перекодувати, де 0 – означає відсутність захворювання або звички, а 1 – позначає наявність захворювання або шкідливої звички. Крім того дані містять багато пропущених значень, які позначено «97, 98, 99, 9999-99-99». Усі ці значення потрібно видалити, так як вони можуть значно спотворити результати аналізу.

3. Критерії порівняння алгоритмів ML

У даній роботі для порівняння побудованих моделей машинного навчання використано *Accuracy* або точність моделі машинного навчання – це вимірювання, яке використовується для визначення, яка модель найкраще визначає зв'язки та закономірності між змінними в наборі даних на основі вхідних або навчальних даних (тренувальних). Чим краще модель може узагальнити «невидимі» дані, тим кращі прогнози та ідеї вона може дати, що, у свою чергу, принесе більшу цінність.

4. Опис результатів

Для початку всі змінні, окрім Age, приведемо до факторного типу. Після видалення пропущених значень та вибору 20000 пацієнтів для аналізу маємо основний показник – 12,5% пацієнтів, які потрапили до лікарні, не вижили після захворювання Covid-19.

4.1. Результати моделі логістичної регресії (LR)

Поділимо дані на навчальну (60%) та тестову (40%) вибірки та побудуємо логістичну регресійну модель спочатку для тренувальної вибірки. Отже, згідно з результатами моделі логістичної регресії можна зробити наступні висновки:

- Найбільший позитивний вплив на смерть пацієнта мають наявність діабету, пневмонії та вік пацієнта. Чим більшим є вік пацієнта – тим більшою є ймовірність смерті пацієнта від Covid-19.
- Пацієнти жінки мають меншу ймовірність смертності від Covid-19. Стать пацієнта, наявність діабету, пневмонії та вік пацієнта є найбільш значущими показниками, які впливають на змінну died.

- Наявність гіпертонії та інших захворювань є менш значущими змінними за попередні. Тобто ці показники мають менший вплив на ймовірність смертності пацієнта.
- Цікавим висновком є ще і той факт, що астма у пацієнта та його куріння не мають впливу на змінну died. Тобто курці можуть бути спокійними, ця шкідлива звичка не наражає їх на більшу небезпеку.
- Застосувавши побудовану модель логістичної регресії до тестової вибірки, бачимо, що AIC показник зменшився майже на 2000, що є показником адекватності моделі.

За допомогою функції predict спрогнозуємо ймовірність летальності для пацієнта жінки 80 років з усіма наявними захворюваннями:

```
predict(logmodel, data.frame(sex = "1", diabetes = "1", tobacco = "1",
pneumonia = "1", age = c(80), asthma = "1", other_disease = "1", hypertension =
"1", cardiovascular = "1"), type = "response")
1
0.6538167
```

Отже, ймовірність смерті такої пацієнтки – 65%. А ймовірність летальності для 80 річної пацієнтки без жодного захворювання – 9%. Для 30 річної пацієнтки без жодного захворювання – ймовірність летальності становить 0,9%.

Таким чином, отримавши дану логістичну модель можна прогнозувати перебіг захворювання для пацієнтів різних вікових груп з різними захворюваннями. Таким чином можна встановити, до яких наслідків може призвести вірус Covid-19, якщо не приймати заходів вакцинації та ізоляції населення. Такий дослідницький підхід якраз і дозволяє заглянути наперед, знаючи відсоток населення з різними супутніми захворюваннями, та встановити невідворотні наслідки завчасно.

Точність або акуратність побудованої моделі становить 89,68%. Тобто дисперсія залежної змінної died на 89,68% пояснюється незалежними змінними. Такий результат акуратності моделі на практиці є досить високим.

4.2. Результати методу опорних векторів (SVM)

Аналогічно до моделі логістичної регресії, алгоритм машинного навчання SVM є класифікатором [1]. Тобто наше завдання полягає в тому, щоб побудована модель класифікації максимально точно класифікувала вхідні дані. Завдяки моделі логістичної регресії ми уже зробили висновки про те, що найнебезпечнішими показниками для пацієнта з вірусом Covid-19 є наявність пневмонії, діабету та поважний вік пацієнта. Отже, не потрібно довго лікуватися вдома від Covid-19, так як доведення захворювання до розвитку пневмонії збільшує ймовірність летальності у 9,7 разів. Наявність діабету у пацієнта збільшує ймовірність летальності у 1,6 разів, а наявність астми – у 0,75 разів.

Переходимо до навчання моделі SVM. Скористаємося методом trainControl(). Цей метод дозволить контролювати всі обчислення. Класифікатор поділив тестову вибірку на 2 групи – 0 та 1 за змінною died. Тобто 0 – пацієнт з певними захворюваннями не помре, 1 – пацієнт помре. Точність побудованої моделі SVM становить 88.9%.

4.3. Результати випадкового лісу (Random Forest)

Алгоритм RF – це алгоритм ML з учителем для класифікації та регресії. Як видно з назви, цей алгоритм випадковим чином створює ліс із кількома деревами (Decision Trees, DT). Аналогічно, у класифікаторі RF, чим більше дерев у лісі, тим більша точність результатів [2]. Простими словами, RF будує кілька DT (так званий ліс) і склеює їх разом, щоб отримати більш точний прогноз. Ліс, який він будує, являє собою набір дерев рішень.

Крос-валідація або перехресна перевірка використовується для оцінки ефективності моделі за допомогою навчальних даних. Починаємо з випадкового поділу даних навчання на 5 частин однакового розміру, які називаються «folds». Далі модель навчається на 4/5 даних і перевіряється її точність на 1/5 даних, які було пропущено. Потім цей процес повторюється з кожним розділенням даних. Зрештою, відсоток точності для п’яти різних розділів даних усереднюється, щоб отримати середню точність. Бібліотека Caret якраз і виконує описаний алгоритм. Caret автоматично вибирає значення гіперпараметра «mtry», яке є найбільш точним при перехресній перевірці. На виході з mtry = 2 середня точність становить 0,8921117, або приблизно 89.2%. Це найвище значення, тому Caret вибирає це значення.

4.4. Результати методу k-середніх (k-means)

Алгоритм кластеризації *k*-means наступний:

1. Завантажити дані та опрацювати їх (привести до відповідного типу, видалити пропущені значення).
2. Обираємо 20000 даних для дослідження та стовпці 2 і 4 (died, age) для проведення кластеризації даних смертності по віку від Covid-19.
3. Визначаємо кількість кластерів за допомогою методу ліктя.

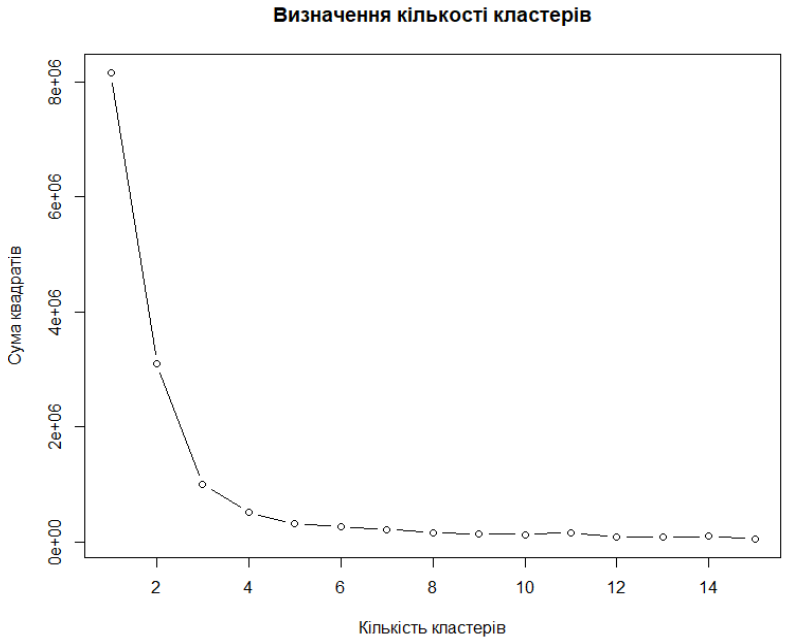

Рисунок 1 - Визначення кількості кластерів

Для визначення кількості кластерів використано метод ліктя. Необхідна кількість кластерів дорівнює 3, так як при цьому значенні відбувається значний перегин кривої.

4. Ділимо вхідні дані на 3 кластери:

Рисунок 2 - Результат кластеризації

Центри кластерів є наступними:

```
        died       age
1 0.05322156 40.696462
2 0.01042287  6.808219
3 0.26476578 63.948614
```

Метод *k*-середніх кластеризує дані з точністю 87,6%. Цей відсоток є мірою загальної дисперсії в наборі даних. Наступна таблиця показує отримані акуратності кожної з побудованих моделей. Таким чином, можна зробити висновок, що найкраще для розглянутих у роботі даних працює модель випадкового лісу, точність якого становить 89,2%.

Таблиця 2 – Результати алгоритмів

Model	Model Accuracy
Logistic Regression	0.8912824
Support Vector Machine	0.8885
Random Forest	0,8921117
K-Means	0.876

Висновки.

Методи машинного навчання стрімко розвиваються – щодня країни з найбільш розвиненими економіками світу вкладають багато ресурсів у впровадження комп'ютерних «мізків» у життєдіяльність людини. Таким чином спеціалістам стали відомі досі невідомі взаємозалежності в даних. Важливо, застосовувати ці знання для розвитку медицини. Адже, медична галузь за останні 30 років, на жаль, не набула такого стрімкого розвитку, як технічний прогрес.

У даній статті проведено порівняння ефективності методів машинного навчання на прикладі даних поширення Covid-19. Розглянуто наступні

алгоритми машинного навчання: логістична регресія (Logistic Regression), метод опорних векторів (Support Vector Machine), випадковий ліс (Random Forest), метод -середніх (K-Means). Для порівняння моделей використано акуратності побудованих моделей.

Література:

[1] Anandhanathan, Praveen & Gopalan, Priyanka. (2021). Comparison of Machine Learning algorithm for COVID-19 Death Risk Prediction. 10.21203/rs.3.rs-196077/v1.

[2] Edureka courses https://www.edureka.co/blog/random-forest-classifier/ [Електронний ресурс]

Abstract. *The article compares the effectiveness of Machine Learning Algorithms on the example of Covid-19 distribution data. The following Machine Learning algorithms are considered: Logistic Regression, Support Vector Machine, Random Forest, K-Means. The accuracy of the constructed models was used to compare the Algorithms. All statistical studies were conducted in the R Programming language.*

Keywords: *Machine Learning, Random Forest, Support Vector Machine, Logistic Regression, K-means.*

Статья отправлена: 20.01.2022 г.