УДК 004.8

ВПЛИВ ШУМУ ВХІДНИХ ДАНИХ НА РЕЗУЛЬТАТИ ПРОГНОЗУВАННЯ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Пилипенко В.І., старший викладач
Київський національний університет технологій та дизайну
Антонюк Ю.Д., студентка
Київський національний університет технологій та дизайну
Цітелашвілі О.В., студентка
Київський національний університет технологій та ди

Ключові слова: машинне навчання, прогнозування, точність моделі, шум вхідних даних, фільтрація шуму, ансамблеві моделі, RMSE.

Сучасний етап розвитку інтелектуальних систем характеризується експоненційним зростанням обсягів даних, а також підвищенням їхньої гетерогенності та рівня зашумленості. У таких умовах машинне навчання (ML) виступає ключовим інструментом для виявлення латентних закономірностей і побудови прогнозних моделей. Дослідження показують, що при зростанні рівня шуму на кожні 10%, середньоквадратична помилка (RMSE) регресійних моделей зростає в середньому на 3,2%, а точність класифікаторів RandomForest знижується на 3–5% [1]. Критичним аспектом є диференціація чутливості архітектур до спотворень. Наївний байєсівський класифікатор демонструє відносну стабільність при високому рівні шуму, тоді як складні нейронні мережі потребують жорсткої регуляризації. Встановлено, що моделі типу LSTM зберігають високу точність лише при відношенні сигналу до шуму (SNR) вище -12 дБ, тоді як падіння показника до -26 дБ спричиняє обвал ефективності до 65% [2]. Мінімізація впливу неінформативних параметрів досягається через препроцесинг (PCA, методи фільтрації) та використання ансамблевих підходів. Застосування технік селекції ознак дозволяє стабілізувати результати, забезпечуючи точність навіть у складних датасетах. Системне поєднання адаптивних алгоритмів та процедур крос-валідації є необхідною умовою створення робастних систем, здатних надійно функціонувати в умовах високої невизначеності та варіативності реальних даних [3]. Водночас ефективність цих моделей безпосередньо залежить від якості вхідних даних, оскільки реальні датасети практично завжди містять стохастичні викривлення [4]. Шум у даних можна формалізувати як випадкову компоненту ε, що додається до істинного сигналу:

$$Xobserved = Xtrue + \varepsilon, \quad (1)$$

де ε зазвичай моделюється як нормально розподілена величина N(0,σ2).

Зростання дисперсії шуму призводить до збільшення зміщення оцінок (bias), зростання варіації (variance) та погіршення узагальнюючої здатності моделей. Ключова проблема полягає в тому, що шум у вхідних даних є системним фактором деградації якості прогнозування. Навіть незначний рівень зашумлення (5–10%) може спричинити суттєве зміщення статистичних оцінок і спотворення ваг ознак. Це породжує класичну дилему: недонавчання (через маскування інформативного сигналу) або перенавчання (overfitting), коли модель інтерпретує випадкові коливання як закономірності особливо коли питання полягає в точності прогнозування моделі [5].

Емпіричні дослідження показують, що залежність похибки від рівня шуму має квазі-лінійний характер. Зокрема, середньоквадратична помилка (RMSE) може бути апроксимована як:

$$RMSE(\alpha) \approx RMSE0(1 + k \cdot \alpha), \quad (2)$$

де α – рівень шуму, k ≈ 0.032

Збільшення шуму на кожні 10% призводить у середньому до зростання RMSE приблизно на 3,2%, тоді як точність класифікаційних моделей (зокрема ансамблевих, таких як RandomForest) знижується на 3–5%. На рисунку 1 представлено вплив шуму на точність моделі.

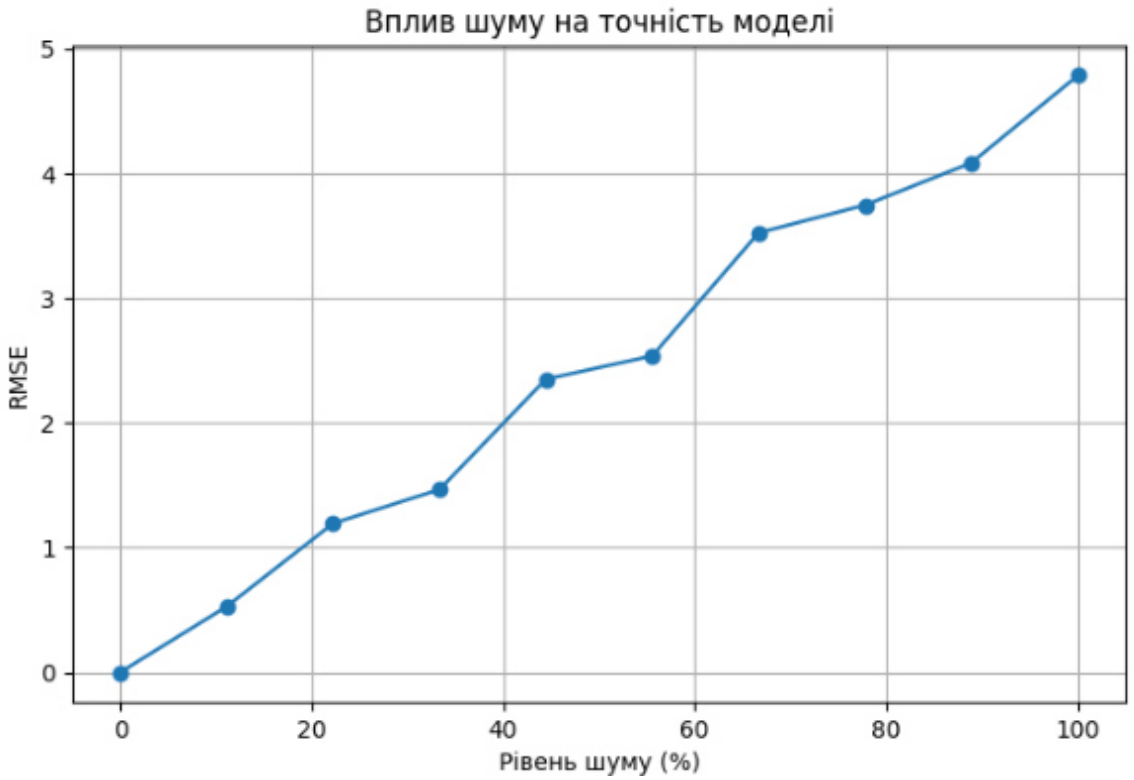

Рисунок 1 - Вплив шуму на точність моделі

Важливим результатом є підтвердження того, що застосування методів зменшення шуму, зокрема зниження розмірності та видалення викидів, суттєво підвищує ефективність застосування моделей.

Підвищення робастності моделей досягається за рахунок застосування методів попередньої обробки даних:

1. зменшення розмірності (PCA)
2. фільтрація шуму
3. виявлення викидів (наприклад, Isolation Forest, Z-score)
4. селекція ознак

Комбіноване використання цих підходів дозволяє мінімізувати вплив як атрибутивного, так і класового шуму. Додатково, застосування ансамблевих методів і процедур крос-валідації сприяє стабілізації результатів і підвищенню узагальнюючої здатності моделей.

Список використаних джерел

1. Пилипенко В. (2025). Прогнозування високого рівня академічної успішності студентів з використанням машинного навчання. Наука і техніка сьогодні, 8(45), 1634–1649. https://doi.org/10.52058/2786-6025-2025-8(49)-1634-1649

2. Saseendran, Arun & Setia, Lovish & Chhabria, Viren & Chakraborty, Debrup & Barman Roy, Aneek. (2019). Impact of Noise in Dataset on Machine Learning Algorithms. 10.13140/RG.2.2.25669.91369.

3. Abhinav Atla, Rahul Tada, Victor Sheng, and Naveen Singireddy. 2011. Sensitivity of different machine learning algorithms to noise. J. Comput. Sci. Coll. 26, 5 (May 2011), 96–103.

4. Стаценко, В., & Пилипенко, В. (2024). Оцінювання ефективності моделі прогнозування успішності методами машинного навчання. Herald of Khmelnytskyi National University. Technical sciences, 331(1), 271-276, https://doi.org/10.31891/2307-5732-2024-331-41

5. S. Volodymyr, V. Pylypenko, V. Skidan and A. Volivach, "Investigation of the Accuracy of Machine Learning Methods in Prediction of Students Success," 2024 IEEE 5th KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2024, pp. 1-4, doi: 10.1109/KhPIWeek61434.2024.10877975.