**КЛАСИФІКАЦІЯ МУНІЦИПАЛЬНИХ ОБ'ЄКТІВ ЕНЕРГОСПОЖИВАННЯ
З ВИКОРИСТАННЯМ МЕТОДІВ МАШИННОГО НАВЧАННЯ****А. Л. Перекрест, В. О. Огарь**

Кременчуцький національний університет імені Михайла Остроградського

вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: pkg13@gmail.com, vita.ogar@gmail.com

О. В. Вовна

Донецький національний технічний університет

пл. Шибанкова, 2, м. Покровськ, 85300, Україна. E-mail: Oleksandr.Vovna@donntu.edu.ua

Проведено порівняння методів машинного навчання для класифікації різних муніципальних споживачів енергетичних ресурсів. Методи машинного навчання допомагають в аналізі параметрів систем управління енергоспоживанням та підтримці прийняття ефективних рішень енергоменеджерами. Виконано аналіз особливостей застосування різних методів машинного навчання в задачах аналізу даних: Naive Bayes, CART, Bagged CART, Randoom Forest, Adaboost, eXtreme Gradient Boosting, Support Vector Machine з лінійним ядром та з радіально базисною функцією. Обрано метрики для оцінки якості моделей, що використовувались для класифікації споживачів. Отримано оптимальні значення гіперпараметрів для кожної з моделей. Проведено вибір моделей, що мають найкращі результати відповідно до обраних метрик окремо для кожної групи користувачів: школи, дитячі садки, бібліотеки тощо. Так, для вирішення задачі класифікації даних про теплоспоживання дошкільних навчальних закладів обрано метод eXtreme Gradient Boosting, для класифікації даних про теплоспоживання загальноосвітніх шкіл – bagged CART, для класифікації даних про теплоспоживання закладів позашкільної освіти, бібліотек тощо – Randoom Forest. Отримані результати дозволяють підвищити якість аналізу ефективності використання енергетичних ресурсів для теплозабезпечення будівель різного призначення та можуть впроваджуватись в системи муніципального енергомоніторингу.

Ключові слова: цивільні будівлі, енергоспоживання, аналіз даних, машинне навчання, класифікація.

**КЛАССИФИКАЦИЯ МУНИЦИПАЛЬНЫХ ОБЪЕКТОВ ЭНЕРГОПОТРЕБЛЕНИЯ
С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ****А. Л. Перекрест, В. А. Огарь**

Кременчугский национальный университет имени Михаила Остроградского

ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: pkg13@gmail.com, vita.ogar@gmail.com

А. В. Вовна

Донецкий национальный технический университет

пл. Шибанкова, 2, г. Покровск, 85300, Украина. E-mail: oleksandr.vovna@donntu.edu.ua

Проведено сравнение методов машинного обучения для классификации различных муниципальных потребителей энергетических ресурсов. Методы машинного обучения помогают в анализе параметров систем управления энергопотреблением и поддержке принятия эффективных решений энергоменеджерами. Выполнен анализ особенностей применения различных методов машинного обучения в задачах анализа данных: Naive Bayes, Cart, Bagged Cart, Randoom Forest, Adaboost, eXtreme Gradient Boosting, Support Vector Machine с линейным ядром и с радиально базисной функцией. Выбраны метрики для оценки качества моделей, используемых для классификации потребителей. Получены оптимальные значения гиперпараметров для каждой из моделей. Проведен выбор моделей, имеющих лучшие результаты в соответствии с выбранными метриками отдельно для каждой группы пользователей: школы, детские сады, библиотеки и др. Так, для решения задачи классификации данных о теплоснабжении дошкольных учебных заведений выбран метод eXtreme Gradient Boosting, для классификации данных о теплоснабжении общеобразовательных школ – Bagged Cart, для классификации данных о теплоснабжении учреждений внешкольного образования, библиотек и т.д. – Randoom Forest. Полученные результаты позволяют повысить качество анализа эффективности использования энергетических ресурсов для теплообеспечения зданий различного назначения и могут внедряться в системы муниципального энергомониторинга.

Ключевые слова: гражданские здания, энергопотребление, анализ данных, машинное обучение, классификация.

АКТУАЛЬНІСТЬ РОБОТИ. За сучасних умов постає питання ефективного застосування енергетичних ресурсів, що використовуються в муніципальних будівлях різного призначення для забезпечення необхідних мікрокліматичних умов. На опалення та гаряче водопостачання таких будівель витрачається біля 40 % від первинних енергетичних ресурсів [1]. При цьому, застосування

комп'ютеризованих систем моніторингу та керування енергоспоживанням будівель забезпечує підвищення ефективності використання енергоресурсів та дозволяє накопичувати значний обсяг даних про особливості функціонування інженерних систем та будівель в цілому [2, 3]. Отримані дані разом з інтелектуальними методами їх аналізу [4–6] є основою для визначення

неочевидних шаблонів поведінки об'єктів енергоспоживання, діагностики нештатних ситуацій, прогнозування витрат енергетичних ресурсів тощо.

Метою даної роботи є обґрунтування вибору методів класифікації для інформаційного ранжування муніципальних об'єктів різного призначення на основі даних про їх енергоспоживання.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. Для визначення типових патернів поведінки комунальних закладів в ході дослідження виконано кластеризацію (метод *k-means*) та класифікацію різними методами комунальних споживачів центрального опалення м. Дрогобич. Дані отримані з порталу відкритих даних Дрогобицької міської ради, що був реалізований за підтримки «ТАPAS / Прозорість та підзвітність у державному управлінні та послугах», який фінансується Агенцією США з міжнародного розвитку (USAID) й урядом Великої Британії та здійснюється Фондом Євразія.

Проведений аналіз дозволив представити споживання теплової енергії будинками кількома кластерами. Так, для трьох груп будівель різного призначення отримана наступна кількість кластерів: дошкільні навчальні заклади – 6, загальноосвітні школи – 8 та заклади, що не входять до попередніх груп (заклади позашкільної освіти, бібліотеки тощо) – 6.

У якості методів машинного навчання обрані Naive Bayes, CART, Bagged CART, Random Forest (RF), AdaBoost, XGBoost та Support Vector Machine (SVM) [7–14].

Naive Bayes. Наївний байєсовський класифікатор (*naive Bayes*) – алгоритм, оснований на теоремі Байєса з припущенням про незалежність ознак, тобто алгоритм вважає, що наявність будь-якої ознаки в класі не пов'язана з будь-якою іншою ознакою у класі [7].

CART. У основі алгоритму класифікації та регресії (CART) лежить зменшення міри неоднорідності (задача класифікації) або обраної міри похибки оцінки прогнозування (задача регресії) в кожному з вузлів дерева [7]. Алгоритм бінарного дерева рішень є жадібним – при виборі наступного шляху проходження відбувається орієнтування на рішення поточного етапу.

Bagged CART. Bagging (від *bootstrap aggregation*) – це один з перших і найпростіших видів ансамблів, створений Лео Брейманом в 1994 році [8, 9]. Беггінг заснований на статистичному методі *bootstrap*. Метод бутстрапа полягає в багаторазовому випадковому створенні повторних вибірок з емпіричного розподілу.

Random Forest. RF – алгоритм машинного навчання, що полягає у використанні методу *bagging* на множині дерев рішень з використанням методу випадкових підпросторів та в змозі вирішувати завдання класифікації та регресії [10]. Особливість

побудови дерев у RF – використання максимальної глибини: оскільки розглядається середній результат множини дерев, відсікається необхідність у налаштуванні кожного з дерев (не виконується *padding*). Тренування окремих дерев у RF відбувається незалежно одне від одного, паралельно та на різних підмножинах, що отримані завдяки *bagging* [4].

XGBoost. XGBoost – одна з найефективніших реалізацій градієнтного бустингу, розроблена Тяньцзи Ченом (Tianqi Chen) [11, 12]. Градієнтний бустинг – ансамбль слабких прогнозних моделей, навчання яких проводиться послідовно. На кожній ітерації розраховуються відхилення уже навченого ансамблю, а наступна модель буде передбачувати ці відхилення. Зменшення середнього відхилення моделі (таргету оптимізаційної задачі) проводиться шляхом додавання прогнозу нового дерева до прогнозу уже навченого ансамблю.

SVM. Метод опорних векторів (SVM) – метод аналізу даних, що використовується для задач класифікації та регресії [13, 14].

Метрики для знаходження оптимальних параметрів моделей

Для обґрунтування вибору методів класифікації комунальних об'єктів за встановленими кластерами застосовано певні як цільові значення при визначенні оптимальних параметрів моделей класифікації.

Більшість метрик, таких як швидкість помилок, специфічність, чутливість та точність виводяться з матриці помилок. Крім того, на них ґрунтується кілька вдосконалених метрик, таких як ROC та точність відклику [15]. Розглянемо базовий приклад на основі бінарного класифікатора.

Двійковий класифікатор на виході має два значення класів або міток, наприклад Так/Ні та 1/0, для заданих вхідних даних. Клас, що викликає інтерес з точки зору рішення задачі аналізу даних зазвичай позначають як «позитивний», а інший – як «негативний».

Тестовий набір даних, що використовується для оцінки продуктивності, повинен містити правильні мітки (спостережувані мітки) для всіх екземплярів даних. Ці мітки використовуються для порівняння з прогнозованими мітками для оцінки продуктивності після класифікації.

Передбачувані мітки зазвичай збігаються з частиною спостережуваних міток, адже, зазвичай, створення ідеального класифікатора неможливе.

Матриця помилок формується з чотирьох результатів, отриманих у результаті двійкової класифікації.

Двійковий класифікатор прогнозує всі екземпляри даних тестового набору даних як позитивні чи негативні. Ця класифікація (або прогнозування) дає чотири результати – справжній позитивний, справжній негативний, хибно позитивний і хибно негативний:

– справжній позитивний (TP): правильний

позитивний прогноз;

– хибно позитивний (FP): неправильний позитивний прогноз;

– справжній негативний (TN): правильний негативний прогноз;

– хибно негативний (FN): неправильний негативний прогноз.

Матриця помилок бінарної класифікації – це таблиця два на два, утворена завдяки підрахунку числа чотирьох результатів двійкового класифікатора (табл. 1).

При вирішенні задачі мультикласової класифікації матриця помилок буде містити набір результатів «один проти всіх». Наприклад, у задачі трьох класів чутливість першого класу обчислюється проти всіх зразків у другому та третьому класах тощо.

Таблиця 1 – Матриця помилок бінарної класифікації

Значення	Прогнозовані	
Розмічені	Позитивний	Негативний
Позитивний	TP (A)	FN (C)
Негативний	FP (B)	TN (D)

Коефіцієнт помилок (ERR) обчислюється як кількість усіх неправильних прогнозів, поділених на загальну кількість набору даних. Найкращий показник помилок – 0,0, а найгірший – 1,0.

$$ERR = \frac{B + C}{A + D + C + B}. \quad (1)$$

Точність (Accuracy) обчислюється як кількість усіх правильних прогнозів, поділених на загальну кількість набору даних. Найкраща точність – 1,0, а найгірша – 0,0.

$$Acc = \frac{A + D}{A + D + C + B}. \quad (2)$$

Чутливість (Sensitivity, Recall або True positive rate) обчислюється як кількість правильних позитивних прогнозів, розділених на загальну кількість позитивних. Дану метрику також називають відгуком (recall) або справжньою позитивною пропорцією (TPR). Найкраща чутливість – 1,0, тоді як найгірша – 0,0.

$$Sens = \frac{A}{A + C}. \quad (3)$$

Специфічність (Specificity або True negative rate) обчислюється як кількість правильних негативних прогнозів, поділених на загальну кількість негативних. Її також називають справжньою негативною пропорцією (TNR). Найкраща специфічність – 1,0, тоді як найгірша – 0,0.

$$Spes = \frac{D}{B + D}. \quad (4)$$

Точність (Precision) обчислюється як кількість правильних позитивних прогнозів, поділених на загальну кількість позитивних. Найкраща точність – 1,0, тоді як найгірша – 0,0.

$$Prec = \frac{A}{B + A}. \quad (5)$$

F-оцінка – це гармонійне середнє значення між точністю та відгуком.

$$F1 = \frac{(1 + \beta^2) \cdot Prec \cdot recall}{(\beta^2 \cdot Prec) + recall}. \quad (6)$$

Додаткові показники. Норма відсутності інформації (no-information rate) – припущення, що інформація не виходить за рамки загального розподілу класів, які необхідно передбачити. Це найбільша пропорція спостережуваних класів (у тестовому наборі було більше даних одного класу, ніж іншого). Тест гіпотези також обчислюється для оцінки того, чи є загальна точність більшою, ніж величина найбільшого класу. Також додатково обчислюється поширеність (Prevalence) «позитивної події», рівень виявлення (Detection Rate) (правильний позитивний запис також передбачається як правильний позитивний) та поширеності виявлення (Detection Prevalence) (поширеність передбачуваних позитивних).

Переважає (Prev):

$$Prev = \frac{A + C}{A + B + C + D}. \quad (7)$$

Позитивне прогностичне значення (Positive predictive value):

$$PPV = \frac{Sens \cdot Prev}{((Sens \cdot Prev) + (1 - Spes) \cdot (1 - Prev))}. \quad (8)$$

Негативне прогностичне значення (Negative predictive value):

$$NPV = \frac{Spes \cdot (1 - Prev)}{(((1 - Sens) \cdot Prev) + (Spes \cdot (1 - Prev)))}. \quad (9)$$

Рівень виявлення (Detection Rate):

$$Det Rate = \frac{A}{A + B + C + D}. \quad (10)$$

Поширеність виявлення (Detection Prevalence):

$$Det Prev = \frac{A + B}{A + B + C + D}. \quad (11)$$

Врівноважена точність (Balanced Accuracy):

$$Bal Acc = \frac{(Sens + Spes)}{2}. \quad (12)$$

ROC-крива та AUC. ROC-крива (receiver operating characteristic) – графік, що керує оцінкою якості класифікації, показує співвідношення між Sensitivity і Specificity. Кожна точка кривої ROC виникає із значень в матриці помилок, пов'язаних із застосуванням конкретного відсічення на прогнозах класифікатора.

Кількісна інтерпретація ROC – AUC (area under ROC curve) – площа, обмежена ROC-кривою і віссю Specificity. Вищий показник AUC вказує на якісніший класифікатор, при цьому значення 0,5 демонструє непридатність вибраного методу класифікації (відповідає рандомному класифікатору).

Крива *Precision-Recall* та *prAUC*. Криві точності-відгуку будуються у координатах позитивного прогностичного значення (PPV, у-вісь) та правильної позитивної оцінки (TPR або Recall, вісь x).

Кількісна інтерпретація PRC – prAUC (area under PRC curve) – площа, обмежена PRC-кривою і віссю Recall.

Логістична функція втрат (Logistic Loss). Окремо визначається логістична функція втрат як:

$$\log Loss = -\frac{1}{l} \sum_{i=1}^l (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)). \quad (13)$$

де $\hat{y}$ – це відповідь алгоритму на i -му об'єкті, y – справжня мітка класу на i -му об'єкті, а l – розмір вибірки.

Інтуїтивно можна уявити мінімізацію logloss як завдання максимізації асигуру шляхом штрафу за неправильні прогнози. Однак необхідно відзначити, що logloss вкрай сильно штрафує за впевненість класифікатора в неправильній відповіді.

Карра коефіцієнт Коена. Карра коефіцієнт Коена схожий на точність класифікації, за винятком того, що нормалізується за базовою лінією випадкових шансів у наборі даних. Іншими словами, це статистика, яка використовується для вимірювання надійності між рейтингами для якісних позицій. Дану метрику найдоцільніше використовувати для даних, які мають дисбаланс у класах (наприклад, 70–30 розбиття для класів 0 і 1, при якому можна досягти 70 % точності, передбачивши, що всі екземпляри відносяться до класу 0).

Карра коефіцієнт Коена характеризує зв'язок двох рейтингів, що класифікують N предметів у категорії C та є несумісними. Визначається за наступною формулою:

$$Kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}, \quad (14)$$

де p_0 – відносна дотримана згода між «переможцями» (ідентична точності), а p_e – гіпотетична ймовірність випадкової «перемоги», при використанні спостережуваних даних для обчислення ймовірностей кожного спостерігача, який рандомно бачить кожну категорію.

Результати класифікації на прикладі даних про теплоспоживання дошкільних навчальних закладів

Тренування моделей для класифікації даних дошкільних навчальних закладів проводилося на навчальній вибірці розміром у 1179 записів.

Наївний Байєс (naive Bayes)

Результати ресемплінга за параметрами налаштування моделі наведені в табл. 2. Матриця помилок класифікатора на тестувальній вибірці наведена в табл. 3.

Таблиця 2 – Результати ресемплінга за параметрами налаштування моделі

Параметри та метрики	Значення	
usekernel	FALSE	TRUE
logLoss	0,3019	0,1791
AUC	0,9941	0,997
prAUC	0,8928	0,9149
Accuracy	0,9186	0,9381
Kappa	0,8973	0,9219
Mean_F1	0,9072	0,9269
Mean Sensitivity	0,9260	0,9449
Mean Specificity	0,9839	0,988
Mean Pos. Pred. Value	0,8989	0,9173
Mean Neg. Pred. Value	0,9828	0,9869
Mean Precision	0,8989	0,9173
Mean Recall	0,9260	0,9449
Mean Detection Rate	0,1531	0,1564

Таблиця 3 – Матриця помилок (Confusion Matrix) для класифікатора naive Bayes

Відповідь алгоритму / Реальна мітка класу	C1	C2	C3	C4	C5	C6
C1	51	0	0	2	0	4
C2	0	129	0	0	2	2
C3	0	0	30	0	0	5
C4	0	0	0	131	0	0
C5	0	0	0	12	38	1
C6	1	2	2	0	0	91

Статистика по окремим класам для класифікатора на тестувальній вибірці наведена в табл. 4.

Таблиця 4 – Статистика по окремим класам для класифікатора naive Bayes

Метрика / Мітка класу	C1	C2	C3	C4	C5	C6
Sensitivity	0,9808	0,9847	0,9375	0,9034	0,9500	0,8835
Specificity	0,9867	0,9892	0,9894	1,0000	0,9719	0,9875
Pos. Pred. Value	0,8947	0,9699	0,8571	1,0000	0,7451	0,9479
Neg. Pred. Value	0,9978	0,9946	0,9957	0,9624	0,9956	0,9705
Prevalence	0,1034	0,2604	0,0636	0,2883	0,0795	0,2048
Detection Rate	0,1014	0,2565	0,0596	0,2604	0,0756	0,1809
Detection Prevalence	0,1133	0,2644	0,0696	0,2604	0,1014	0,1909
Balanced Accuracy	0,9837	0,9870	0,9634	0,9517	0,9610	0,9355

На рисунках 1–3 наведені графіки густини ймовірностей деяких змінних для кожного кластеру.

Рисунок 1 – Графік густини імовірності площі для кожного кластеру

Рисунок 2 – Графік густини імовірності зовнішньої температури для кожного кластеру

Рисунок 3 – Графік густини імовірності величини теплоспоживання для кожного кластеру

CART

Результати ресемплінга за параметрами налаштування моделі наведені в табл. 5. Параметр моделі, що налаштовується – параметр складності (complexity parameter, cp), що використовується для контролю розміру дерева та вибору його оптимального розміру. Це мінімальне вдосконалення моделі, що є необхідним для кожного вузла (сума помилок класифікації у кожному термінальному вузлі та добутку штрафу на кількість розщеплень). Оптимальне значення $cp = 0,05210284$.

На рисунку 4 наведено вигляд дерева класифікації.

Рисунок 4 – Дерево класифікації даних про теплоспоживання дошкільних навчальних закладів

Bagged CART

Результати моделі на тестовій вибірці наведені в табл. 5.

Таблиця 5 – Результати моделі на тестовій вибірці

Параметри та метрики	Значення
logLoss	0,08095029
AUC	0,9926007
prAUC	0,09251797
Accuracy	0,9884836
Kappa	0,9863085
Mean F1	0,9837099
Mean Sensitivity	0,9642871
Mean Specificity	0,9983627
Mean Pos. Pred. Value	0,965601
Mean Neg. Pred. Value	0,9978816
Mean Precision	0,965601
Mean Recall	0,9642871
Mean Detection Rate	0,1235605
Mean Balanced Accuracy	0,9810633

RF

Результати ресемплінга за параметрами налаштування моделі наведені в табл. 6. Параметром, що потребує налаштування, є кількість змінних, що випробовуються при кожному розділенні ($mtry$). Оптимальне значення – 4.

Помилка ООВ дорівнює 1,53 % на навчальній вибірці. На рис. 5 наведено графік значень помилок у залежності від кількості дерев у лісі. Оптимальна кількість дерев ($ntrees$) – 500.

Таблиця 6 – Результати ресемплінга за параметрами налаштування моделі

Параметри та метрики	Значення		
mtry	2	4	6
logLoss	0,0599448	0,039353	0,0637073
AUC	0,9997642	0,999764	0,9990008
prAUC	0,8141958	0,434700	0,1273856
Accuracy	0,9838972	0,986454	0,9847519
Kappa	0,9795085	0,982763	0,9806022
Mean F1	0,9812669	0,983913	0,9823203
Mean Sensitivity	0,9790008	0,982260	0,9808717
Mean Specificity	0,9965941	0,997148	0,9967998
Mean Pos. Pred. Value	0,9851747	0,986966	0,9852258
Mean Neg. Pred. Value	0,9968463	0,997333	0,9969769
Mean Precision	0,9851747	0,986966	0,9852258
Mean Recall	0,9790008	0,982260	0,9808717
Mean Detection Rate	0,1639829	0,164409	0,1641253
Mean Balanced Accuracy	0,9877974	0,989704	0,9888358

AdaBoosting

Результати ресемплінга за параметрами налаштування моделі наведені в табл. 7. Дана модель показує найкращі результати при таких значеннях параметрів, що налаштовуються: кількість ітерацій (mfinal) – 100, максимальна глибина (maxdepth) – 3.

Рисунок 5 – Графік значень помилок у залежності від кількості дерев у лісі

eXtreme Gradient Boosting

Результати ресемплінга за гіперпараметрами, що потребують налаштування та їх оптимальне значення:

- кількість ітерацій бустингу (nrounds) – 100;
- максимальна глибина дерева (max_depth) – 3;
- регуляризаційний параметр (η), який

використовується для запобігання перенаванчню: після кожного кроку існує можливість безпосередньо отримувати ваги нових функцій, і η зменшує ваги функцій, щоб зробити процес бустингу більш стабільним (eta) – 0,4;

– мінімальне зменшення втрат, необхідне для подальшого розділення на листковому вузлі дерева (gamma) – 0;

– співрозмірне співвідношення стовпців при побудові кожного дерева, виділення підвибірок відбувається один раз для кожного побудованого дерева (colsample_bytree) – 0,6;

– мінімальна сума ваги екземпляра (гессіна), необхідна дочірньому вузлу: якщо на етапі розділення дерева з'явиться листковий вузол з сумою ваги екземпляра менше гессіна, тоді процес побудови дерева зупиниться (min_child_weight) – 1;

Таблиця 7 – Результати ресемплінга за параметрами налаштування

Параметри та метрики	Значення				
maxdepth	1	2	2	3	3
mfinal	150	100	150	100	150
logLoss	13,32959	5,276322	5,27702	0,230228	0,2297778
AUC	0,634133	0,811754	0,811706	0,994639	0,9946325
prAUC	0,002350	0,060366	0,062627	0,119041	0,1220996
Accuracy	0,546237	0,832015	0,829473	0,969446	0,9685844
Kappa	0,377472	0,779841	0,776399	0,961029	0,9599327
Mean F1	–	–	–	0,961959	0,9609934
Mean Sensitivity	0,332353	0,640011	0,635845	0,955914	0,9552197
Mean Specificity	0,896823	0,963712	0,963148	0,993471	0,9933109
Mean Pos. Pred. Value	–	–	–	0,973750	0,9725753
Mean Neg. Pred. Value	0,924197	0,970909	0,970432	0,994042	0,9938611
Mean Precision	–	–	–	0,973750	0,9725753
Mean Recall	0,332353	0,640011	0,635845	0,955914	0,9552197
Mean Detection Rate	0,091040	0,138669	0,138245	0,161574	0,1614307
Mean Balanced Accuracy	0,614588	0,801862	0,799497	0,974693	0,9742653

– додаткове співвідношення навчальних примірників – якщо його значення 0,5 – XGBoost випадково відбирає половину навчальних даних перед побудовою дерев, і це дозволить запобігти перенаванчання, виділення підвбірок відбувається на кожній ітерації бустингу (subsample) – 0,75.

SVM

SVM (Linear kernel)

Результати ресемплінга за параметрами налаштування моделі наведені в табл. 8. Дана модель показує найкращі результати при такому значенні гіперпараметру $C = 1$. Кількість опорних векторів – 219.

Таблиця 8 – Результати ресемплінга за параметрами налаштування моделі

Параметри та метрики	Значення
logLoss	1,0549
AUC	0,9365
prAUC	0,7214
Accuracy	0,5987
Kappa	0,5147
Mean F1	–
Mean Sensitivity	0,6439
Mean Specificity	0,9214
Mean Pos. Pred. Value	–
Mean Neg. Pred. Value	0,9293
Mean Precision	–
Mean Recall	0,6439
Mean Detection Rate	0,0998
Mean Balanced Accuracy	0,7827

SVM (Radial Basis Function Kernel)

Результати ресемплінга за параметрами налаштування моделі наведені в табл. 9. Дана модель показує найкращі результати при таких значеннях гіперпараметрів: $\xi = 0,2828508$ та $C = 1$. Кількість опорних векторів – 393.

Таблиця 9 – Результати ресемплінга за параметрами налаштування моделі

Параметри та метрики	Значення		
C	0,25	0,5	1
logLoss	1,1135	1,0837	1,0741
AUC	0,9076	0,9139	0,9137
prAUC	0,6948	0,7048	0,7049
Accuracy	0,5487	0,5581	0,5589
Kappa	0,4562	0,4667	0,4674
Mean F1	0,5330	0,5435	0,5467
Mean Sensitivity	0,6367	0,6407	0,6370
Mean Specificity	0,9110	0,9129	0,9131
Mean Pos. Pred. Value	0,6672	0,6694	0,6680
Mean Neg. Pred. Value	0,9190	0,9205	0,9206
Mean Precision	0,6672	0,6694	0,6680
Mean Recall	0,6367	0,6407	0,6370
Mean Detection Rate	0,0915	0,0930	0,0932
Mean Balanced Accuracy	0,7738	0,7768	0,7750

Вибір найкращої моделі для класифікації

Узагальнені результати по метрикам, що не були наведені при оцінці якостей моделей на тестувальній вибірці для даних про теплоспоживання дошкільних навчальних закладів, наведені в табл. 10.

Таблиця 10 – Результати accuracy, no information rate, p-value, kappa

Модель / метрика	Accuracy	No Information Rate	P-Value [Acc > NIR]	Kappa
Naive Bayes	0,9344	0,2883	< 2,2e-16	0,9172
CART	0,7376	0,2883	< 2,2e-16	0,6462
Bagged CART	0,992	0,2883	< 2,2e-16	0,9899
RF	0,9881	0,2883	< 2,2e-16	0,9848
Adaboost	0,9801	0,2883	< 2,2e-16	0,9747
XGBoosting	0,995	0,2883	< 2,2e-16	0,9899
Svm Linear	0,5845	0,2883	< 2,2e-16	0,4980
Svm RBF	0,5726	0,2883	< 2,2e-16	0,4874

Аналіз комплексу вибраних метрик дозволяє зробити висновок, що для вирішення задачі класифікації даних про теплоспоживання дошкільних навчальних закладів необхідно у якості моделі обрати eXtreme Gradient Boosting.

Для даних про теплоспоживання загальноосвітніх шкіл м. Дрогобич обрано наступні гіперпараметри для моделей:

- CART: $cp = 0,1141975$;
- RF: $ntrees = 500$, $mtry = 4$;
- Adaboost: $mfinal = 150$, $maxdepth = 3$;
- XGBoosting: $nrounds = 150$, $max_depth = 1$, $eta = 0,3$, $gamma = 0$, $colsample_bytree = 0,8$, $min_child_weight = 1$, $subsample = 0,5$;
- Svm Linear: $C = 1$, кількість опорних векторів – 361;
- Svm RBF $\xi = 0,49588$, $C = 1$, кількість опорних векторів – 502.

Для даних про теплоспоживання закладів позашкільної освіти, бібліотек тощо м. Дрогобич вибрано наступні гіперпараметри для моделей:

- CART: $cp = 0,1360691$;
- RF: $ntrees = 50$, $mtry = 2$;
- Adaboost: $mfinal = 50$, $maxdepth = 3$;
- XGBoosting: $nrounds = 100$, $max_depth = 2$, $eta = 0,3$, $gamma = 0$, $colsample_bytree = 0,8$, $min_child_weight = 1$, $subsample = 0,5$;
- Svm Linear: $C = 1$, кількість опорних векторів – 245;
- Svm RBF $\xi = 0,98205$, $C = 1$. Кількість опорних векторів – 339.

На підставі результатів моделей для класифікації даних про теплоспоживання загальноосвітніх шкіл можемо стверджувати, що найкращим методом, з урахуванням комплексу вибраних метрик для оцінки, є bagged CART.

Результати моделей для класифікації даних про теплоспоживання закладів позашкільної освіти,

бібліотек тощо дозволяють стверджувати, що найдоцільнішим методом для класифікації інформації про дану групу є RF.

ВИСНОВКИ. Обрано метрики для оцінки якості моделей, що використовуватимуться для класифікації споживачів: logLoss, AUC, prAUC, Accuracy, Kappa, Mean F1, Sensitivity, Specificity, Pos. Pred. Value, Neg. Pred. Value, Precision, Recall, Detection Rate, Balanced Accuracy. Здійснено аналіз математичних апаратів та порівняння методів машинного навчання, що дозволяють на основі даних про теплоспоживання цивільних будівель класифікувати споживачів за класами, які були виділені на попередньому етапі дослідження. Методи, що використано у дослідженні: Naive Bayes, CART, Bagged CART, RF, Adaboost, eXtreme Gradient Boosting, SVM з лінійним ядром та з радіально базисною функцією. Знайдено оптимальні значення гіперпараметрів для кожної з моделей. Проведено вибір моделей, що мають найкращі результати відповідно до вибраних метрик окремо для кожної групи користувачів. Так, для вирішення задачі класифікації даних про теплоспоживання дошкільних навчальних закладів – eXtreme Gradient Boosting, для класифікації даних про теплоспоживання загальноосвітніх – bagged CART, для класифікації даних про теплоспоживання закладів позашкільної освіти, бібліотек тощо – RF.

ЛІТЕРАТУРА

1. Руководство по энергетической статистике. Международное энергетическое агентство. IEA Publications: 9, rue de la Federation. France by Stedi. 2007, 192 с.
2. Bishop C. M. Pattern Recognition and Machine Learning. New York : Springer, 2006. 738 с.
3. Грас Дж. Data Science: Наука о данных с нуля: пер. с англ. Санкт-Петербург : БХВ-Петербург, 2017. 336 с.
4. Ситник В. Ф., Краснюк М. Т. Интеллектуальный анализ данных. Київ : КНЕУ, 2007. 376 с.
5. Perekrest A., Chorny O., Mur O., Nikolenko A., Kuznetsov V., Kuznetsova Y. Preparation and preliminary analysis of data on energy consumption by municipal buildings. *Eastern-European Journal of Enterprise Technologies*. Vol. 6. № 8–96. 2018. P. 32–42.
6. Vovna O., Laktionov I., Sukach S., Kabanets M., Cherevko E. Method of adaptive control of effective energy lighting of greenhouses in the visible optical range. *Bulgarian Journal of Agricultural Science*. 2018. Vol. 24 (2). P. 335–340.
7. Domingos P. M., Pazzani M. J. Beyond independence: conditions for the optimality of the simple Bayesian classifier. *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning (ICML '96)*, Lorenza Saitta (Ed.). San Francisco, CA : Morgan Kaufmann Publishers Inc., 1996. P. 105–112.
8. Loh Wei-Yin. Classification and Regression Trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2011. Vol. 1. P. 14–23. DOI: 10.1002/widm.8.
9. Leo Breiman. Bagging predictors. *Machine Learning*. 1996. Vol. 24. № 2. P. 123–140. DOI: 10.1007/BF00058655.
10. Leo Breiman. Random Forests. *Machine Learning*. 2001. Vol. 45. № 1. P. 5–32. DOI: 10.1023/A:1010933404324.
11. Freund Y., Schapire R. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*. 1999. Vol. 14. P. 771–780.
12. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 2016. New York : ACM, 2016. P. 785–794. DOI: 10.1145/2939672.2939785.
13. Bartlett P., Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers. *Advances in Kernel Methods*. 1998. Cambridge, USA: MIT Press, 1999. P. 43–54. URL: <https://pdfs.semanticscholar.org/1f22/85a73e640ac6ce1f713b61bed838.pdf>. (дата звернення: 13.05.2020).
14. Vapnik V., Chapelle O. Bounds on error expectation for support vector machines. *Neural Computation*. 2000. Vol. 12 (9). P. 2013–2036. URL: <https://pdfs.semanticscholar.org/ca38/dc7c67ead75dac4b1a2c766cdf0c.pdf>. (дата звернення: 13.05.2020).
15. Hossin M., Sulaiman M. Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining and Knowledge Management Process*. 2015. Vol. 5 (2). P. 1–11. DOI: 10.5121/ijdkp.2015.5201.

CLASSIFICATION OF MUNICIPAL ENERGY CONSUMPTION FACILITIES WITH THE USE OF MACHINE LEARNING METHODS

A. Perekrest, V. Ogar

Kremenchuk Mykhailo Ostrohradskyi National University

vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E-mail: pksk13@gmail.com, vita.ogar@gmail.com

O. Vovna

Donetsk National Technical University

pl. Shibankova, 2, Pokrovsk, 85300, Ukraine. E-mail: Oleksandr.Vovna@donntu.edu.ua

Purpose. A comparison of machine learning methods for the classification of different municipal consumers of energy resources. Machine learning methods help in analyzing the parameters of the energy management system and support effective decision making by energy managers. **Methodology.** The analysis of different methods application

peculiarities of machine learning in data analysis problems was performed: Naive Bayes, CART, Bagged CART, Random Forest, Adaboost, eXtreme Gradient Boosting, Support Vector Machine with linear core and with radial basis function. **Results.** Metrics were selected to assess the quality of the models used to classify consumers. The optimal values of hyperparameters for each of the models were obtained. The models with the best results were selected according to the selected metrics separately for each user group: schools, kindergartens, libraries, etc. **Originality.** Thus, the eXtreme Gradient Boosting method was chosen to solve the problem of heat consumption data of preschool educational institutions, bagged CART was used to classify data on heat consumption of secondary schools, Random Forest was used to classify data on heat consumption of preschool institutions, libraries, etc. **Practical value.** The obtained results allow to improve the efficiency analysis quality of energy resources use for heat supply of buildings of different purposes and can be implemented in the municipal energy monitoring systems. Figures 5, tables 10, references 15.

Key words: civil buildings, energy consumption; data analysis; machine learning, classification.

REFERENCE

1. Guide to Energy Statistics. (2007). International Energy Agency. IEA Publications: 9, rue de la Federation. France by Stedi, 192. [in France].
2. Vovna, O., Laktionov, I., Sukach, S., Kabanets, M., & Cherevko, E. (2018). Method of adaptive control of effective energy lighting of greenhouses in the visible optical range. *Bulgarian Journal of Agricultural Science*, 24 (2), 335–340.
3. Perekrest, A., Chorny, O., Mur, O., Nikolenko, A., Kuznetsov, V., & Kuznetsova, Y. (2018). Preparation and preliminary analysis of data on energy consumption by municipal buildings. *Eastern-European Journal of Enterprise Technologies*, 6, 32–42.
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. NY: Springer.
5. Grus, Joel. *Data Science from Scratch* (2017). SPB.: BHV-Peterburg. [in Russian]
6. Sutnyk, V. F., & Krasnuk, M. T. (2007). *Intelektualnyi analiz danih [Data mining]*. Kyiv: KNEY. [in Ukrainian]
7. Domingos, P. M., & Pazzani, M. J. (1996). Beyond independence: conditions for the optimality of the simple Bayesian classifier. *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning (ICML '96)*, San Francisco, CA: Morgan Kaufmann Publishers Inc., 105–112.
8. Loh, Wei-Yin. (2011). Classification and Regression Trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 14–23. DOI: 10.1002/widm.8.
9. Leo, Breiman (1996). Bagging predictors. *Machine Learning*, 24, 123–140. DOI: 10.1007/BF00058655.
10. Leo, Breiman (2001). Random Forests. *Machine Learning*, 45, 5–32. DOI: 10.1023/A:1010933404324.
11. Freund, Y., & Schapire, R. A. (1999). Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14, 771–780.
12. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, NY: ACM, 785–794. DOI: 10.1145/2939672.2939785.
13. Bartlett, P., & Shawe-Taylor, J. (1998). *Generalization performance of support vector machines and other pattern classifiers. Advances in Kernel Methods*. Cambridge, USA: MIT Press, 43–54. URL: <https://pdfs.semanticscholar.org/1f22/85a73e640ac6ce1f713b61bed838.pdf>.
14. Vapnik, V., & Chappelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation*, 12 (9), 2013–2036. URL: <https://pdfs.semanticscholar.org/ca38/dc7c67ead75dac4b1a2c766cdf0c.pdf>.
15. Hossin, M., & Sulaiman, M. (2015). Review Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining and Knowledge Management Process*, 5 (2), 1–11. DOI: 10.5121/ijdkp.2015.5201.

Стаття надійшла 26.05.2020.