

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

**Чорновол Олена Вікторівна**

УДК 004.89

**ІНТЕЛЕКТУАЛЬНА СИСТЕМА ПРОГНОЗУВАННЯ  
ЕНЕРГОВИРОБНИЦТВА АЕС**

124 – Системний аналіз

Автореферат  
магістерської наукової роботи на здобуття освітньої кваліфікації  
«Магістр системного аналізу»

Миколаїв – 2020

Магістерська наукова робота є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: д.т.н., професор, професор кафедри інтелектуальних інформаційних систем  
Кондратенко Юрій Пантелійович

Рецензент: к.ф.-м.н., доцент кафедри інженерії програмного забезпечення  
Пузирьов Сергій Володимирович

Захист відбудеться «26» лютого 2020 р. о 9<sup>30</sup> год. на засіданні екзаменаційної комісії (ауд. 2-403) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

З магістерською науковою роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «\_\_\_» лютого 2020 р.

Секретар  
екзаменаційної комісії,  
к.пед.н., доцент

Н. М. Болубаш

## ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

*Актуальність* дослідження визначається складністю розрахунку дисбалансів попиту енергії та низький рівень управління енергоефективністю атомних електростанцій.

*Метою* магістерської наукової роботи є прогнозування енерговиробництва на прикладі атомної електроенергії в реальному часі методами машинного навчання.

*Об'єктом* дослідження є інтелектуальна система прогнозування енерговиробництва АЕС методами машинного навчання.

*Предметом* дослідження є аналіз вибраних моделей машинного навчання для прогнозування енергії: множинна лінійна регресія (MLR), опорно-верторна регресія (SVR), модель випадкового лісу (Random Forest) та модель екстремального градієнтного бустінгу (XGBoost).

*Практичне значення* даної магістерської наукової роботи полягає у можливості застосування моделей машинного навчання для інформаційно-аналітичної системи в галузі атомної енергетики.

Результати даної магістерської наукової роботи було надруковано у тезах науково-методичної конференції «Могилянські читання – 2019» у секції Комп'ютерні науки, а також у тезах Всеукраїнської науково-практичної конференції «Інтелектуальні інформаційні системи – 2020».

Магістерська наукова робота складається із вступу, 6 розділів, висновків, додатків. Загальний обсяг роботи складає 139 сторінки, 48 рисунків, 11 таблиць та 85 посилань на літературні джерела.

## ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі магістерської наукової роботи обґрунтовано актуальність обраної теми, сформульовано мету і задачі дослідження, визначено предмет та об'єкт дослідження.

У першому розділі наведено огляд атомної енергетики в Україні, подано економічну оцінку методом розрахунку Levelised Cost of Energy (LCOE), та виокремлені перспективи розвитку галузі. Аналіз сектору показав, що розвиток атомної енергетики є перспективним напрямом у досягненні цілей Паризької угоди, а саме підвищення енергоефективності, енергозбереження, впровадження сучасних методів енергоменеджменту, енергоаудиту.

У другому розділі здійснено опис машинного навчання для регулювання енергії, наведено узагальнену класифікацію моделей машинного навчання та особливостей їх використання. Охарактеризовано гіперпараметри моделей машинного навчання та їх налаштування за допомогою GridSearchCV.

Технології генерації електроенергії сильно залежать від умов навколишнього середовища, і підлягають плановому та позаплановому обслуговуванню. Завдання оператора енергомережі полягає в тому, щоб вирішити дефіцит наявних ресурсів порівняно з фактичним попитом. Існує три рішення щодо дефіциту електроенергії: побудувати більше електростанцій на базовому навантаженні (цей процес може зайняти багато років і десятиліть планування та будівництва), придбати та імпортувати електроенергію з інших регіональних енергосистем (цей вибір може бути дуже дорогим і обмежений передачею електроенергії з енергосистем з надлишковою потужністю), або підключення невеликих станцій максимальної потужності. Оскільки операторам мережі потрібно швидко реагувати на дефіцит електроенергії, щоб уникнути відключення електроенергії, оператори мережі покладаються на комбінацію останніх двох варіантів. Методи машинного навчання допомагають прогнозувати дефіцит енергії та покращити регулювання енерговиробництва.

Системи машинного навчання можна розділити на більш лінійні (high bias), які мають порівняно малу здатність до формування складних інтерполяційних

кривих, і системи з високою варіативністю (high variance), які здатні формувати криві складної форми. Поведінка цих моделей різниться при збільшенні числа тренувальних прикладів. Перші, як правило, узагальнюють результати, часто не враховуючи деяких, можливо істотних, відмінностей між навчальними прикладами. Другі, навпаки, «відстежують» всі нюанси, можливо випадкові, але, в той же час, недостатньо узагальнюють. Для перших характерна недотренованість, в той час як для других перетренованість. Оцінити здатності моделі при одноразовому експерименті, як правило, неможливо, оскільки і перші, і другі можуть давати близькі показники помилок.

Коефіцієнт кореляції Пірсона ( $r$ -Пірсона) застосовується для дослідження взаємозв'язку двох змінних, вимірних в метричних шкалах на одній і тій же вибірці. Він дозволяє визначити, наскільки пропорційна мінливість двох змінних.

Для задач енергоменеджменту найчастіше використовують середнє квадратичне відхилення, яке так само, як і середнє лінійне відхилення, показує, на скільки в середньому відхиляються конкретні значення ознаки від середнього їх значення. Середнє квадратичне відхилення завжди більше середнього лінійного відхилення. Мале значення стандартного відхилення вказує, що дані точок скупчені ближче до середнього значення вибірки, в той час як великі значення стандартного відхилення вказують, що точки розподілені в більш широкому діапазоні значень.

Крім того, що стандартне відхилення характеризує мінливість вибірки, воно зазвичай використовується як міра достовірності статистичних висновків. Наприклад, межа похибки для даних опитування визначається за допомогою розрахунку очікуваного стандартного відхилення в результатах за умови, якби те саме опитування було проведене декілька разів. Таке виведення стандартного відхилення часто називають «стандартною похибкою» оцінювання або «стандартною похибкою середнього», якщо мова йде про середнє. Вона визначає стандартне відхилення усіх середніх значень, отримані для даної генеральної сукупності на основі вибірки.

Програмування моделей машинного навчання у дослідженні відбувається мовою Python. Вбудовані структури даних високого рівня в поєднанні з динамічним

набором тексту та динамічним зв'язуванням роблять її дуже привабливою для швидкого розвитку додатків, а також для використання в якості мови сценаріїв для з'єднання існуючих компонентів разом. Python має повністю динамічну систему типу і використовує автоматичне управління пам'яттю. Вона підтримує об'єктивізовані, імперативні, функціональні та процедурні парадигми програмування.

Бібліотека Prophet надає можливість робити прогнози часових рядів з хорошою точністю, використовуючи прості інтуїтивно зрозумілі параметри, і підтримує включення впливу спеціальної сезонності та свят. Коли модель прогнозування не працює, як планувалося, ми хочемо мати можливість налаштувати параметри методу відповідно до конкретної проблеми. Налаштування цих методів вимагає глибокого розуміння того, як працюють основні моделі часових рядів. Наприклад, першими вхідними параметрами для автоматизованої ARIMA є максимальні порядки розрізнення, авторегресивні компоненти та компоненти середнього значення.

Модель розкладного часового ряду може використовуватися з трьома основними компонентами: тенденція, сезонність та свята. Вони поєднуються в рівнянні:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t,$$

де  $g(t)$  – кусково-лінійна або логістична крива зростання для моделювання неперіодичних змін у часових рядах;  $s(t)$  – періодичні зміни (наприклад, щотижнева/річна сезонність);  $h(t)$  – наслідки відпусток (надано користувачем) з нерегульованим графіком;  $\varepsilon_t$  – термін помилки пояснює будь-які незвичайні зміни, не пристосовані моделлю.

Використовуючи час в якості регресора, Prophet намагається встановити декілька лінійних і нелінійних функцій часу як складових. Моделювання сезонності – це той самий підхід, що застосовується експоненціальним згладжуванням у техніці Холт-Вінтерса. Насправді ми задаємо проблему прогнозування як вправу на

встановлення кривих, а не чітко дивимось на залежність кожного спостереження за часовим рядом, залежно від часу.

Тенденція моделюється за рахунок підгонки детальної лінійної кривої над трендом або неперіодичною частиною часового ряду. Лінійність ряду показує мінімальний вплив зміни даних. Маючи в своєму розпорядженні знання про домен, аналітик може потім визначити змінну потужність  $C(t)$  для прогнозів часових рядів, які він намагається зробити.

Для відповідності та прогнозування наслідків сезонності Prophet покладається на серію Фур'є, щоб забезпечити гнучку модель. Сезонні ефекти  $s(t)$  апроксимуються наступною функцією:

$$s(t) = \sum_{n=1}^N (a_n \cos(\frac{2\pi n t}{P}) + b_n \sin(\frac{2\pi n t}{P})),$$

де  $P$  – це період (365 для щорічних даних та 7 для тижневих даних). Параметри  $[a_1, b_1, \dots, a_N, b_N]$  необхідно оцінити для даного  $N$  для моделювання сезонності.

Порядок фур'є  $N$ , який визначає, чи дозволяється моделювати зміни високої частоти, є важливим параметром для встановлення тут. Якщо за певний часовий ряд користувач вважає, що високочастотні компоненти - це лише шум, і їх не слід розглядати для моделювання, він / вона може встановити значення  $N$  від нижчого значення. Якщо ні,  $N$  можна налаштувати на більш високе значення та встановити, використовуючи точність прогнозу. Функція Prophet () використовується для визначення моделі прогнозування Prophet в Python.

**В третьому розділі** розглянуто класифікацію моделей машинного навчання: множинна лінійна регресія (MLR), опорно-векторна регресія (SVR), модель випадкового лісу (Random Forest) та модель екстремального градієнтного бустінгу (XGBoost).

**Метод опорних векторів (SVR)** це алгоритм машинного навчання, який може бути використаний як для вирішення проблем класифікації, так і для регресії. Однак він здебільшого використовується в проблемах класифікації. У цьому алгоритмі побудується кожен елемент даних як точка в  $n$ -мірному просторі (де  $n$  -

кількість функцій), причому значення кожної функції є значенням певної координати. Потім ми проводимо класифікацію, знаходячи гіперплощину, яка дуже добре розмежує два класи. Вектори підтримки - це просто координати індивідуального спостереження. Векторна машина підтримки - це межа, яка найкраще розділяє два класи (гіперплощина / лінія).

Незважаючи на те, що менш популярний, ніж SVM, алгоритм SVR виявився ефективним інструментом для оцінки функцій у реальному значенні. Як підхід під контрольним навчанням, SVR тренується, використовуючи симетричну функцію втрат, яка однаково карає високі та низькі оцінки. Використовуючи  $\epsilon$ -підхід Вапніка, гнучка трубка мінімального радіуса формується симетрично навколо оціночної функції, так що абсолютні значення помилок, менші за певний поріг  $\epsilon$  ігноруються як вище, так і нижче оцінки. Таким чином, точки поза трубкою накладаються штрафними санкціями, але ті, що знаходяться в трубці, вище або нижче функції, не отримують штрафних санкцій. Однією з головних переваг SVR є те, що його обчислювальна складність не залежить від розмірності вхідного простору. Крім того, він має чудові можливості узагальнення, з високою точністю прогнозування.

SVR формулює задачу наближення цієї функції як проблему оптимізації, яка намагається знайти найвужчу трубку, зосереджену навколо поверхні, мінімізуючи помилку прогнозування, тобто відстань між прогнозованим та бажаним результатами.

$$y = f(x) = \langle w, x \rangle + b = \sum_{j=1}^M w_j x_j + b, y, b \in R, x, w \in R^M$$

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ l \end{bmatrix} = w^T x + b, x, w \in R^{M+1}.$$

Колишня умова виробляє об'єктивну функцію в рівнянні, де  $\|w\|$  - величина нормального вектора до наближеної до поверхні:

$$\min_w \frac{1}{2} \|w\|^2.$$

Щоб уявити, як величину ваг можна інтерпретувати як міру плоскості, розглянемо наступний приклад:

$$f(x, w) = \sum_{j=1}^M w_j x^j, x \in R, x, w \in R^M,$$



де  $M$  - порядок многочлена, який використовується для наближення функції. Зі збільшенням величини вектора  $w$  більша кількість  $w_j$  є ненульовою, що призводить до рішень вищого порядку. Горизонтальна лінія є поліноміальним рішенням 0-го порядку і має дуже велике відхилення від бажаних виходів, а значить, велику помилку. Лінійна функція, поліном 1-го порядку, дає кращі наближення для частини даних, але все ще недоотримує навчальні дані. Рішення шостого порядку забезпечує найкращий компроміс між площиною функції та помилкою прогнозування. Рішення найвищого порядку має нульову помилку, але велику складність і, швидше за все, перевищить рішення на даних, які ще не побачать. Величина  $w$  виступає як регуляризуючий термін і забезпечує оптимізацію контролю задачі над площинністю рішення.

Почати опис алгоритму **випадкового лісу (Random Forest)** слід з дерева прийняття рішень, як з основного структурного елементу лісу, адже саме від того, яким чином побудовано кожне дерево, серйозно залежить якість роботи і стійкість всієї фінальної композиції.

Розглянемо розмічену вибірку об'єктів  $\{(x_i, y_i)\}_{i=1}^N$ , де  $x_i \in R^2$  - опис об'єкта в двовимірному просторі, а  $y_i \in \{0,1\}$  - мітка класу. Незважаючи на те, що в середині об'єкти різних класів сильно перемішані, за допомогою дерева рішень з такою вибіркою досить зручно працювати: на кожному кроці необхідно вибирати ознаку і значення порога, за якими відбувається оптимальне по заданому критерію розбиття. При вирішенні прикладних задач часто використовуються наступні критерії:

- для задач класифікації іGain:

$$\sum_{v \in \{L,R\}} \frac{|S_v|}{|S|} H(S_v),$$

$$H(S) = - \sum_{c \in C} p_c \log_2(p_c),$$

де  $C$  - безліч класів даної задачі, а  $p_c$  - ймовірність класу  $c$  для безлічі об'єктів  $S$ ;

- для задач регресії аналогічний іGain критерій з використанням дисперсій:

$$iGain(S) = |S|Var(S) - \sum_{v \in \{L,R\}} |S_v|Var(S_v),$$

де  $Var(S)$  - дисперсія відгуків об'єктів з множини  $S$ .

При кожному діленні всі об'єкти діляться на дві більш дрібні групи, тобто розглядати в кожному з вузлів завдання розбивається на дві більш дрібні підзадачі. Завданням максимального числа об'єктів в вершині-листі дерева встановлюється один з можливих критеріїв зупинки для алгоритму. Таким чином, можна досить якісно класифікувати розглянуту вибірку об'єктів за допомогою всього одного дерева рішень, якщо в якості відповіді для тестового об'єкта, який потрапив в осередок  $A_i$ , видавати номер, що найбільш часто зустрічається в цьому осередку класу.

**Множинна лінійна регресія (MLR)**, відома також як множинна регресія, є статистичною методикою, яка використовує кілька пояснювальних змінних для прогнозування результату змінної відповіді. Метою множинної лінійної регресії (MLR) є моделювання лінійної залежності між пояснювальною (незалежною) змінною та змінною реакції (залежною). Лінійне відношення (або лінійна асоціація) - це статистичний термін, що використовується для опису прямолінійного зв'язку між змінною та постійною. Лінійні зв'язки можуть бути виражені або в графічному форматі, де змінна та константа з'єднані через пряму, або в математичному форматі, де незалежна змінна множиться на коефіцієнт нахилу, доданий на постійну, яка визначає залежну змінну. Лінійне відношення може протиставлятися поліноміальним або нелінійним (кривим) відношенням.

По суті, множинна регресія - це розширення звичайної регресії найменших квадратів (OLS), яка включає більш ніж одну пояснювальну змінну. Звичайна регресія найменших квадратів (OLS) - це статистичний метод аналізу, який оцінює зв'язок між однією або декількома незалежними змінними та залежною змінною; метод оцінює співвідношення шляхом мінімізації суми квадратів у різниці між спостережуваними та передбачуваними значеннями залежної змінної, сконфігурованої у вигляді прямої. У цьому введенні регресія OLS буде обговорюватися в контексті біваріантної моделі, тобто моделі, в якій є лише одна

незалежна змінна (X), яка передбачає залежну змінну (Y). Однак логіка регресії OLS легко поширюється на багатоваріантну модель, в якій є дві або більше незалежних змінних. Формула для множинної лінійної регресії є:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon,$$

де для  $i = n$  спостережень:  $y_i$  - залежна змінна,  $x_i$  - розширювальні змінні,  $\beta_0$  - y-змінна (постійний член),  $\beta_p$  - коефіцієнти нахилу для кожної пояснювальної змінної,  $\epsilon$  - термін помилки моделі (також відомий як залишки).

Лінійна (OLS) регресія порівнює відповідь залежної змінної з урахуванням зміни деякої пояснювальної змінної. Однак рідко буває, що залежна змінна пояснюється лише однією змінною. У цьому випадку аналітик використовує множинну регресію, яка намагається пояснити залежну змінну, використовуючи більш ніж одну незалежну змінну. Множинні регресії можуть бути лінійними та нелінійними.

Множинні регресії засновані на припущенні, що існує лінійна залежність як залежної, так і незалежної змінних. Вона також не передбачає значної кореляції між незалежними змінними. Таку регресію можна розглядати як численні регулярні лінійні регресійні моделі, оскільки ви просто порівнюєте співвідношення між ознаками для заданої кількості ознак. Передбачається, що існує лінійна залежність між залежною змінною та незалежною змінною. Це також передбачає, що змінні та ознаки - це всі постійні значення, а не дискретні значення.

В основі моделі **XGBoost** лежить алгоритм градієнтного бустінга дерев рішень. Метод бустінгу - це техніка машинного навчання для задач класифікації і регресії, яка будує модель передбачення в формі ансамблю слабких пророчих моделей, зазвичай дерев рішень. Навчання ансамблю проводиться послідовно на відміну, наприклад від беггінга. На кожній ітерації обчислюються відхилення прогнозів вже навченого ансамблю на навчальній вибірці. Наступна модель, яка буде додана в ансамбль буде передбачати ці відхилення. Таким чином, додавши передбачення нового дерева до пророкувань навченого ансамблю ми можемо зменшити середнє відхилення моделі, котре є таргетом оптимізаційної задачі. Нові

дерева додаються в ансамбль до тих пір, поки помилка зменшується, або поки не виконується одне з правил «ранньої зупинки».

Найважливішим фактором успіху XGBoost є його масштабованість у всіх сценаріях. Система працює більш ніж у десять разів швидше, ніж існуючі популярні рішення на одній моделі. Масштабованість XGBoost зумовлена кількома важливими системами та алгоритмічними оптимізаціями. Ці нововведення включають: алгоритм навчання нового дерева, призначений для обробки рідких даних; теоретично обґрунтована зважена квантильна ескізна процедура дозволяє обробляти ваги екземплярів у приблизному вивченні дерева; паралельне та розподілене обчислення робить навчання швидшим, що дозволяє швидше вивчити модель.

Функція для оптимізації градієнтного бустінга:

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t),$$

де  $l$  – функція втрат;  $y_i \hat{y}_i$  – значення  $i$ -го елемента навчальної вибірки і сума пророкувань перших  $t$  дерев відповідно;  $x_i$  – набір ознак  $i$ -го елемента навчальної вибірки;  $f_t$  – функція (у нашому випадку дерево), яку ми хочемо навчити на кроці  $t$ ;  $f_t(x_i)$  – передбачення на  $i$ -му елементі навчальної вибірки;  $\Omega(f)$  - регуляризація функції  $f$ .

Реалізація алгоритму була розроблена для ефективності обчислювальних ресурсів часу і пам'яті. Мета проекту полягала в тому, щоб найкращим чином використовувати наявні ресурси для навчання моделі. Деякі ключові функції реалізації алгоритму включають:

- Різні стратегії обробки пропущених даних.
- Блокова структура для підтримки розпаралелювання навчання дерев.
- Продовження навчання для донавчання на нових даних.

XGBoost - одна з найшвидших реалізацій дерев градієнтного бустінгу. Це обумовлено шляхом подолання однієї з основних неефективностей дерев, що підсилюють градієнт: врахування потенційної втрати для всіх можливих розщеплень для створення нової гілки (особливо якщо врахувати випадок, коли існує тисяча особливостей, а отже, тисячі можливих розщеплень). XGBoost вирішує

цю неефективність, дивлячись на розподіл функцій у всіх точках даних на аркуші та використовуючи цю інформацію для зменшення простору пошуку можливих розбитків функцій. Хоча XGBoost реалізує декілька прийомів регуляризації, це прискорення є, безумовно, найбільш корисною особливістю бібліотеки, що дозволяє швидко вивчити багато гіперпараметрів. Модель машинного навчання визначається як математична модель з низкою параметрів, які потрібно засвоїти з даних. Однак є деякі параметри, відомі як гіперпараметри, які потрібно обрати інтуїтивно для власного тренування. Ці параметри виявляють свою важливість завдяки покращенню продуктивності моделі, такої як її складність або ступінь навчання. Моделі можуть мати багато гіперпараметрів, і пошук найкращого поєднання параметрів може розглядатися як проблема пошуку.

**У четвертому розділі** наводиться формування набору даних для дослідження прогнозу споживання енергії, тренування моделей та опис розробленої інформаційно-аналітичної системи.

Погода з найближчої метеостанції аеропорту (аеропорт міста Миколаїв) була завантажена з публічного набору даних із надійного прогнозу (gr5.ua) та об'єднана разом з експериментальними наборами даних з <https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction>. Дві випадкові змінні були включені в набір даних для тестування регресійних моделей та для фільтрації непередбачуваних атрибутів (параметрів).

Споживання енергії було розглянуто як проблему часових рядів. З цієї причини було реалізовано такі методи часового ряду, такі як ARIMA з статистичних моделей бібліотеки statsmodels та Seasonal Decomposition за допомогою бібліотеки Prophet.

Спостереження набору даних відносяться до періоду, який менший за один рік, тому можна стримувати сезонність. Цю сезонність можна виразити фіктивними змінними (dummy variables), які пояснюють залежну змінну (споживання). Ми використовуємо 3 моделі для тренування наших алгоритмів. Нам потрібно оптимізувати гіперпараметри за допомогою GridSearchCV. З цієї причини ми

розділили початковий набір даних на набори для тренування, валідації та тестування.

Після налаштування гіперпараметрів при встановленні валідація ми робимо першу оцінку і порівнюємо бал середньоквадратичної похибки (RMSE) між моделями. Ми використовуємо Standard Scaler для стандартизації для лінійної регресії. Далі проводимо кілька лінійних регресій, і отримуємо показники моделі прогнозування. Ця модель використовується для валідації даних, для тренування ми будемо використовувати моделі опорно-векторної регресії (SVR), Random Forest, та XGBoost.

Отримавши результати 12 тестів для 4 моделей машинного навчання (Multiple Linear Regression, SVR, Random Forest, XGBoost) формуємо порівняльну таблицю (Таблиця 1) з найкращими показниками для кожної моделі.

Таблиця 1. Порівняння моделей машинного навчання для прогнозування енерговиробництва

Назва моделі	Параметри	Значення	$R^2$	RMSE
Multiple Linear Regression	-	-	0.252	0.569
SVR	regressor__gamma	[0.001, 0.01, 0.1]	0,6	0.434
	regressor__C	[50,60,70]		
Random Forest	n_estimators	[100, 110, 130, 140]	0,61	0.407
	max_depth	[0.001, 0.01, 0.1]		
	max_features	[50,60,70]		
XGBoost	n_estimators	[4,5,6]	0.61	0.443
	max_depth	[100, 110, 130, 140]		
	learning_rate	[0.1, 0.5, 0.7]		

З наведених результатів робимо висновок, що модель випадкового лісу (Random Forest) дає найкращий результат для вирішення задач прогнозування енергоспоживання. Ми вибираємо Random Forest для подальшого тренування нашого набору даних.

Для аналізу отриманого набору даних прогнозованої кількості споживання енергії, та даних видобутку електроенергії на Південноукраїнської АЕС, ми створюємо інтерфейс за допомогою пакету розробки Predix, використовуючи додаток Remote Monitoring & Diagnostics (RMD) Reference. Платформа Predix дозволяє створювати програми з фокусом на промисловому Інтернеті (Industrial Internet) та дозволяє керувати та масштабувати ці програми.

Додаток RMD показує кращу практику архітектури прикладних мікросервісів, що інтегрується з усіма основними службами Predix. Мікросервіси Reference-Front і Back-End демонструють, як використовувати Predix PAAS для створення індустріальної програми для енергоменеджменту та аудиту.

В архітектурному відношенні програма організована у чотири рівні (презентація, доставка, агрегація та зберігання) та підтримує три потоки даних (введення, аналіз, візуалізація):

- Рівень презентації - шар інтерфейсу та мікросервіси;
- Рівень доставки - кешування, мобільний телефон, персоналізація;
- Рівень агрегації - склад послуг та бізнес-логіка;
- Зберігання рівня - послуги Predix PAAS.

**У спеціальній частині** магістерської наукової роботи з «Охорони праці та безпеки життєдіяльності» розглянуто мікрокліматичні умови праці на робочих місцях на предмет виробничого освітлення та дотримання вимог експлуатації ПК в офісі компанії «Template monster». В результаті розрахунків встановлено, що передбаченої кількості вікон, тобто їх загальної площі вистачає для забезпечення вимог санітарних норм щодо природного освітлення приміщення для якого проводився аналіз умов праці. Були запропоновані заходи, що дозволяють знизити навантаження роботи на 15 пунктів, що призведе до загального приросту продуктивності персоналу на 5%. Покращення умов праці персоналу не потребують великих матеріальних витрат, і можуть бути легко введені в дію РМ відділом.

**У методичній частині** розроблено практичні роботи на теми «Класична лінійна регресія» та «Застосування методу DEMATEL для прийняття рішень на основі обробки експертних суджень і нечіткої логіки в середовищі MATLAB».

## ЗАГАЛЬНІ ВИСНОВКИ

У даній магістерській науковій роботі досліджена проаналізовано сучасний стан проблеми аналітики даних, що застосовується для управління атомною енергією. Наведено загальну характеристику розвитку атомної енергетики, яка будучи однією з низьковуглецевих технологій, існуючих на даний момент, здатна допомогти країнам вирішити подвійне завдання – отримання енергії та охорона клімату. При виробленні атомної енергетики не утворюється діоксид вуглецю (CO<sub>2</sub>), тому вона прирівнюється за мінімальним обсягом викидів з відновлювальними джерелами енергії. Інструментами досягнення покращення екологічної ситуації ми виокремили підвищення енергоефективності, впровадження заходів з енергозбереження, впровадження сучасних методів енергоменеджменту, та енергоаудиту. Радикальне покращення показників енергоефективності сприятиме скороченню обсягів споживання енергії, збереженню ресурсів, зменшенню обсягів викидів вуглекислого газу (CO<sub>2</sub>), а також уповільненню зростання цін на електроенергію для промисловості та населення. Розглянуто чотири моделі машинного навчання: множинна лінійна регресія (MLR), опорно-верторна регресія (SVR), модель випадкового лісу (Random Forest) та модель екстремального градієнтного бустінгу (XGBoost). Модель випадкового лісу (Random Forest) показала найкращі результати для задачі прогнозування попиту енергії. Для створення і моделювання було використано мову програмування Python, а також бібліотеки PySpark. Для аналізу та обробки даних був використаний метод EDA (Exploratory Data Analysis). Ми проаналізували дані з ресурсу [gp5.ua](http://gp5.ua) у поєднанні із експериментальними наборами даних з ресурсу [archive.ics.uci.edu](http://archive.ics.uci.edu). Для розрахунку тривалості дня в Миколаївській області було використано ресурс [sunrise-and-sunset.com](http://sunrise-and-sunset.com).

В рамках дослідження для тренування алгоритмів використовується GridSearchCV для оптимізації гіперпараметрів моделей. Дослідивши процес настройки гіперпараметрів з метою визначення оптимальних значень для даних моделей, ми прийшли до висновку, що найбільше продуктивність на заданих значеннях гіперпараметра ґрунтується модель випадкового лісу (Random Forest).



Створено прототип інформаційно-аналітичної системи пакетом розробки, що дозволить експлуатуючим організаціям атомної галузі виробити ефективнішу і діючу на випередження стратегію управління, приймати критичні рішення швидше і точніше завдяки інноваційній зв'язці інституційних знань, аналітичних технологій та машинного навчання, а також оптимізувати логістичні процеси через швидке і легке прогнозування продуктивності та стану активів.

У методичній частині магістерської роботи розроблено дві практичні роботи на теми «Класична лінійна регресія» та «Застосування методу DEMATEL для прийняття рішень на основі обробки експертних суджень і нечіткої логіки в середовищі MATLAB».

У спеціальній частині магістерської роботи з «Охорони праці та безпеки в надзвичайних ситуаціях» розглянуто мікрокліматичні умови праці на робочих місцях на предмет виробничого освітлення та дотримання вимог експлуатації ПК в офісі компанії «Template Monster». В результаті розрахунків встановлено, що передбаченої кількості вікон, тобто їх загальної площі вистачає для забезпечення вимог санітарних норм щодо природного освітлення приміщення, для якого проводився аналіз умов праці. Були запропоновані заходи, що дозволяють знизити навантаження роботи на 15 пунктів, що призведе до загального приросту продуктивності персоналу на 5%. Покращення умов праці персоналу не потребують великих матеріальних витрат, і можуть бути легко введені в дію РМ відділом.

## АНОТАЦІЯ

**Чорновол Олена Вікторівна. Інтелектуальна система прогнозування енерговиробництва АЕС.** – На правах рукопису.

Магістерська наукова робота на здобуття освітньої кваліфікації «Магістр системного аналізу». – Чорноморський національний університет імені Петра Могили, Миколаїв, 2020.

Дана магістерська наукова робота присвячена дослідженню методів прогнозування енерговиробництва.

Метою є дослідження та реалізація прогнозування енергоефективності на прикладі атомної електроенергії в реальному часі методами машинного навчання.

Об'єктом дослідження є інтелектуальна система прогнозування енерговиробництва АЕС методами машинного навчання.

Предметом дослідження є аналіз вибраних моделей машинного навчання для прогнозування енергії: множинна лінійна регресія (MLR), опорно-верторна регресія (SVR), модель випадкового лісу (Random Forest) та модель екстремального градієнтного бустінгу (XGBoost).

Фахова частина магістерської наукової роботи складається з наступних розділів: атомна енергетика як майбутнє енергетичних ринків; поняття машинного навчання (МН) для регулювання енергії; побудова моделей машинного навчання для прогнозування споживання енергії.

Задачі, які були виконані в процесі роботи:

- аналіз останніх наукових публікацій;
- аналіз предметної області, дослідження стану розвитку атомної енергетики;
- дослідження поняття машинного навчання, моделі та бібліотеки мови програмування Python для вирішення задачі часових рядів;
- сформування та тестування моделей машинного навчання на основі гіперпараметрів;
- порівняльний аналіз результатів роботи моделей;
- розробка інформаційно-аналітичної системи енергоменеджменту.

У спеціальній частині магістерської роботи з «Охорони праці та безпеки в надзвичайних ситуаціях» розглянуто мікрокліматичні умови праці на робочих місцях на предмет виробничого освітлення та дотримання вимог експлуатації ПК в офісі компанії «Template Monster». В результаті розрахунків встановлено, що передбаченої кількості вікон, тобто їх загальної площі вистачає для забезпечення вимог санітарних норм щодо природного освітлення приміщення, для якого проводився аналіз умов праці.

У методичній частині розроблено практичні роботи з курсу багатокритерійного прийняття рішень на теми «Класична лінійна регресія» та «Застосування методу DEMATEL для прийняття рішень на основі обробки експертних суджень і нечіткої логіки в середовищі MATLAB».

Робота складається з 139 сторінок, 48 рисунків, 11 таблиць та 85 посилань на літературні джерела.

**Ключові слова:** машинне навчання, Random Forest, SVR, XGBoost, MLR.

## ABSTRACT

### **Chornovol Olena. Predicting energy efficiency at NPPs using ML methods. –**

On the rights of the manuscript.

Master's scientific work for obtaining an educational qualification "Master of Systems Analysis". – Petro Mohyla Black Sea National University, Mykolaiv, 2020.

The master's research paper is devoted to the study methods of forecasting energy production.

The *purpose* of this paper is to research and implement energy efficiency forecasting using the example of atomic electricity in real time by machine learning methods.

The *object* is an intelligent system for predicting the power generation of nuclear power plants by machine learning methods.

The *subject* of the study is the analysis of selected machine learning models for energy prediction: Multiple Linear Regression (MLR), Supported Vectors Regression (SVR), Random Forest and Extreme Gradient Boosting (XGBoost).

The professional part of master's research paper consists of the following sections: atomic energy as the future of energy markets; the concept of machine learning (ML) for energy regulation; building machine learning models to predict energy consumption.

Tasks that were completed during the work:

- analysis of recent scientific publications;
- domain analysis, research into the state of development of nuclear energy;
- exploring the concept of machine learning, models and libraries of Python programming language to solve the time series problem;
- formation and testing of machine learning models based on hyperparameters;
- comparative analysis of model performance;
- development of an information-analytical system of energy management.

In the special part of the Master's thesis on Occupational Safety and Security in the workplace, microclimatic working conditions for work lighting and compliance with the requirements of PC operation in the Template Monster office are considered. As a result of the calculations it is established that the estimated number of windows, that is, their total area is sufficient to meet the requirements of sanitary standards for the natural lighting of the premises for which the working conditions were analyzed.

In the methodical part practical works on the course of multicriteria decision-making on the topics "Classical linear regression" and "Application of the DEMATEL method for decision-making based on expert judgment processing and fuzzy logic in MATLAB environment" is developed.

The work consists of 139 pages, 48 figures, 11 tables and 85 references to literary sources.

*Keywords: machine learning, Random Forest, SVR, XGBoost, MLR.*