

УДК 004.4:378

DOI <https://doi.org/10.32782/2663-5941/2025.1.2/39>**Ясінський А.М.**

ПВНЗ «Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука»

Джунь Й.В.

ПВНЗ «Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука»

Кундос М.Г.

ПВНЗ «Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука»

Соловей Л.Я.

ПВНЗ «Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука»

ПОРІВНЯЛЬНИЙ АНАЛІЗ ПРОГНОЗУВАННЯ ЗА ДОПОМОГОЮ ЛІНІЙНОЇ РЕГРЕСІЇ ТА НЕЙРОННОЇ МЕРЕЖІ ЗАСОБАМИ БІБЛІОТЕКИ NEURALNET

У статті проведено порівняльний аналіз двох підходів до прогнозування даних: лінійної регресії та нейронної мережі, створеної за допомогою бібліотеки Neuralnet. Дослідження спрямоване на оцінку точності прогнозів кожної моделі з використанням середнього квадратичного відхилення як основного показника ефективності. Вивчено вплив архітектури нейронної мережі, параметрів навчання, а також розміру та характеристик вхідних даних на точність прогнозування. Особливу увагу приділено порівнянню відхилень прогнозованих результатів від реальних даних, що дозволяє оцінити практичну цінність кожної моделі для задач, що потребують високої точності. Методика дослідження включає побудову та тренування моделей на одних і тих самих наборах даних, що забезпечує коректність порівняння. Для лінійної регресії використано традиційні статистичні підходи до моделювання залежності між змінними, тоді як для нейронної мережі реалізовано багатошаровий перцептрон із різними кількостями прихованих шарів і нейронів. Аналіз результатів проводився з урахуванням змін точності моделей залежно від складності їхньої архітектури та обчислювальних ресурсів, необхідних для навчання.

Отримані результати демонструють, що нейронна мережа може досягати більш високої точності у прогнозуванні складних нелінійних залежностей, тоді як лінійна регресія є більш ефективною для задач із лінійними взаємозв'язками між змінними. Також відзначено, що обчислювальні витрати на навчання нейронної мережі значно перевищують витрати для лінійної регресії, що слід враховувати при виборі методу для реальних задач. Результати дослідження можуть бути корисними для фахівців у галузі аналізу даних, машинного навчання та розробки програмного забезпечення, а також для студентів і науковців, які займаються вивченням методів прогнозування.

Стрімкий прогрес у комп'ютерних технологіях прокладає шлях до створення нейрокомп'ютерів шостого покоління. Експерти зі штучного інтелекту прогнозують, що ці системи будуть функціонувати за принципами, подібними до роботи біологічних нейронних мереж, зокрема людського мозку. Це зумовлює зростання попиту на нейромережеві технології, які дедалі активніше впроваджуються у різні галузі та сфери застосування.

Ключові слова: Neuralnet, лінійна регресія, машинне навчання, штучна нейронна мережа, нормалізація даних, середньо квадратичне відхилення (MSE).

Постановка проблеми. Лінійна регресія є одним із найбільш поширених і зрозумілих методів прогнозування, але її можливості обмежені у моделюванні складних нелінійних залежностей. Натомість нейронні мережі здатні працювати з такими залежностями завдяки своїй багатошаровій архітектурі та здатності до нелінійного перетворення даних. Проте використання нейронних

мереж вимагає значних обчислювальних ресурсів і складної конфігурації, що може бути не завжди виправданим.

Таким чином, постає проблема оцінки точності та ефективності цих двох підходів у контексті реальних задач прогнозування. Необхідно визначити, у яких випадках лінійна регресія є достатньою, а в яких нейронна мережа виправдовує свої

обчислювальні витрати. Для цього важливо провести порівняння моделей за єдиним критерієм, наприклад, середнім квадратичним відхиленням (MSE), яке дозволяє оцінити відхилення прогнозованих значень від реальних.

Рішення цієї проблеми сприятиме більш обґрунтованому вибору моделей у задачах прогнозування та значущості як для наукових досліджень, так і для прикладного використання в різних сферах.

Аналіз останніх досліджень і публікацій. Теоретичні аспекти машинного навчання та аналізу даних для прогнозування були досліджені різними вітчизняними та зарубіжними науковцями.

У статті [1, с. 25-30] розглядаються різні методи машинного навчання, що застосовуються для прогнозування, кожен з яких охарактеризовано за обраними критеріями з метою порівняння їх ефективності у прогнозуванні.

Робота [2] надає доступне введення в основні концепції машинного навчання. Автор розглядає різні типи машинного навчання, а також основні етапи процесу машинного навчання, такі як підготовка даних, вибір моделі, навчання, оцінка та впровадження. Дослідження також підкреслює важливість розуміння основних понять і процесів машинного навчання для ефективного застосування цих методів у реальних задачах.

Автори книги [3, с. 30] надають детальне введення в основи проектування та навчання нейронних мереж, крім того, особливу увагу приділено методам навчання для як прямих мереж (включаючи багатосарові та радіально-базисні мережі), так і рекурентних мереж.

У дослідженні [4, с. 83] розглядається застосування машинного навчання та аналізу даних, як ключових інструментів для прогнозування із застосування регресійного аналізу в програмному середовищі R-Studio.

Ряд авторів стверджують, що Neuralnet є оптимальним методом для моделювання будь-яких неперервних та нелінійних функцій, які не припускають жодних умов до вхідної інформації [5, с. 8].

Розроблений авторами С. Фрітчем, Ф. Гюнтером, М. Н. Райтом, М. Сулінгом і С. М. Мюллером документ [6, с. 1-15] є технічною документацією до пакету Neuralnet для моделювання даних та роботи з нейронними мережами в середовищі R.

Постановка завдання. Метою статті є порівняння результатів прогнозування отриманих із використанням лінійної регресії та нейронної мережі. Оцінити точність кожної із моделей опра-

цювання даних, створеної з використанням різних технологій, за результатами обчислення середнього квадратичного відхилення (MSE). Оцінити наскільки прогнозовані результати далекі від реальних даних.

Виклад основного матеріалу. Нейронна мережа (або штучна нейронна мережа, ШНМ) може бути реалізована у вигляді програмного або апаратного забезпечення. Її будують за аналогією з біологічними нейронними мережами, які складаються з нервових клітин живих організмів. Основна ідея полягає в моделюванні процесів, що відбуваються в мозку.

Оцінка моделі є ключовим етапом у процесі її розробки. Для моделей, метою яких є прогнозування, MSE є надійним індикатором точності. Вона показує, наскільки близько лінія регресії проходить до набору точок даних. MSE виступає функцією ризику, яка вимірює втрати на основі квадратичної похибки.

Обчислення MSE передбачає визначення середнього значення квадратів різниць між спостережуваними та передбаченими значеннями. Це метрика, яка кількісно оцінює середню квадратну різницю між цими значеннями. У випадку, коли модель ідеальна, значення MSE дорівнює 0. Однак зі зростанням помилок у моделі зростає і значення MSE. Середньоквадратичну похибку також часто називають середньоквадратичним відхиленням (MSD). Цей показник є одним із ключових критеріїв для оцінки точності алгоритмів прогнозування.

У 2022 та 2023 роках відбувався справжній бум у розвитку генеративного штучного інтелекту. Стрічками технологічних медіа ширилися новини про ChatGPT та Midjourney, багато технологічних гігантів інтегрували AI у свої продукти, а дослідники з OpenAI забили на сполох, попередивши правління про небезпеку, яку становить нове відкриття у галузі штучного інтелекту. Велику роль у розвитку нейронних мереж, зокрема двох найпопулярніших, зіграла архітектура нейронних мереж Transformer. Розроблена в Google для машинного перекладу, вона замінила не настільки ефективну архітектуру LSTM.

Сучасні середовища програмування забезпечують користувачів надійними засобами аналізу даних. Проте, щоб продемонструвати переваги нейронних мереж при опрацюванні даних ми пропонуємо використати середовище RStudio. Мова програмування R має достатньо засобів для проведення регресійного аналізу даних через використання добре відпрацьованих алгоритмів,

а також з використанням внутрішніх функцій та спеціалізованої бібліотеки нейронної мережі.

У відкритих ресурсах мережі Інтернет можна отримати доступ до великої кількості різнопланових даних. Для отримання даних ми використали платформу [7]. Дані на цьому сайті зберігаються у форматі csv файлів. Запропонований набір даних можна використати для побудови моделі та дослідницького аналізу та вивчення робочих процесів машинного навчання.

На наступному етапі опрацювання було розділено дані на навчальну та тестову частини. Це дозволило сформулювати модель лінійної регресії на навчальній сукупності та перевірити її на тестовому наборі. Чітких вимог, щодо пропорцій сформованих таким чином наборів даних, не існує. В ході дослідження проводились обчислення із декількома наборами даних. Суттєвого впливу на кінцеві результати не спостерігалось.

На етапі побудови моделі лінійної регресії використовувалась внутрішня функція RStudio. Параметрами функції `lm()` є перелік імен полів на яких будується лінійна регресія та масив даних. В цій частині роботи обчислювали прогнозовані результати та знаходили значення параметра MSE.

```
lm.fit <- glm(Grades ~ Socioeconomic.Score +
Study.Hours + Sleep.Hours + Attendance...,
data=real_data)
summary(lm.fit)
pr.lm <- predict(lm.fit, test_data)
pr.lm
MSE.lm <- sum((pr.lm - test_data$Grades)^2)/
nrow(test_data)
MSE.lm
```

Результатом роботи цієї частини алгоритму є значення MSE для моделі реалізованої на використанні функції `lm()`.

```
> MSE.lm
[1] 20.28519
```

Однією з найважливіших процедур при формуванні нейронної мережі є нормалізація даних. Це передбачає коригування даних до загального масштабу, щоб точно порівнювати прогнозовані та фактичні значення. Опрацювання не нормалізованих даних призводить до того, що прогнозоване значення залишається незмінним для всіх спостережень, незалежно від вхідних значень.

З метою є підвищення точності прогнозування та нівелювання впливати окремих даних на

передбачення через великий діапазон числових значень проводиться процедура нормалізації або масштабування. Мова програмування R дозволяє провести нормалізацію двома способами – масштабування за допомогою функції мови R, або перетворення даних за допомогою техніки `max/min` нормалізації.

Технологія нормалізації `min-max` має певні недоліки, вона має тенденцію доводити дані до середнього значення. Щоб досягти кращих результатів використана також вбудована функція масштабування

```
dfNormZ <- as.data.frame( scale(df[1:2] ))
```

В середовищі RStudio для побудови нейронних мереж можна використовувати велику кількість пакетів. Серед них можна виділити такі, як `Neuralnet` та `Nnet`. Деякі автори рекомендують створювати моделі нейромереж із застосуванням пакетів `AMORE`, `kohonen` та інші. Пакет `Nnet` при побудові нейромережі дозволяє використовувати тільки один проміжний шар. Навчання нейронної мережі проводилось із різними значеннями параметрів `hidden` та `linear.output`.

```
library(neuralnet)
n <- names(real_data_nor)
f <- as.formula(Grades ~ Socioeconomic.Score +
Study.Hours + Sleep.Hours + Attendance...)
nn <- neuralnet(f, data = real_data_nor,
hidden = c(6,4,2), linear.output=T)
plot(nn))
```

Параметр нейронної мережі `hidden` задає конфігурацію внутрішніх шарів моделі. Цей параметр визначає кількість нейронів в кожному прихованому шарі. Наприклад, `hidden = c(6, 4, 2)` означає, що в першому прихованому шарі є 6 нейронів, у другому – 4, а в третьому – 2. Встановленої методики для розрахунку кількості скритих шарів нейронної моделі не існує. Оптимізуючи параметри розробленої мережі за параметром MSE вводились параметри (5,3,2), (5,4,2), (6,4,3). Зміни параметра `hidden` відображались на показниках помилки (`Error`) та кількості кроків (`Steps`). Для формування кінцевих результатів дослідження встановлено `hidden = c(6,4,2)`. Графічне відображення отриманої моделі штучної мережі, розробленої засобами RStudio, подано Рис. 1.

Чорні лінії відображають зв'язки між кожним шаром і вагами на кожному з'єднанні, тоді як сині

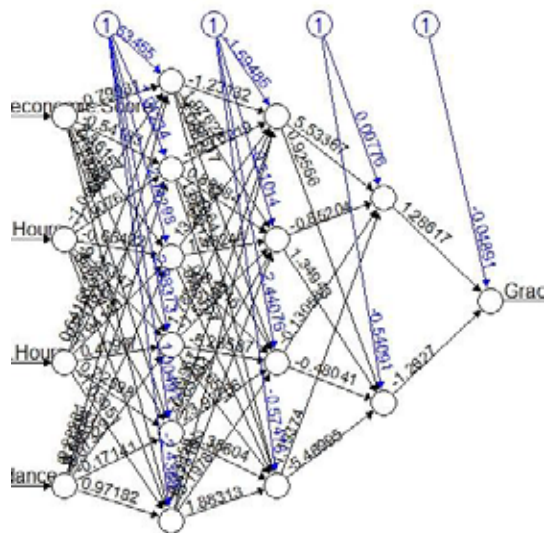


Рис. 1. Графічне відображення моделі штучної мережі

лінії вказують на доданий на кожному кроці член зміщення. Зміщення можна розглядати як **перехоплення** (intercept) лінійної моделі.

Параметр функції `linear.output=T` задає вимогу побудувати регресійну функцію у вигляді лінійного рівняння. Зміна цього параметру на `linear.output=F` не змінює суттєво характеристик ефективності моделі побудованої на конкретних даних нашого дослідження.

Мережа є, по суті, «чорною скринькою», тому детально описати процес навчання неможливо. За результатами навчання можна стверджувати, що алгоритм навчання успішно збігся, а отже, модель готова до використання.

Для оцінки ефективності досліджуваної нейронної мережі, побудованої із використанням бібліотеки `Neuralnet`, обчислено MSE використовуючи нормалізовані дані.

```
pr.nn <- compute(nn,test_data_nor[,1:4])
pr.nn_ <- pr.nn$net.
result*(max(mydani$Grades)-min(mydani$Grades))
+min(mydani$Grades)
test_data.r <- (test_data_nor$Grades)*(max(mydani$Grades)-
min(mydani$Grades)) +min(mydani$Grades)
MSE.nn <- sum((test_data.r - pr.nn_)^2)/
nrow(test_data_nor)
print(paste('MSE_line_funk=',MSE.lm,', MSE_
Naure_net=', MSE.nn))
```

За результатами роботи цієї частини дослідження отримано важливі показники для оцінки розроблених моделей:

«MSE_line_funk=20.285186341023, MSE_Naure_net=1.34815824403662»

При значній величині MSE точки даних сильно розкидані від центрального моменту. Якщо дані щільно групуються навколо середнього значення то MSE буде невеликим (середнім) значенням. У такому випадку значення даних розподілені нормально, менше помилок, відсутня асиметрія. Менше значення MSE вказує на менші похибки та краще оцінювання результатів прогнозування.

За нашим дослідженням нейронна мережа перевершує лінійну модель у прогнозуванні Grades. Такий результат залежить від обраного способу поділу даних на навчальну та тестову вибірки.

Для підтвердження результатів дослідження проведемо швидку перехресну перевірку отриманих моделей, будемо графіки прогнозованих даних (Рис. 2), отриманих з використанням лінійної регресії та моделі штучної нейронної мережі та проведемо їх якісний аналіз.

```
par(mfrow=c(1,2))
plot(test_data$Grades, pr.nn_, col='blue',
main='Результати мережі', pch=20, cex=0.5)
abline(reg = lm(test_data$Grades ~ pr.nn_,
mydani))
legend('bottomright', legend='Naure_
net',pch=18,col='blue', bty='n')
plot(test_data$Grades, pr.lm, col='green',
main='Результати лінійн._функц', pch=20,
cex=0.7)
abline(reg = lm(test_data$Grades ~ pr.lm,
mydani))
legend('bottomright', legend='Line_f',
pch=20,col='green', bty='n', cex=.95)
plot(test_data$Grades, pr.nn_,
col='blue',main='Мережа', pch=20,cex=0.7)
points(test_data$Grades,pr.lm,col='green',
pch=18,cex=0.7)
abline(reg = lm(test_data$Grades ~ pr.nn_,
mydani))
legend('topright', legend=c('Naure_net',
'Line_f'), pch=18,col=c('blue','green'))
```

Ліва частина Рис. 2 відображає сукупність прогнозованих результатів отриманих при використанні моделі нейронної мережі побудованої із використанням бібліотеки `Neuralnet`. Точки щільно прилягають до лінії регресійного рівняння, візуально підтверджують лінійну залежність. Середнє квадратичне відхилення обчислення становить 1.34815824403662.

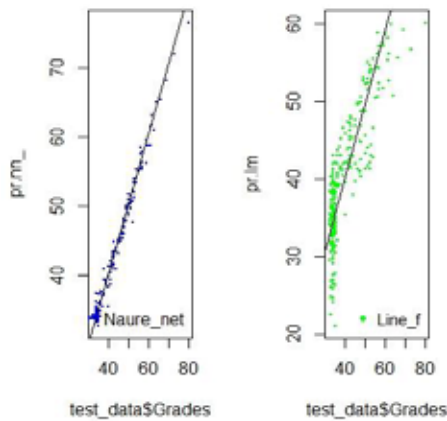


Рис. 2. Графіки прогнозованих даних отриманих з використанням моделі штучної нейронної мережі та лінійної регресії

Множини точок правої частини Рис. 2 відображає прогнозовані результати отримані за допомогою моделі побудованої із використанням вну-

трішньої функції lm . На графіку спостерігається значне розсіювання та слабо виражена лінійна залежність. Середнє квадратичне відхилення обчислення становить 20.285186341023.

Висновки. У статті було проведено порівняльний аналіз ефективності прогнозування даних за допомогою двох підходів: лінійної регресії та нейронної мережі, реалізованих із використанням бібліотеки *Neuralnet*. Отримані результати дозволяють зробити висновки, що нейронна мережа продемонструвала вищу точність прогнозування у випадках, коли вхідні дані мали складну структуру та містили нелінійні залежності. Водночас лінійна регресія краще працює на простих наборах даних із лінійними зв'язками між змінними.

Практика оцінки ефективності моделей прогнозування передбачає використання ряду методик, алгоритмів та технологій, що може стати завданням наступних досліджень.

Список літератури:

1. Бажан Т. О. Порівняльний аналіз методів машинного навчання для побудови прогнозів. *Сучасний захист інформації*. 2024. № 4(60). С. 125–130. <https://doi.org/10.31673/2409-7292.2024.040013>
2. Brownlee J. *Data Preparation for Machine Learning: data cleaning, feature selection, and data*. Machine Learning Mastery. 2020. 398p. URL: <https://machinelearningmastery.com>
3. Martin T. Hagan, Howard B. Demuth, Mark Hudson Beale, Orlando De Jesús *Neural network design*. 2014. 800p.
4. Пінцак І. Використання машинного навчання та аналізу даних для прогнозування тенденцій у електронній комерції. *Information Technology: Computer Science, Software Engineering and Cyber Security*. 2024. № 1. С. 80–88.
5. Xiang Li, Ding Qian, Sun J.Q. Remaining Useful Life Estimation in Prognostics Using Deep Convolution Neural Networks. *Reliability Engineering and System Safety Elsevier*, vol. 172(C). p.1-11.
6. Stefan Fritsch, Frauke Guenther, Marvin N. Wright, Marc Suling, Sebastian M. Mueller Package 'neuralnet'. 2022. 15p. URL: <https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf>
7. Data science competition platform and online community for data scientists and machine learning practitioners under Google LLC. URL: <https://www.kaggle.com/> (дата звернення: 06.01.2025)

Yasinsky A.M., Dzhun I.V., Kundos M.G., Solovei L.Ya. COMPARATIVE ANALYSIS OF FORECASTING USING LINEAR REGRESSION AND NEURAL NETWORK USING THE NEURALNET LIBRARY

The article presents a comparative analysis of two approaches to data forecasting: linear regression and a neural network created using the Neuralnet library. The study is aimed at assessing the accuracy of each model's predictions using the mean square deviation as the main performance indicator. The influence of the neural network architecture, training parameters, and the size and characteristics of the input data on the accuracy of the forecast is studied. Particular attention is paid to comparing the deviations of the predicted results from real data, which allows assessing the practical value of each model for tasks requiring high accuracy. The research methodology includes building and training models on the same data sets, which ensures the correctness of the comparison. For linear regression, traditional statistical approaches to modeling the dependence between variables were used, while for the neural network, a multilayer perceptron with different numbers of hidden layers and neurons was implemented. The analysis of the results was carried out taking into account changes in the accuracy of the models depending on the complexity of their architecture and the computational resources required for training.

The results obtained demonstrate that the neural network can achieve higher accuracy in predicting complex nonlinear dependencies, while linear regression is more effective for problems with linear relationships between variables. It is also noted that the computational costs of training a neural network significantly

exceed the costs for linear regression, which should be taken into account when choosing a method for real-world problems. The results of the study can be useful for specialists in the field of data analysis, machine learning and software development, as well as for students and scientists studying forecasting methods.

Rapid progress in computer technology paves the way for the creation of sixth-generation neurocomputers. Artificial intelligence experts predict that these systems will operate on principles similar to those of biological neural networks, including the human brain. This is driving the demand for neural network technologies, which are increasingly being implemented in various industries and applications.

Key words: *Neuralnet, linear regression, machine learning, artificial neural network, data normalization, mean square error (MSE).*