

Невінський Денис Володимирович*Національний університет «Львівська політехніка», Львів*

ORCID 0000-0002-0962-072X

Мартьянов Дмитро Ігорович*Національний університет «Львівська політехніка», Львів*

ORCID 0009-0003-3919-4412

Виклюк Ярослав Ігорович*Національний університет «Львівська політехніка», Львів*

ORCID 0000-0003-4766-4659

Сем'янів Ігор Олександрович*Буковинський державний медичний університет, Чернівці*

ORCID 0000-0003-0340-0766

Верещак Кирило Геннадійович*Одеський національний університет імені І. І. Мечникова, Одеса*

ORCID 0009-0009-8836-4143

АНАЛІЗ ВАЖЛИВОСТІ ФАКТОРІВ ТА ЕФЕКТИВНОСТІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ У ПРОГНОЗУВАННІ ВИПАДКІВ ТУБЕРКУЛЬОЗУ В ІНДІЇ

Анотація: Туберкульоз (ТБ) залишається однією з найсерйозніших інфекційних хвороб у світі, а Індія є країною з найвищим тягарем цього захворювання за даними Всесвітньої організації охорони здоров'я (ВООЗ). Висока щільність населення, нерівний доступ до медичних послуг, соціально-економічні умови та супутні захворювання, такі як діабет, створюють сприятливе середовище для поширення туберкульозу. У цьому дослідженні проведено аналіз важливості факторів, що впливають на захворюваність на туберкульоз в Індії, та оцінено ефективність моделей машинного навчання для прогнозування випадків захворювання. Метою роботи є виявлення ключових детермінант поширення туберкульозу та розробка науково обґрунтованих рекомендацій для зменшення епідеміологічного навантаження в регіоні.

Дослідження базується на даних за 2019–2022 роки, зібраних із відкритих джерел, зокрема баз ВООЗ та урядових звітів Індії. Набір даних включає 126 записів і 25 змінних, серед яких показники діагностики (кількість виявлених випадків ТБ, мультирезистентний ТБ, ТБ-ВІЛ коінфекція), соціальні фактори (вживання тютюну, алкоголю), інфраструктура медичних закладів, а також результати лікування (успішність, смертність, переривання терапії). Для аналізу використано описову статистику, кореляційний аналіз, множинну лінійну регресію з $L1/L2$ регуляризацією (Ridge, Lasso) та нелінійні методи машинного навчання, зокрема дерева рішень (Decision Tree), метод найближчих сусідів (KNN), метод опорних векторів (SVM) і випадковий ліс (Random Forest). Точність моделей оцінювалась за допомогою крос-валідації, використовуючи метрики R^2 (коефіцієнт детермінації) та MSE (середньоквадратична помилка).

Результати кореляційного аналізу показали відсутність сильних лінійних зв'язків між факторами та загальною кількістю випадків ТБ, що вказує на нелінійну природу залежностей. Множинна лінійна регресія продемонструвала низьку пояснювальну здатність ($R^2 \approx 0.3$), а регуляризовані методи (Lasso з $\alpha = 0.01$) децю покращили узагальнення ($R^2 = 0.1007$). Найвищу точність серед лінійних моделей показали фактори, пов'язані з кількістю зареєстрованих випадків серед чоловіків і жінок, а також діагностикою мультирезистентного ТБ. Нелінійні моделі виявилися ефективнішими: початковий аналіз показав, що Random Forest ($R^2 = 0.4595$ на тестових даних)

переверишує KNN і SVM, тоді як Decision Tree страждає від перенавчання ($R^2 = -0.3044$ на тестових даних).

Для підвищення точності було запропоновано нову цільову змінну – нормалізовану кількість випадків ТБ на 100 000 населення (*total_inf*), що враховує чисельність населення штатів. Після цього точність моделей значно зростає: Decision Tree досяг $R^2 = 0.8724$, а Random Forest – $R^2 = 0.8378$ на тестових даних. Факторний аналіз підтвердив, що ключовими предикторами є діагностика мультирезистентного ТБ (MDR/RR TB Diagnosed) та інфраструктура центрів лікування (PMDT-Infrastructure), що відображає важливість медичних ресурсів і своєчасного виявлення резистентних форм захворювання.

Ключові слова: туберкульоз, Індія, машинне навчання, множинна лінійна регресія, випадковий ліс, дерево рішень, SHAP-аналіз, соціально-економічні фактори, мультирезистентний туберкульоз, прогнозування захворюваності.

Denys Nevinskyi

Lviv Polytechnic National University, Lviv

ORCID 0000-0002-0962-072X

Dmytro Martjanov

Lviv Polytechnic National University, Lviv

ORCID 0009-0003-3919-4412

Yaroslav Vykyuk

Lviv Polytechnic National University, Lviv

ORCID 0000-0003-4766-4659

Ihor Semianiv

Bukovinian State Medical University, Chernivtsi

ORCID 0000-0003-0340-0766

Kyrylo Vereshchak

Odesa I.I.Mechnikov National University, Odesa

ORCID 0009-0009-8836-4143

ANALYSIS OF FEATURE IMPORTANCE AND EFFECTIVENESS OF MACHINE LEARNING MODELS IN PREDICTING TUBERCULOSIS CASES IN INDIA

Abstract: Tuberculosis (TB) remains one of the most severe infectious diseases globally, with India bearing the highest burden according to the World Health Organization (WHO). High population density, unequal access to healthcare, socioeconomic conditions, and comorbidities such as diabetes create a conducive environment for TB spread. This study analyzes the importance of factors influencing TB incidence in India and evaluates the effectiveness of machine learning models in predicting cases. The aim is to identify key determinants of TB spread and develop evidence-based recommendations to reduce the epidemiological burden in the region.

The study utilizes data from 2019–2022, sourced from open databases, including WHO and Indian government reports. The dataset comprises 126 records and 25 variables, encompassing diagnostic indicators (e.g., detected TB cases, multidrug-resistant TB, TB-HIV coinfection), social factors (e.g., tobacco and alcohol use), healthcare infrastructure, and treatment outcomes (e.g., success, mortality, treatment interruption). The analysis employed descriptive statistics, correlation analysis, multiple linear regression with L1/L2 regularization (Ridge, Lasso), and nonlinear machine learning methods, including Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest. Model accuracy was assessed via cross-validation using R^2 (coefficient of determination) and MSE (mean squared error) metrics.

Correlation analysis revealed no strong linear relationships between factors and total TB cases, suggesting nonlinear dependencies. Multiple linear regression showed low explanatory power ($R^2 \approx 0.3$), while regularized methods (Lasso with $\alpha = 0.01$) slightly improved generalization ($R^2 = 0.1007$). Among

linear models, factors related to case notifications by gender and multidrug-resistant TB diagnosis were most significant. Nonlinear models proved more effective: initial analysis indicated Random Forest ($R^2 = 0.4595$ on test data) outperformed KNN and SVM, while Decision Tree suffered from overfitting ($R^2 = -0.3044$ on test data).

To enhance accuracy, a new target variable—normalized TB cases per 100,000 population (total_inf)—was introduced, accounting for state population sizes. This adjustment significantly improved model performance: Decision Tree achieved $R^2 = 0.8724$, and Random Forest reached $R^2 = 0.8378$ on test data. Factor analysis confirmed that multidrug-resistant TB diagnosis (MDR/RR TB Diagnosed) and treatment center infrastructure (PMDT-Infrastructure) are key predictors, highlighting the critical role of medical resources and timely detection of resistant strains.

Keywords: tuberculosis, India, machine learning, multiple linear regression, random forest, decision tree, SHAP analysis, socioeconomic factors, multidrug-resistant tuberculosis, incidence prediction.

1. Вступ

Туберкульоз (ТБ) є одним з найбільш серйозних інфекційних захворювань у світі, що залишається значною загрозою для громадського здоров'я. За даними Всесвітньої організації охорони здоров'я (ВООЗ), Індія є країною з одним із найвищих рівнів захворюваності на туберкульоз, що робить вивчення його поширення вкрай актуальним завданням. Висока щільність населення, нерівний доступ до медичних послуг та соціально-економічні чинники сприяють поширенню цієї хвороби.

Аналіз факторів, що впливають на поширення туберкульозу, є ключовим для розробки ефективних заходів профілактики та контролю. Серед таких факторів можна виділити соціально-економічні умови, супутні захворювання (зокрема, діабет), доступ до медичних послуг та рівень вакцинації населення. Використання методів машинного навчання та статистичних моделей дозволяє не лише оцінити вплив цих факторів, а й прогнозувати майбутні тенденції поширення хвороби.

У цьому дослідженні проведено аналіз поширення туберкульозу в Індії із застосуванням регресійних моделей. Основна мета роботи – виявити ключові детермінанти захворюваності та оцінити можливості прогнозування на основі зібраних даних. Це дозволить розробити науково обгрунтовані рекомендації для зменшення тягаря хвороби в регіоні.

2. Аналіз літературних даних і постановка проблеми

Дослідження проблеми туберкульозу активно ведеться у світовій науковій спільноті. ВООЗ у своїх щорічних звітах [1] наголошує на зростанні кількості випадків туберкульозу серед вразливих груп населення, а також на необхідності впровадження нових методів діагностики та лікування. Щорічні звіти Національної програми ліквідації туберкульозу Індії (NTEP) [2] надають детальний огляд епідеміологічної ситуації, прогресу в діагностиці та лікуванні туберкульозу, включно з ко-інфекціями, зокрема цукровим діабетом. У звітах узагальнено ключові показники на національному та регіональному рівнях, що дозволяє оцінити динаміку змін, ефективність втручань і залишкові виклики. Дані використовуються як основа для планування політик та моніторингу реалізації стратегії "End TB" у межах Цілей сталого розвитку. Сучасні дослідження активно розглядають застосування методів машинного навчання у медицині, зокрема у прогнозуванні та діагностиці інфекційних захворювань.

Прогнозування та діагностика інфекційних захворювань

У роботі [2] досліджено моделі прогнозування денге, засновані на методах машинного навчання. Аналіз охопив три основні підходи: діагностичний, епідеміологічний та інтервенційний. Найбільш поширеним методом у діагностичному підході є логістична регресія, тоді як у просторовому аналізі епідеміологічного підходу – лінійна регресія. У дослідженні [3] запропоновано метод машинного навчання для покращення функцій оцінки докінгу в процесі перепрофілювання лікарських засобів. Використання методів опорних векторів дозволило підвищити ефективність відбору потенційних інгібіторів *Mycobacterium tuberculosis*.

Аналіз резистентності та терапевтичні підходи

У роботі [4] використано ансамбль алгоритмів штучного інтелекту, включаючи випадкові ліси, для виявлення критичних концентраційних порогів лікарських препаратів, що

асоціюються з невдачею терапії або летальними наслідками. У дослідженні [10] розглянуто методики швидкого та точного прогнозування резистентності *Mycobacterium tuberculosis* до антибіотиків на основі геномних даних. Використання глибоких згорткових нейронних мереж дозволило значно покращити точність класифікації.

Автоматизована діагностика туберкульозу

У роботі [5] запропоновано метод автоматичного виявлення туберкульозу за рентгенівськими знімками легень із використанням ансамблевого навчання та гібридних дескрипторів ознак, що демонструє високу точність діагностики. Дослідження [6] присвячене використанню методів машинного навчання для аналізу резистентності *Mycobacterium tuberculosis*. Було протестовано кілька класифікаторів, серед яких логістична регресія та градієнтний бустинг дерев, що покращило точність прогнозування.

Штучний інтелект у респіраторній медицині

У дослідженні [7] розглянуто застосування штучного інтелекту у респіраторній медицині, зокрема в аналізі торакальних знімків, гістологічних препаратів та даних функціональних тестів легень. Встановлено, що глибокі нейронні мережі перевершують людських експертів у розпізнаванні патологій. У роботі [8] порівняно ефективність методів множинної лінійної регресії та нейронних мереж у прогнозуванні антимікобактеріальної активності хімічних сполук. Асоціативні нейронні мережі забезпечили вищу точність прогнозів.

Класифікація та контроль поширення туберкульозу

У дослідженні [10] розроблено критерії класифікації туберкульозного увеїту за допомогою машинного навчання, що підвищило точність класифікації. У роботі [11] використано дерева рішень для аналізу факторів ризику туберкульозу у великої рогатої худоби. Було встановлено, що ключовими предикторами поширення захворювання є наявність заражених сусідів та кількість перевезень худоби. У дослідженні [12] запропоновано використання методів машинного навчання для прогнозування ефективності малих молекул проти *Mycobacterium tuberculosis*. Використання байесівських класифікаторів і методів опорних векторів дозволило виявити перспективні сполуки для подальших доклінічних досліджень. У роботі [14] розглянуто роль вдосконалених методів діагностики у зниженні глобального тягаря туберкульозу. Запропоновано впровадження швидких і точних тестів, що можуть суттєво покращити раннє виявлення та контроль захворювання.

Попри значну кількість досліджень, присвячених прогнозуванню туберкульозу в країнах із високим епідеміологічним навантаженням, більшість із них зосереджені або на клінічних даних, або на аналізі зображень (рентген, КТ). Значно менше уваги приділено комплексному аналізу соціальних, демографічних, інфраструктурних і поведінкових факторів, що впливають на поширення захворювання на макрорівні (регіони, штати).

У контексті сучасних досліджень у сфері медичного прогнозування ця робота займає нішу між традиційною епідеміологією та новітніми технологіями інтелектуального аналізу даних. Вона пропонує інтерпретовані та прикладні результати, які можуть бути використані для формування цільових рекомендацій на рівні системи охорони здоров'я.

3. Мета і задачі дослідження

Метою дослідження є аналіз факторів поширення туберкульозу в Індії та оцінка моделей машинного навчання для прогнозування випадків із розробкою рекомендацій щодо профілактики.

Для досягнення поставленої мети вирішено такі завдання:

- Зібрано дані про туберкульоз в Індії за 2019–2022 роки з відкритих джерел.
- Проведено кореляційний аналіз для виявлення зв'язків між факторами та захворюваністю.
- Протестовано моделі лінійної регресії (Ridge, Lasso) для прогнозування.
- Застосовано нелінійні методи (Decision Tree, KNN, SVM, Random Forest) для порівняння точності.

- Удосконалено моделі шляхом нормалізації цільової змінної.
- Визначено ключові фактори, такі як мультирезистентний ТБ, за допомогою факторного аналізу.
- Розроблено рекомендації для покращення профілактики туберкульозу в Індії.

Наукова новизна цієї роботи полягає в тому, що вперше на прикладі даних про туберкульоз в Індії:

- застосовано системний підхід до аналізу нелінійних взаємозв'язків між факторами за допомогою машинного навчання;
- запропоновано метод нормалізації цільової змінної (total_inf), що враховує демографічні особливості регіонів;
- здійснено порівняння лінійних і нелінійних моделей та проведено глибокий аналіз важливості ознак;
- продемонстровано, як сучасні ML-алгоритми можуть використовуватися як інструмент для прийняття управлінських рішень у сфері охорони здоров'я.

4. Методологія дослідження

У межах даного дослідження було не лише протестовано низку моделей машинного навчання, а й здійснено порівняльний аналіз можливих підходів до побудови класифікаторів та вибору релевантних ознак. Це дозволило обґрунтовано вибрати оптимальні методи прогнозування та інтерпретації результатів.

Джерела даних

Дані для аналізу отримані з відкритих джерел, зокрема бази даних ВООЗ та урядових звітів Індії. До аналізу включено такі змінні:

- Рівень захворюваності на туберкульоз;
 - Частка хворих на діабет серед пацієнтів з туберкульозом;
 - Соціально-економічні фактори (рівень доходу, доступ до медичних послуг);
 - Географічні особливості регіонів.
- Методи побудови моделей

Описова статистика – використана для узагальнення характеристик вибірки та первинної оцінки тенденцій у даних.

Кореляційний аналіз – проведено для визначення залежностей між змінними.

Регресійний аналіз – використано множинну лінійну регресію з використанням L1/L2 регуляризації для виявлення найбільш значущих факторів, що впливають на рівень захворюваності.

Методи машинного навчання – застосовано алгоритми дерева рішень (Decision Tree) та випадкових лісів (Random Forest) для підвищення точності прогнозування.

Запропоновано метод перетворення даних що дозволить підняти точність пропонуванних моделей.

Крім моделей, наведених у статті (Decision Tree, Random Forest, SVM, KNN), додатково розглядалися такі алгоритми:

- Градієнтний бустинг (Gradient Boosting, XGBoost) — потужний ансамблевий метод, здатний моделювати складні нелінійні залежності навіть на невеликих вибірках;
- Логістична регресія — як базовий інтерпретований підхід для класифікації супутніх захворювань та демографічних категорій;
- Нейронні мережі — попри складність інтерпретації, були розглянуті як тестова альтернатива для перевірки точності.

Методи вибору та інтерпретації важливості ознак

З метою підвищення достовірності висновків були застосовані або проаналізовані наступні методи:

- Лінійні статистичні методи:
 - χ^2 (хи-квадрат) — для категоріальних ознак;
 - ANOVA F-оцінки — для визначення дисперсійної значущості;

- Mutual Information — міра спільної залежності змінних.
- Інтерпретовані ML-методи:
 - Permutation Feature Importance — оцінює вплив ознак шляхом випадкової перестановки значень;
 - SHAP (SHapley Additive exPlanations) — сучасний універсальний підхід до пояснення як лінійних, так і нелінійних моделей, дозволяє будувати локальні та глобальні інтерпретації впливу кожної змінної.

Метрики оцінки моделей

Для оцінки точності побудованих моделей використано метод крос-валідації (k-fold cross-validation). Оцінювання точності виконувалося за кількома метриками, що враховують різні аспекти якості моделей:

- R^2 (коефіцієнт детермінації) — основна метрика для оцінки пояснювальної здатності;
- MSE (mean squared error) — середньоквадратична похибка;
- MAE (mean absolute error) — середня абсолютна похибка;
- RMSE (root mean squared error) — квадратний корінь із MSE, чутливий до великих відхилень.

Застосування вищезгаданих підходів дало змогу не лише підвищити точність прогнозування, але й забезпечити прозорість результатів, необхідну для подальшого використання у системах охорони здоров'я. З міркувань обмеженого обсягу статті було представлено лише ті результати, які демонстрували найкращу якість моделі та найкращу інтерпретованість у контексті поставленого завдання.

5. Результати та обговорення

5.1. Джерела даних

У цьому дослідженні використовуються офіційні дані щодо туберкульозу в Індії, зібрані з 2019 по 2022 рік для всіх штатів країни. Дані включають інформацію про кількість випадків туберкульозу, супутні захворювання, соціальні фактори та ефективність лікування [2].

Структура набору даних

Набір містить 126 записів та 25 вхідних змінних. Основні поля включають:

1. Географічні та часові дані:
 - City – місто або регіон Індії
 - Year – рік реєстрації випадків
2. Показники діагностики та виявлення ТБ:
 - Active Case Finding TB cases diagnosed among tested – кількість підтверджених випадків ТБ серед протестованих.
 - Active Case Finding TB Presumptive TB cases tested out of those screened – кількість протестованих пацієнтів із підозрою на ТБ.
 - MDR/RR TB DIAGNOSED MDR/RR patient diagnosed – кількість випадків мультирезистентного туберкульозу.
 - Paediatric TB patients notified – кількість зареєстрованих випадків ТБ серед дітей.
 - TB-HIV co-infected patients Diagnosed – кількість діагностованих випадків коінфекції ТБ-ВІЛ.
 - TB-HIV co-infected patients Put on ART – кількість коінфікованих пацієнтів, які почали антиретровірусну терапію.
3. Соціальні фактори та поведінкові особливості пацієнтів:
 - TB patients with known Tobacco usage status – кількість пацієнтів із відомим статусом вживання тютюну.
 - TB patients with known Alcohol usage status – кількість пацієнтів із відомим статусом вживання алкоголю.
 - TB-COVID 19 patients detected – кількість випадків ТБ серед пацієнтів, які також хворіли на COVID-19.
 - Pregnant TB patients identified – кількість вагітних жінок із ТБ.
4. Показники лікування та результатів терапії:

- Treatment outcome of TB patients notified in (% Lost to follow up) – частка пацієнтів, які припинили лікування.
 - Treatment outcome of TB patients notified in (Death Rate) – рівень смертності серед пацієнтів із ТБ.
 - Treatment outcome of TB patients notified in (Success Rate) – рівень успішного лікування.
 - Treatment outcome of TB patients notified in (Treatment Failure Rate) – рівень невдалого лікування.
 - Treatment outcome of TB-HIV patients notified in (Death Rate) – смертність серед коінфікованих пацієнтів.
 - Treatment outcome of TB-HIV patients notified in (Success Rate) – рівень успішного лікування коінфікованих пацієнтів.
 - Treatment outcome of Paediatric TB patients notified in (Death Rate) – смертність серед дітей із ТБ.
 - Treatment outcome of Paediatric TB patients notified in (Success Rate) – успішне лікування серед дітей.
5. Інфраструктура медичних закладів та супутні захворювання:
- PMDT- Infrastructure No. of Nodal DR-TB centres – кількість центрів лікування лікарсько-стійкого туберкульозу.
 - TB- DM patients initiated on Anti-diabetic treatment – кількість пацієнтів із ТБ і діабетом, які почали лікування діабету.
 - TB - DM Patients diagnosed among tested – кількість пацієнтів із діабетом серед протестованих на ТБ.

6. Цільова змінна (вихідне поле):

- TB case notification total – загальна кількість зареєстрованих випадків туберкульозу.

Описова статистика та кореляційний аналіз

Для оцінки варіативності випадків туберкульозу було побудовано розподіл значень цільового показника TB case notification total, який представлено на Рисунку 1. Як видно з графіка, найбільша кількість зареєстрованих випадків знаходиться в діапазоні 50–250 на 10 000 населення. Це вказує на відносно рівномірний рівень захворюваності в більшості регіонів.

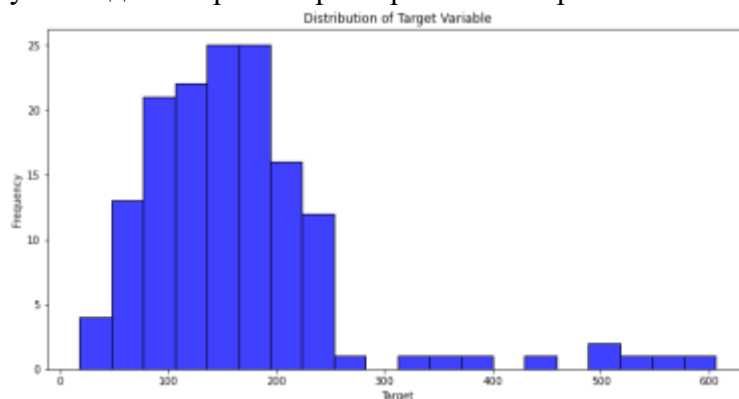


Рис 1. Розподіл випадків туберкульозу

Однак у певних штатах Індії зафіксовано значно вищий рівень захворюваності, що може бути пов'язано із соціально-економічними умовами, доступністю до медичної допомоги, кліматичними факторами або іншими причинами. Ці відмінності вимагають подальшого аналізу, зокрема дослідження впливу окремих чинників на динаміку поширення туберкульозу.

5.2. Кореляційний аналіз факторів

Для визначення можливих зв'язків між вхідними змінними та цільовим полем TB case notification total було обчислено матрицю кореляцій, що представлена на Рисунку 2.

З аналізу кореляційної таблиці видно, що:

- Відсутні сильні лінійні залежності між вхідними змінними та цільовим показником. Це свідчить про те, що жоден із факторів самостійно не має прямого пропорційного або оберненого впливу на рівень захворюваності.

• Не спостерігається жодного від'ємного зв'язку, що означає, що всі змінні мають або позитивний, або нейтральний вплив на вихідну змінну.

• Оскільки лінійні зв'язки між факторами та цільовим показником відсутні, можна припустити, що вплив цих факторів є нелінійним, що вимагає застосування методів машинного навчання для моделювання залежностей.

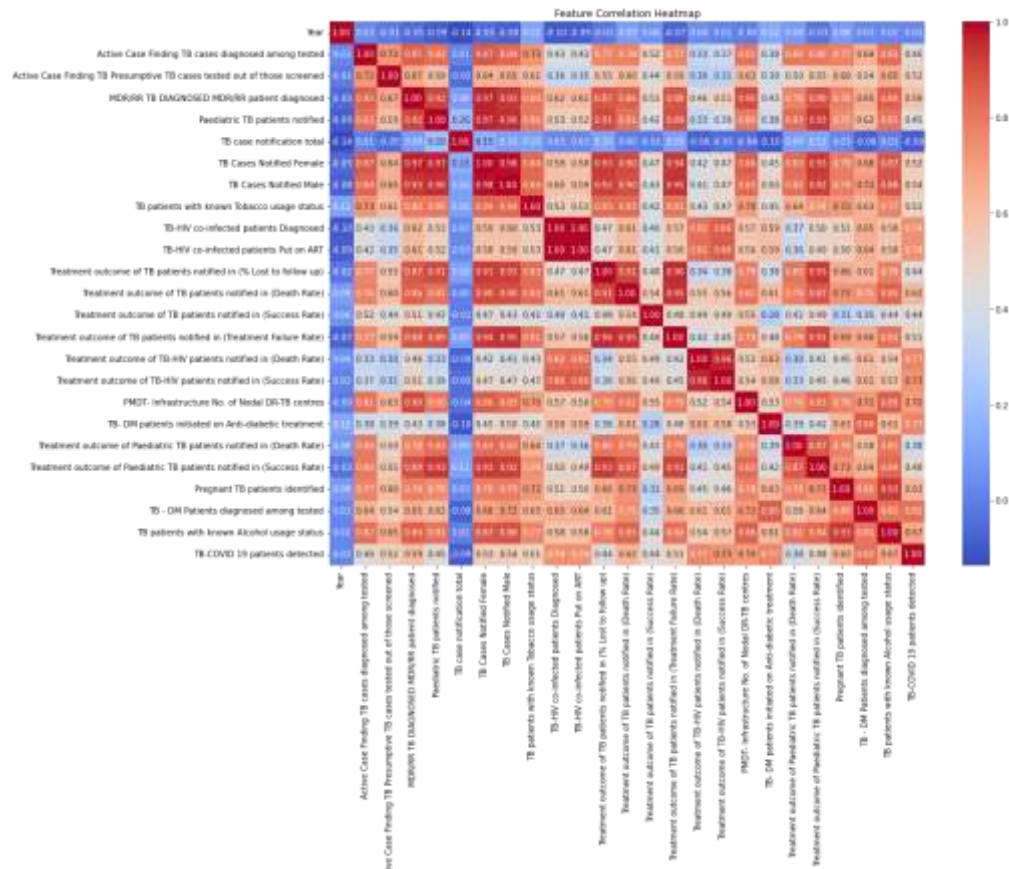


Рис. 2. Кореляційна матриця взаємозв'язків між факторами.

Отримані результати підтверджують необхідність подальшого використання нелінійних моделей, які можуть краще ідентифікувати складні взаємозв'язки між факторами та поширеністю туберкульозу в різних регіонах Індії. З іншого боку невелика кількість даних дає змогу використати не лінійні моделі за рахунок великої кількості емпіричних параметрів. Тому в дослідженні були використані відносно прості моделі машинного навчання для встановлення зв'язку між факторами та цільовим показником.

5.3. Регресійний аналіз

Для виявлення найбільш значущих факторів, що впливають на рівень захворюваності на туберкульоз, було використано множинну лінійну регресію. Дослідження передбачало оцінку ефективності базової лінійної моделі, а також використання регуляризованих методів Ridge та Lasso для покращення якості прогнозування.

Оцінка множинної лінійної регресії

Для оцінки точності моделі дані були розділені на тренувальну (80%) та тестову (20%) вибірки. Додатково було застосовано крос-валідацію із параметрами $n_splits=5$, $shuffle=True$.

Точність моделі оцінювалася за такими метриками:

• Mean Squared Error (MSE) – середньоквадратична помилка, що оцінює середню величину помилок моделі.

• R-squared (R^2) – коефіцієнт детермінації, що показує, яка частка варіації залежної змінної пояснюється незалежними змінними.

Результати регресійної моделі подано в Таблиці 1.

Таблиця 1

Оцінка точності множинної лінійної регресії

Метрика	Тренувальні дані	Тестові дані	Середнє значення (крос-валідація)
Mean Squared Error (MSE)	0.6005	0.9802	5.1966
R-squared (R^2)	0.3097	0.3007	-5.3676

Як видно з таблиці, модель демонструє низький рівень пояснювальної здатності ($R^2 \approx 0.3$), що свідчить про те, що лінійна регресія не є ефективною для прогнозування рівня захворюваності. Крім того, негативний показник середньої точності в крос-валідації ($R^2 = -5.37$) вказує на перенавчання моделі на тренувальній вибірці та її слабку узагальнюючу здатність.

5.4. Використання регуляризованих методів Ridge та Lasso

Оскільки базова лінійна модель не продемонструвала задовільних результатів, було застосовано регуляризовані методи Ridge та Lasso. Для налаштування гіперпараметрів було використано GridSearchCV з перебором значень alpha:

{'alpha': [1, 0.1, 0.01, 0.001, 0.0001]}

Оптимальні значення параметрів та результати моделювання наведені в Таблиці 2.

Таблиця 2

Найкращі результати Ridge та Lasso

Модель	Best Alpha (MSE)	Best Score (MSE)	Best Alpha (R^2)	Best Score (R^2)
Lasso	0.01	-0.8152	0.01	0.1007
Ridge	1.00	-0.9429	1.00	-0.0505

Отримані результати свідчать про те, що метод **Lasso-регресії з параметром $\alpha = 0.01$** продемонстрував найкраще узагальнення серед усіх протестованих моделей ($R^2 = 0.1007$). Водночас Ridge-регресія не змогла покращити результати базової моделі.

5.5. Оцінка вагомості факторів

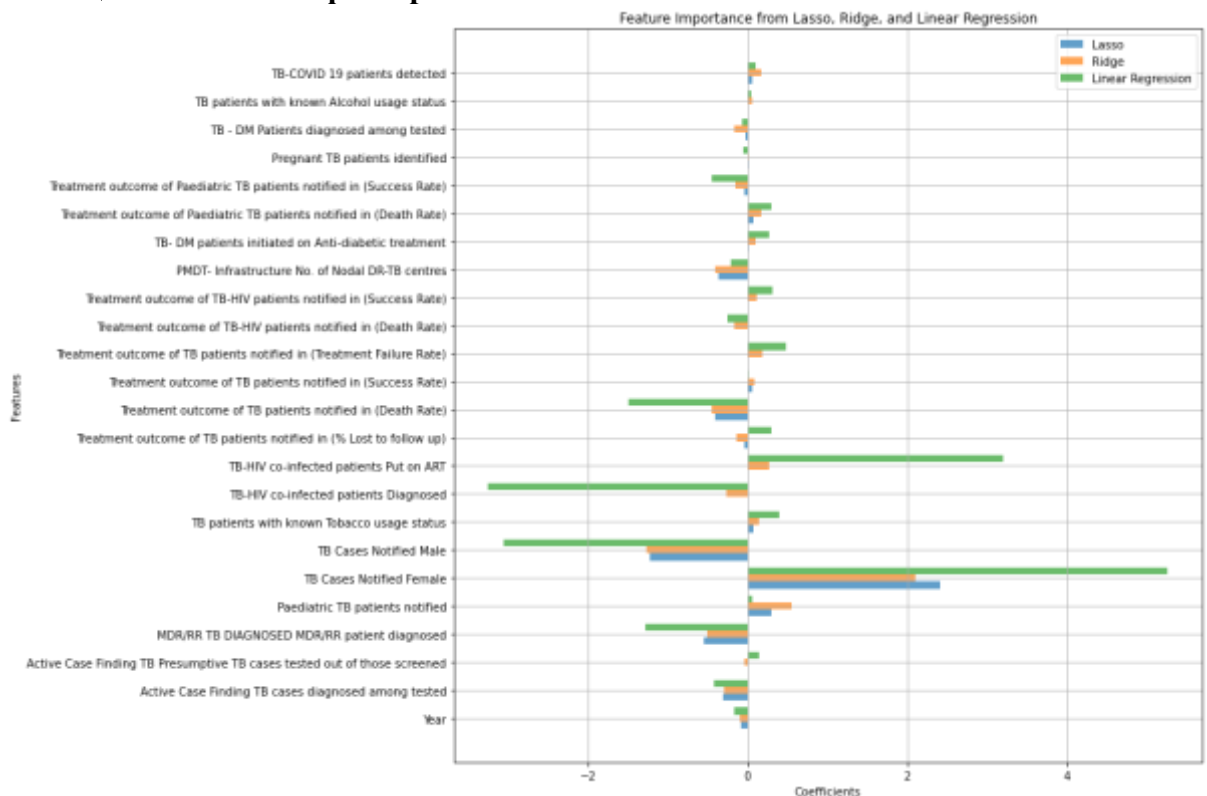


Рис. 3. Оцінка важливості факторів для моделі множинної регресії

Для оцінки важливості факторів використовувалися коефіцієнти регресійних моделей. Оскільки всі вхідні змінні були **нормалізовані**, це дозволяє безпосередньо порівнювати їхню вагу та вплив у межах різних моделей. Даний аналіз дозволяє визначити **ключові фактори**,

що впливають на рівень захворюваності, а також порівняти ці результати з іншими методами машинного навчання у подальших дослідженнях.

На Рисунку 3 представлено графічне порівняння значущості факторів у різних моделях, а результати кількісного аналізу наведено у Таблиці 3.

Таблиця 3

Важливість факторів моделей множинної регресії

Feature	Lasso Coefficient	Ridge Coefficient	Linear Regression Coefficient
TB Cases Notified Female	2.419684	2.107322	5.267556
TB-HIV co-infected patients Put on ART	0.000000	0.270044	3.208259
Treatment outcome of TB patients notified ...	0.000000	0.183351	0.480899
TB patients with known Tobacco usage status	0.065768	0.140842	0.403452
Treatment outcome of TB-HIV patients notified ...	-0.000000	0.108874	0.306359
Treatment outcome of Paediatric TB patients no...	0.076777	0.175554	0.296422
Treatment outcome of TB patients notified in (...)	-0.034294	-0.135819	0.292421
TB- DM patients initiated on Anti-diabetic tre...	0.008892	0.098152	0.270214
Active Case Finding TB Presumptive TB cases te...	-0.000000	-0.037047	0.143527
TB-COVID 19 patients detected	0.057069	0.167566	0.105165
Paediatric TB patients notified	0.300905	0.545590	0.062164
TB patients with known Alcohol usage status	0.000000	0.061360	0.041262
Treatment outcome of TB patients notified in (...)	0.056771	0.083941	0.008838
Pregnant TB patients identified	0.021074	0.011369	-0.060003
TB - DM Patients diagnosed among tested	-0.021035	-0.165178	-0.065678
Year	-0.077361	-0.095665	-0.166723
PMDT- Infrastructure No. of Nodal DR-TB centres	-0.367928	-0.403824	-0.204416
Treatment outcome of TB-HIV patients notified ...	-0.000000	-0.167968	-0.253432
Active Case Finding TB cases diagnosed among t...	-0.311439	-0.293686	-0.414591
Treatment outcome of Paediatric TB patients no...	-0.038630	-0.158344	-0.445773
MDR/RR TB DIAGNOSED MDR/RR patient diagnosed	-0.553758	-0.510831	-1.288347

Feature	Lasso Coefficient	Ridge Coefficient	Linear Regression Coefficient
Treatment outcome of TB patients notified in (...)	-0.401312	-0.449354	-1.494065
TB Cases Notified Male	-1.231524	-1.271014	-3.061572
TB-HIV co-infected patients Diagnosed	0.000000	-0.268894	-3.250397

5.6. Порівняння важливості факторів у різних моделях

Як видно з таблиці та рисунку, розподіл факторів залишається однаковим для всіх моделей, проте їхня абсолютна величина дещо змінюється. Це пов'язано з тим, що різні методи оцінки (Lasso, Ridge, Linear Regression) використовують різні підходи до регуляризації та вибору вагових коефіцієнтів.

Найбільш вагомими факторами для прогнозування рівня захворюваності на туберкульоз є:

- “TB Cases Notified Female” та “TB Cases Notified Male” – ці змінні мають найбільші позитивні та негативні коефіцієнти відповідно. У моделі лінійної регресії їхні коефіцієнти мають найбільші значення (5.27 та -3.06), що свідчить про значний вплив на вихідну змінну.
- “MDR/RR TB DIAGNOSED MDR/RR patient diagnosed” та “Treatment outcome of TB patients notified in (...)” – також мають значні негативні коефіцієнти у всіх моделях, що може свідчити про зв'язок із високим рівнем смертності або складністю лікування.
- “TB-HIV co-infected patients Diagnosed” – у моделі лінійної регресії має високий негативний вплив (-3.25), проте в Lasso ця змінна була відфільтрована, що може вказувати на її низьку стійкість до регуляризації.

Деякі змінні, наприклад, “TB-COVID 19 patients detected” або “Treatment outcome of Paediatric TB patients”, мають менший вплив у порівнянні з основними факторами, проте можуть відігравати важливу роль у моделюванні залежно від вибору алгоритму.

Загалом, результати показують, що хоча різні методи оцінки коефіцієнтів можуть давати трохи різні значення, структура важливості змінних залишається незмінною. Це свідчить про стабільність впливу ключових факторів і дозволяє використовувати ці результати для подальшого дослідження більш складними моделями машинного навчання.

5.7. Методи машинного навчання

Для прогнозування рівня захворюваності на туберкульоз були обрані наступні методи машинного навчання:

- Дерево рішень (Decision Tree)
- Метод найближчих сусідів (K-Nearest Neighbors, KNN)
- Метод опорних векторів (Support Vector Machine, SVM)
- Випадковий ліс (Random Forest)

Основною причиною вибору цих алгоритмів є невеликий розмір вибірки. Лінійні моделі, такі як множинна лінійна регресія, показали низьку точність через відсутність лінійних залежностей між факторами та цільовою змінною. Тому було прийнято рішення використати нелінійні методи, які можуть краще виявляти складні взаємозв'язки між змінними навіть при обмеженій кількості записів.

• Дерево рішень – добре працює на малих наборах даних, оскільки воно не потребує масштабування ознак і може виявляти нелінійні закономірності.

• Метод KNN – підходить для невеликих вибірок, адже його ефективність значно падає на великих наборах даних через необхідність зберігати всі зразки для порівняння.

• Метод опорних векторів (SVM) – ефективний у розпізнаванні складних взаємозв'язків між змінними та працює добре навіть на невеликих вибірках.

• Випадковий ліс (Random Forest) – ансамблевий метод, який поєднує кілька дерев рішень, що дозволяє отримати більш **стійкі** результати, уникнувши переобучення.

5.8. Порівняння точності моделей

Оцінка моделей проводилася за допомогою розбиття даних на навчальний (80%) та тестовий (20%) набори. Основними метриками для порівняння були середньоквадратична помилка (MSE) та коефіцієнт детермінації (R^2). Результати навчання оцінки точності наведені в таблиці 4.

Таблиця 4

Аналіз точності моделей машинного навчання.

Model	Train MSE	Train R^2	Test MSE	Test R^2
Decision Tree	0.0000	1.0000	1.8285	-0.3044
KNN	0.6984	0.1971	1.5141	-0.0802
SVM	0.7383	0.1513	1.3040	0.0697
Random Forest	0.1184	0.8639	0.7577	0.4595

Як видно з таблиці 4:

- Дерево рішень показало ідеальну точність на навчальних даних ($R^2 = 1.0$), що свідчить про переобучення. Однак на тестових даних його продуктивність значно знижується ($R^2 = -0.3044$), що підтверджує його нестійкість.

- KNN та SVM мають кращу генералізацію, але все ще показують низьку точність на тестових даних ($R^2 < 0.1$).

- Випадковий ліс показав найкращу збалансованість між навчальним та тестовим наборами, маючи найвище значення R^2 на тестових даних (0.4595).

На основі отриманих результатів найбільш точними моделями виявилися Decision Tree та Random Forest. Для них було проведено оцінку важливості факторів (Таблиця 5).

Таблиця 5.

Важливість ознак для моделей Decision Tree та Random Forest

Factors	Factors	Decision Tree	Random Forest
Active Case Finding TB Presumptive TB cases tested out of those screened	Active Case Finding TB Presumptive TB cases te...	0.142213	0.105245
PMDT- Infrastructure No. of Nodal DR-TB centres	PMDT- Infrastructure No. of Nodal DR-TB centres	0.164667	0.080200
MDR/RR TB DIAGNOSED MDR/RR patient diagnosed	MDR/RR TB DIAGNOSED MDR/RR patient diagnosed	0.229570	0.069238
TB patients with known Alcohol usage status	TB patients with known Alcohol usage status	0.095303	0.065410
Active Case Finding TB cases diagnosed among tested	Active Case Finding TB cases diagnosed among t...	0.013359	0.065153
Paediatric TB patients notified	Paediatric TB patients notified	0.000015	0.051710
TB Cases Notified Female	TB Cases Notified Female	0.054874	0.050110
TB - DM Patients diagnosed among tested	TB - DM Patients diagnosed among tested	0.077079	0.045450
TB-HIV co-infected patients Put on ART	TB-HIV co-infected patients Put on ART	0.047919	0.043398
Treatment outcome of Paediatric TB patients notified in (Death Rate)	Treatment outcome of Paediatric TB patients no...	0.044695	0.041242

Factors	Factors	Decision Tree	Random Forest
TB- DM patients initiated on Anti-diabetic treatment	TB- DM patients initiated on Anti-diabetic tre...	0.005879	0.039743
Treatment outcome of TB-HIV patients notified in (Success Rate)	Treatment outcome of TB-HIV patients notified ...	0.000309	0.038289
TB-COVID 19 patients detected	TB-COVID 19 patients detected	0.009624	0.035245
TB-HIV co-infected patients Diagnosed	TB-HIV co-infected patients Diagnosed	0.000259	0.034660
Treatment outcome of Paediatric TB patients notified in (Success Rate)	Treatment outcome of Paediatric TB patients no...	0.001263	0.033024
Treatment outcome of TB patients notified in (Death Rate)	Treatment outcome of TB patients notified in (...)	0.000081	0.030597
Treatment outcome of TB patients notified in (Treatment Failure Rate)	Treatment outcome of TB patients notified in (...)	0.022098	0.028695
Treatment outcome of TB-HIV patients notified in (Death Rate)	Treatment outcome of TB-HIV patients notified ...	0.002000	0.026186
TB Cases Notified Male	TB Cases Notified Male	0.005229	0.024557
Treatment outcome of TB patients notified in (% Lost to follow up)	Treatment outcome of TB patients notified in (...)	0.013722	0.022722
Year	Year	0.012904	0.019632
TB patients with known Tobacco usage status	TB patients with known Tobacco usage status	0.048023	0.019236
Treatment outcome of TB patients notified in (Success Rate)	Treatment outcome of TB patients notified in (...)	0.008883	0.019016
Pregnant TB patients identified	Pregnant TB patients identified	0.000034	0.011243

5.9. Порівняння з лінійними моделями

1. Моделі дерев рішень виявили значно більший вплив таких змінних, як “MDR/RR TB DIAGNOSED” та “PMDT Infrastructure”, що відрізняється від лінійних моделей, де найбільші вагові коефіцієнти мали “TB Cases Notified Male” та “TB-HIV co-infected patients Diagnosed”.

2. Лінійні моделі не виділяли такі фактори, як “Active Case Finding TB Presumptive TB cases tested”, тоді як у деревовидних моделях цей фактор займає одну з провідних позицій.

3. Random Forest виявив нижчий вплив окремих змінних, що свідчить про рівномірний розподіл важливості між факторами, тоді як Decision Tree має більш виражене ранжування.

5.10. Аналіз уточненої моделі

З метою підвищення точності моделі було запропоновано використати змінне вихідне поле:

$$\text{data}[\text{'total_inf'}] = 100_000 * \text{data}[\text{'TB case notification total'}] / \text{data}[\text{'population'}]$$

Для цього було створено додаткове поле з кількістю населення для кожного штату залежності від року. Дані отримані із статистичних джерел. Ця формула дозволяє нормалізувати кількість випадків туберкульозу відносно загальної чисельності населення, що сприяє кращій узагальненості моделі.

Кореляційний аналіз після внесених змін показав, що коефіцієнти кореляції практично не змінилися, що свідчить про стабільність факторних залежностей у вибірці.

Оновлені результати точності

Моделі були повторно протестовані:

- Decision Tree:
 - Train - MSE = 0.0000, $R^2 = 1.0000$
 - Test - MSE = 24.4366, $R^2 = 0.8724$
- Random Forest:
 - Train - MSE = 5.9059, $R^2 = 0.9606$
 - Test - MSE = 31.0507, $R^2 = 0.8378$

Аналізуючи оновлені результати, можна зробити наступні висновки:

- Використання нової цільової змінної призвело до суттєвого покращення точності на тестових та навчальних даних.

- Випадковий ліс залишається найбільш узагальнюючою моделлю, хоча деяке переобчислення (збільшення тестової помилки MSE) вказує на необхідність подальшої оптимізації гіперпараметрів.

- Дерево рішень демонструє найкращий коефіцієнт детермінації ($R^2 = 0.8724$), що підтверджує його хорошу адаптацію до нової змінної.

Аналіз факторної важливості

З метою глибшого розуміння значущості факторів у моделі проведено факторний аналіз із використанням двох моделей: дерева рішень та випадкового лісу. Порівняльна таблиця факторної важливості наведена в Таблиці 6.

Таблиця 6

Важливість ознак для моделей Decision Tree та Random Forest для оновленого цільового поля

	Factors	Decision Tree	Random Forest
Factors			
Active Case Finding TB Presumptive TB cases tested out of those screened	Active Case Finding TB Presumptive TB cases te...	0.142213	0.105245
PMDT- Infrastructure No. of Nodal DR-TB centres	PMDT- Infrastructure No. of Nodal DR-TB centres	0.164667	0.080200
MDR/RR TB DIAGNOSED MDR/RR patient diagnosed	MDR/RR TB DIAGNOSED MDR/RR patient diagnosed	0.229570	0.069238
TB patients with known Alcohol usage status	TB patients with known Alcohol usage status	0.095303	0.065410
Active Case Finding TB cases diagnosed among tested	Active Case Finding TB cases diagnosed among t...	0.013359	0.065153
Paediatric TB patients notified	Paediatric TB patients notified	0.000015	0.051710
TB Cases Notified Female	TB Cases Notified Female	0.054874	0.050110
TB - DM Patients diagnosed among tested	TB - DM Patients diagnosed among tested	0.077079	0.045450
TB-HIV co-infected patients Put on ART	TB-HIV co-infected patients Put on ART	0.047919	0.043398

	Factors	Decision Tree	Random Forest
Factors			
Treatment outcome of Paediatric TB patients notified in (Death Rate)	Treatment outcome of Paediatric TB patients no...	0.044695	0.041242
TB- DM patients initiated on Anti-diabetic treatment	TB- DM patients initiated on Anti-diabetic tre...	0.005879	0.039743
Treatment outcome of TB-HIV patients notified in (Success Rate)	Treatment outcome of TB-HIV patients notified ...	0.000309	0.038289
TB-COVID 19 patients detected	TB-COVID 19 patients detected	0.009624	0.035245
TB-HIV co-infected patients Diagnosed	TB-HIV co-infected patients Diagnosed	0.000259	0.034660
Treatment outcome of Paediatric TB patients notified in (Success Rate)	Treatment outcome of Paediatric TB patients no...	0.001263	0.033024
Treatment outcome of TB patients notified in (Death Rate)	Treatment outcome of TB patients notified in (...)	0.000081	0.030597
Treatment outcome of TB patients notified in (Treatment Failure Rate)	Treatment outcome of TB patients notified in (...)	0.022098	0.028695
Treatment outcome of TB-HIV patients notified in (Death Rate)	Treatment outcome of TB-HIV patients notified ...	0.002000	0.026186
TB Cases Notified Male	TB Cases Notified Male	0.005229	0.024557
Treatment outcome of TB patients notified in (% Lost to follow up)	Treatment outcome of TB patients notified in (...)	0.013722	0.022722
Year	Year	0.012904	0.019632
TB patients with known Tobacco usage status	TB patients with known Tobacco usage status	0.048023	0.019236
Treatment outcome of TB patients notified in (Success Rate)	Treatment outcome of TB patients notified in (...)	0.008883	0.019016
Pregnant TB patients identified	Pregnant TB patients identified	0.000034	0.011243

Порівняння факторної важливості

Факторний аналіз для дерева рішень та випадкового лісу після оновлення моделі продемонстрував стабільний розподіл важливості факторів, проте в деяких випадках спостерігається перерозподіл значущості. Зокрема:

- Фактор "MDR/RR TB DIAGNOSED" залишається одним з найважливіших у всіх моделях.

- "PMDT- Infrastructure No. of Nodal DR-TB centres" також стабільно входить до списку найважливіших факторів.

- Роль "TB cases notified female" стала трохи менш важливою в порівнянні з початковим аналізом.

У цілому, результати підтверджують, що перехід до нормалізованої метрики (кількість випадків на 100 000 населення) дозволив підвищити загальну точність моделей і зробити аналіз факторів більш інформативним.

6. Висновки

Було проведено аналіз важливості факторів за допомогою лінійних моделей (Lasso, Ridge, Linear Regression), що дало можливість оцінити внесок окремих змінних у прогнозування цільового показника. Однак, коефіцієнти детермінації R^2 для цих моделей виявилися значно нижчими за аналогічні показники нелінійних методів, що свідчить про слабку здатність лінійних моделей коректно відображати залежності у вибірці.

2. Для побудови прогнозних моделей використано методи машинного навчання (дерево рішень, метод найближчих сусідів, метод опорних векторів, випадковий ліс), що дало можливість порівняти їхню ефективність за критеріями MSE та R^2 . Початковий аналіз показав, що найгірші результати мали KNN (Train $R^2 = 0.1971$, Test $R^2 = -0.0802$) та SVM (Train $R^2 = 0.1513$, Test $R^2 = 0.0697$), у той час як Decision Tree та Random Forest мали кращу узгодженість на тестовій вибірці. Зокрема, Random Forest мав Train $R^2 = 0.8639$ та Test $R^2 = 0.4595$, що вказує на меншу переобучуваність у порівнянні з Decision Tree (Train $R^2 = 1.0000$, Test $R^2 = -0.3044$).

Виконано уточнення моделі шляхом введення нового показника (total_inf) за допомогою методів інженерії ознак. Проведене повторне навчання моделей після застосування нового показника показало зростання точності моделей дерева рішень та випадкового лісу, що дало можливість підтвердити доцільність використання зміненої цільової змінної. Після введення total_inf Decision Tree значно покращився (Test $R^2 = 0.8724$ замість -0.3044), а Random Forest отримав ще більш стабільні результати (Test $R^2 = 0.8378$ замість 0.4595).

Факторний аналіз для деревоподібних моделей дозволив визначити найвпливовіші предиктори, що дало можливість краще зрозуміти природу зв'язків між ознаками та прогнозованим показником. Найвпливовішими факторами для Decision Tree були MDR/RR TB DIAGNOSED MDR/RR patient diagnosed (0.229570) та PMDT- Infrastructure No. of Nodal DR-TB centres (0.164667), що відповідає ключовим медичним параметрам. Аналогічні фактори мали найвищі значення і для Random Forest, що підтверджує їхню важливість.

7. Список використаної літератури

1. World Health Organization. Global tuberculosis report 2024. Retrieved from <https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2024>
2. Ministry of Health and Family Welfare, Government of India. Annual reports [Електронний ресурс] // National Tuberculosis Elimination Programme. — Режим доступу: <https://tbcindia.mohfw.gov.in/annual-reports/> (дата звернення: 13.05.2025).
3. Hoyos, W., Aguilar, J., & Toro, M. (2021). Dengue models based on machine learning techniques: A systematic literature review. *Artificial Intelligence in Medicine*, 119, 102157. <https://doi.org/10.1016/j.artmed.2021.102157>
4. Kinnings, S. L., Liu, N., Tonge, P. J., Jackson, R. M., Xie, L., & Bourne, P. E. (2011). A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of Chemical Information and Modeling*, 51(2), 408-419. <https://doi.org/10.1021/ci100369f>
5. Swaminathan, S., Pasipanodya, J. G., Ramachandran, G., Kumar, A. K. H., Srivastava, S., Deshpande, D., Nuernberger, E., & Gumbo, T. (2016). Drug concentration thresholds predictive of therapy failure and death in children with tuberculosis: Bread crumb trails in random forests. *Clinical Infectious Diseases*, 63, S63-S74. <https://doi.org/10.1093/cid/ciw471>
6. Ayaz, M., Shaukat, F., & Raja, G. (2021). Ensemble learning based automatic detection of tuberculosis in chest X-ray images using hybrid feature descriptors. *Physical and Engineering Sciences in Medicine*, 44(1), 183-194. <https://doi.org/10.1007/s13246-020-00966-0>
7. Kouchaki, S., Yang, Y., Walker, T. M., Walker, A. S., Wilson, D. J., Peto, T. E. A., Crook, D. W., Clifton, D. A., & Hoosdally, S. J. (2019). Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics*, 35(13), 2276-2282. <https://doi.org/10.1093/bioinformatics/bty949>

8. Gonem, S., Janssens, W., Das, N., & Topalovic, M. (2020). Applications of artificial intelligence and machine learning in respiratory medicine. *Thorax*, 75(8), 695-701. <https://doi.org/10.1136/thoraxjnl-2020-214556>
9. Ventura, C., Latino, D. A. R. S., & Martins, F. (2013). Comparison of multiple linear regressions and neural networks based QSAR models for the design of new antitubercular compounds. *European Journal of Medicinal Chemistry*, 70, 831-845. <https://doi.org/10.1016/j.ejmech.2013.10.029>
10. Jabs, D. A., Belfort, R., Bodaghi, B., Graham, E., Gupta, V., Holland, G. N., Lightman, S. L., Oden, N., Palestine, A. G., Smith, J. R., Thorne, J. E., & Trusko, B. E. (2021). Classification criteria for tubercular uveitis. *American Journal of Ophthalmology*, 228, 142-151. <https://doi.org/10.1016/j.ajo.2021.03.040>
11. Kuang, X., Wang, F., Hernandez, K. M., Zhang, Z., & Grossman, R. L. (2022). Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN. *Scientific Reports*, 12(1), 2427. <https://doi.org/10.1038/s41598-022-06449-4>
12. Ekins, S., Pottorf, R., Reynolds, R. C., Williams, A. J., Clark, A. M., & Freundlich, J. S. (2014). Looking back to the future: Predicting in vivo efficacy of small molecules versus *Mycobacterium tuberculosis*. *Journal of Chemical Information and Modeling*, 54(4), 1070-1082. <https://doi.org/10.1021/ci500077v>
13. Romero, M. P., Chang, Y.-M., Brunton, L. A., Parry, J., Prosser, A., Upton, P., Rees, E., Tearne, O., Arnold, M., Stevens, K., & Drewe, J. A. (2020). Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making. *Preventive Veterinary Medicine*, 175, 104860. <https://doi.org/10.1016/j.prevetmed.2019.104860>
14. Keeler, E., Perkins, M. D., Small, P., Hanson, C., Reed, S., Cunningham, J., Aledort, J. E., Hillborne, L., Rafael, M. E., Girosi, F., & Dye, C. (2006). Reducing the global burden of tuberculosis: The contribution of improved diagnostics. *Nature*, 444(Suppl 1), 49-57. <https://doi.org/10.1038/nature05446>