ЧЕРНІВЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ЮРІЯ ФЕДЬКОВИЧА

Навчально-науковий інститут фізико-технічних та комп'ютерних наук
кафедра комп'ютерних наук

Аналіз та прогноз донорства крові

Курсова робота

Рівень вищої освіти – перший (бакалаврський)

Виконала:
студентка 4 курсу, 444А групи
Мельничук Анна Геннадіївна
Керівник:
Кандидат фізико-математичних наук,
доцент Ковальчук М.Л.

Чернівці – 2024

Чернівецький національний університет імені Юрія Федьковича

Навчально-науковий інститут фізико-технічних та комп'ютерних наук

Кафедра: комп'ютерних наук

Спеціальність: 122 Комп'ютерні науки

Рівень вищої освіти: перший (бакалаврський)

Форма навчання денна, курс 4, група 444А

ЗАВДАННЯ

НА КУРСОВУ РОБОТУ СТУДЕНТА

Мельничук Анни Геннадіївни

(прізвище , ім'я, по батькові)

1. Тема роботи

Аналіз та прогноз донорства крові

затверджена протоколом засідання кафедри від «30» серпня 2024 року № 2

2. Термін подання студентом закінченої роботи 4. 11. 2024р.

3. Вхідні дані до роботи

Середовище Jupyter Notebook, Kaggle, GitHub

4. Зміст розрахунково-пояснювальної записки (перелік питань, які треба розробити)

Вступ

Розділ 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

Розділ 2 ПРОЄКТУВАННЯ ТА РЕАЛІЗАЦІЯ МОДЕЛІ

Розділ 3 ВІЗУАЛІЗАЦІЯ ТА РЕЗУЛЬТАТИ

Розділ 4 ПЕРСПЕКТИВИ ВДОСКОНАЛЕННЯ МОДЕЛІ

Висновок

Список використаних джерел

Додатки

5. Перелік графічного, наочного матеріалу

Скріншоти графіків, коду, таблиць

6. Календарний план підготовки курсової роботи

№ з/п	*Етапи роботи*	*Термін виконання*
1	Вибір, погодження й затвердження теми, призначення наукового керівника	1.09.2024
2	Складання календарного плану й розширеного плану-конспекту роботи. Опрацювання джерел.	11.09.2024
3	Організація і проведення теоретичного, емпіричного (експериментального) дослідження	29.09.2024
4	Підготовка складових частин (розділів) роботи	2.10.2024
4.1	Вступ	2.10.2024
4.2	Розділ 1 Аналіз предметної області	5.10.2024
4.3	Розділ 2 Проектування та реалізація моделі	18.10.2024
4.4	Розділ 3 Візуалізація та результати	18.10.2024
4.5	Розділ 4 Перспективи вдосконалення моделі	21.10.2004
4.6	Висновки	22.10.2024
4.7	Список використаних джерел	22.10.2024
4.8	Додатки	23.20.2024
5	Усунення зауважень, урахування рекомендацій наукового керівника, доповнення або скорочення обсягу роботи.	1.11.2024
6	Оформлення тексту роботи, подання роботи науковому керівникові.	4.11.2024
7	Оформлення презентації. Підготовка доповіді на захист.	8.11.2024
8	Захист курсової роботи.	11.11.2024

Студент _____ А.Г.Мельничук

(підпис) *(ініціали, прізвище)*

Науковий керівник _____ М.Л.Ковальчук

(підпис) *(ініціали, прізвище)*

«___» _____ 20__ р.

АНОТАЦІЯ

Стабільне забезпечення медичних установ донорською кров'ю є критично важливим для надання невідкладної медичної допомоги. Оскільки система донорства здебільшого базується на добровільних внесках, забезпечення стабільного поповнення запасів крові залишається непередбачуваним, що часто призводить до їх нестачі.

Використання технологій машинного навчання, може сприяти підвищенню точності прогнозування донорських потоків, оптимізуючи управління запасами крові в медичних установах.

У роботі розроблено та реалізовано модель, призначену для прогнозування ймовірності донорства крові на основі історичних даних про донорів. Процес дослідження включав підготовку набору даних, розробку архітектури моделі, її навчання і перевірку за метриками точності. Отримані результати показують, що модель ефективно прогнозує ймовірність повторних донацій, що може допомогти у зменшенні дефіциту крові та покращенні управління її запасами.

Робота може бути корисною для впровадження в медичних установах як програмний інструмент прогнозування, який допоможе оперативно реагувати на потребу в донорах та знижувати витрати на зберігання.

Ключові слова: ПРОГНОЗУВАННЯ ДОНОРСТВА, МАШИННЕ НАВЧАННЯ, УПРАВЛІННЯ ЗАПАСАМИ КРОВІ, ПОВТОРНІ ДОНОРАЦІЇ.

Курсова робота містить результати власних досліджень. Використання ідей, результатів і текстів наукових досліджень інших авторів мають посилання на відповідне джерело.

_____ А.Г.Мельничук

(підпис) (ініціали та прізвище студента)

ANOTATION

A stable supply of donated blood to medical institutions is critically important for the provision of emergency medical care. Since the donation system is largely based on voluntary contributions, ensuring a stable supply of blood remains unpredictable, often leading to shortages.

The use of machine learning technologies can contribute to increasing the accuracy of predicting donor flows, optimizing the management of blood stocks in medical institutions.

The work developed and implemented model designed to predict the probability of blood donation based on historical data about donors. The research process included the preparation of the data set, the development of the neural network architecture, its training and verification according to various accuracy metrics. The obtained results show that the model effectively predicts the probability of repeated donations, which can help reduce blood shortages and improve blood supply management.

The work may be useful for implementation in medical institutions as a software forecasting tool that will help respond quickly to the need for donors and reduce storage costs.

Keywords: DONATION PREDICTION, MACHINE LEARNING, BLOOD STOCK MANAGEMENT, REPEATED DONATIONS.

The course work contains the results of own research. The use of ideas, results and texts of scientific research of other authors have a link to the appropriate source.

_____ A.G.Melnychuk

(signature) (student's initials and surname)

ЗМІСТ

ВСТУП

Прогнозування донорства крові є важливою та актуальною проблемою у сфері охорони здоров’я, адже стабільність запасів крові критично впливає на своєчасне надання медичної допомоги. Зважаючи на те, що донорство базується на добровільних засадах, забезпечення регулярних поставок крові стає непередбачуваним, що призводить до ризику дефіциту в періоди підвищеного попиту.

Крім цього, неефективне управління запасами може спричиняти надлишкові витрати на зберігання та швидке зниження якості крові, тоді як її нестача може негативно впливати на своєчасність надання медичних послуг. Технології машинного навчання, зокрема нейронні мережі, дозволяють підвищити точність прогнозів завдяки аналізу історичних даних про донорів, допомагаючи медичним закладам ефективніше управляти своїми ресурсами. Отже, дослідження прогнозування ймовірності донації крові з використанням нейронних мереж є актуальним для розвитку галузі "Інформаційні технології" та має практичне значення для оптимізації роботи медичних закладів.

Метою роботи є розробка моделі для прогнозування ймовірності донації крові з огляду на історичні дані, що допоможе підвищити ефективність управління запасами крові у медичних установах. Завдання дослідження включають:

- Аналіз проблеми прогнозування донорства крові та наявних методів у цій сфері.
- Підготовка й обробка даних, що містять інформацію про попередні донації.
- Розробка та тренування моделі для прогнозування ймовірності майбутньої донації.
- Оцінка ефективності моделі за допомогою обраних метрик точності.

– Формування рекомендацій щодо використання моделі у медичних установах для забезпечення оптимального управління запасами.

Об’єктом дослідження є процес донорства крові, особливо щодо питань його регулярності.

Предметом дослідження є модель, розроблена для прогнозування ймовірності донорства на основі історичних даних, що містять інформацію про минулі донації крові.

У роботі застосовані методи аналізу та обробки даних, машинного навчанн а також статистичні методи для оцінки результатів і їх візуалізації.

Запропонована модель може використовуватися у медичних установах для більш точного планування запасів крові, допомагаючи уникати ситуацій дефіциту та знижувати витрати на зберігання невикористаної крові. Вона може бути інтегрована в інформаційні системи медичних установ як програмний інструмент для автоматизованого прогнозування.

Сторінок – 28, рисунків – 22, джерел літератури – 10, додатків – 2.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Аналіз предметної області

Проблема управління запасами крові є надзвичайно важливою для сучасної системи охорони здоров’я. Донорська кров — це цінний, але обмежений ресурс, який потребує ретельного контролю, планування і прогнозування. Під час надзвичайних ситуацій, таких як масштабні аварії чи стихійні лиха, потреба в донорській крові різко зростає, і медичні установи часто стикаються з труднощами у забезпеченні необхідних запасів.

Були зафіксовані випадки, коли внаслідок нестачі донорів пацієнти не могли отримати своєчасну допомогу, що іноді призводило до критичних ситуацій, а інколи — і до летальних наслідків. Наприклад, у деяких країнах світу недолік крові в медичних закладах може виникати навіть через непередбачувані коливання кількості донорів упродовж року. Це ще раз підкреслює необхідність використання автоматизованих систем прогнозування, що дозволяють більш ефективно організовувати роботу банків крові, забезпечуючи прогноз попиту та оптимізуючи роботу з донорами.

Водночас, точність прогнозів ускладнюється через різноманітність факторів, що впливають на донорську активність. Серед них можна відзначити соціальні, економічні, культурні та індивідуальні фактори, кожен з яких може по-своєму впливати на рішення людей про донорство крові. До того ж, термін придатності донорської крові є обмеженим, що створює додаткові труднощі в її управлінні.

1.2 Існуючі рішення та технології

Існуючі підходи до прогнозування донорства крові охоплюють різні методи, зокрема класичні статистичні моделі та сучасні алгоритми машинного навчання. На українському ринку є приклади таких сервісів, як Donorium і DonorUA, які надають деякі аналітичні можливості для прогнозування активності донорів. Однак ці платформи мають певні

обмеження: більшість із них орієнтуються лише на загальні оцінки без врахування індивідуальних і поведінкових особливостей кожного донора, що може знизити точність прогнозів у специфічних ситуаціях.

У загальному контексті сучасних технологій для прогнозування донорства активно використовуються методи регресії. Зокрема, регресійні моделі аналізують залежність між певними характеристиками донорів, як-от вік, частота попередніх здач крові, і ймовірністю повторного донорства. Однак обмеженням таких моделей є складність обробки складних нелінійних залежностей.

На рисунку 1.1 можна побачити з якою схемою працює донорство крові.

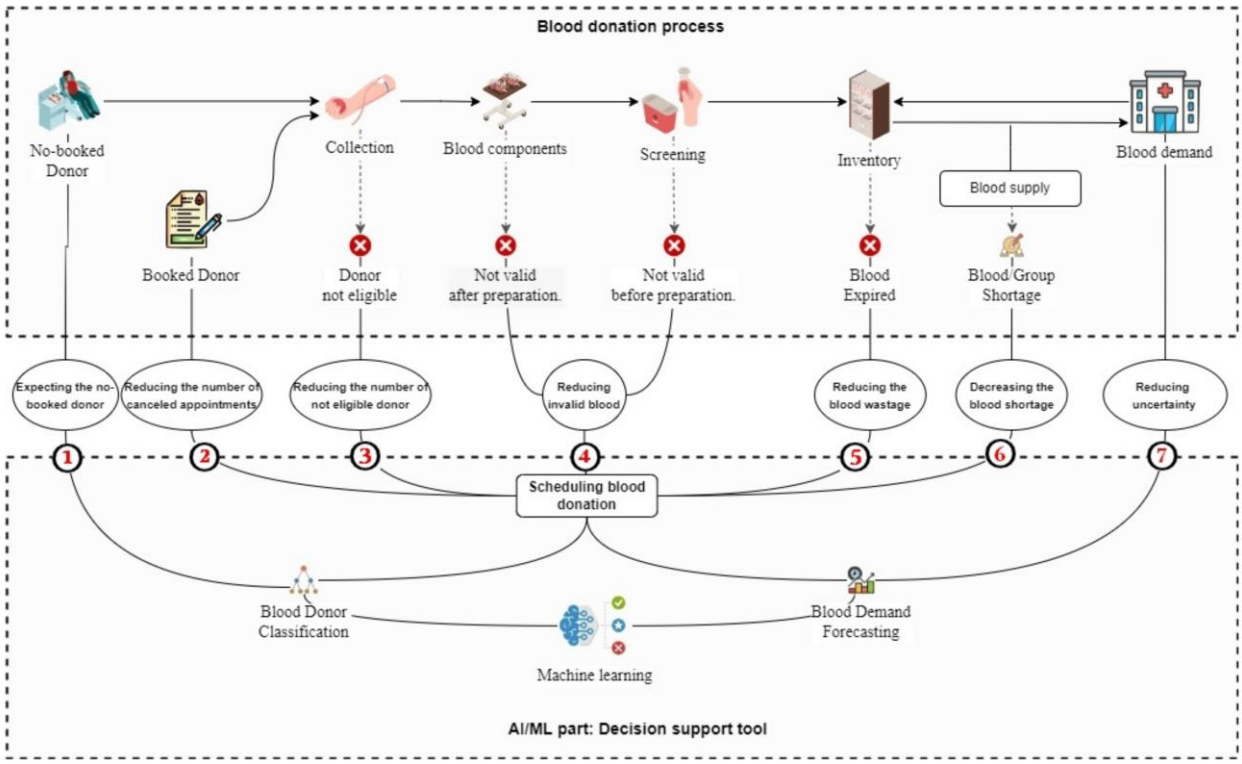


Рисунок 1.2 – Процес донації крові

Іншим напрямом є застосування методів класифікації, таких як дерева рішень, методи найближчих сусідів та вектори підтримки. Ці алгоритми дозволяють аналізувати і класифікувати поведінкові та демографічні характеристики донорів, що може допомогти в прогнозуванні потенційної активності в різні періоди. Проте, і тут існує певний недолік — такі методи можуть бути менш гнучкими при адаптації до нових даних.

Найбільш перспективними для задач прогнозування донорства є нейронні мережі. Ці моделі здатні обробляти складні та часто нелінійні взаємозв’язки між численними факторами, що впливають на донорську активність. Глибокі нейронні мережі, завдяки своїй архітектурі, можуть обробляти великі обсяги даних, враховуючи численні приховані зв’язки між характеристиками донорів. Це робить їх більш надійними у створенні ймовірнісних прогнозів, особливо в задачах з високою динамікою, як у випадку донорства крові.

1.3 Вибір підходу та напрямки подальших досліджень

На основі огляду існуючих методів можна зробити висновок, що для досягнення високої точності у прогнозуванні донорства крові найкраще підходять методи, засновані на нейронних мережах. Їх здатність адаптуватися до змін у структурі даних, виявляти приховані залежності та створювати надійні прогнози дозволяє досягти більш точних результатів порівняно з традиційними підходами.

У цьому дослідженні пропонується модель яка зможе прогнозувати активність донорів на основі аналізу їхніх демографічних та поведінкових характеристик. Основними напрямками подальших досліджень є:

- покращення налаштувань гіперпараметрів моделі для підвищення точності прогнозування;
- використання більших обсягів даних для забезпечення надійності та стабільності моделі;
- інтеграція додаткових факторів, що впливають на донорську активність, як-от сезонні коливання, події соціального значення тощо.

Таке рішення дозволить покращити управління запасами крові в медичних установах, оптимізувати взаємодію з донорами та зменшити ймовірність дефіциту крові в критичних ситуаціях.

ВИСНОВКИ ДО РОЗДІЛУ

Для вирішення задач прогнозування донорства крові доцільно застосовувати нейронні мережі, які дозволяють не лише підвищити точність, але й адаптуватися до динаміки попиту. Це підхід, що не лише оптимізує управління запасами крові, але й сприяє мінімізації ризиків дефіциту крові в критичних ситуаціях.

У подальших дослідженнях доцільно зосередитися на створенні та вдосконаленні гіперпараметрів моделей, збільшенні обсягів навчальних даних та врахуванні додаткових факторів, що впливають на активність донорів, таких як сезонні коливання та соціальні події.

Це дозволить створити надійну систему прогнозування, що покращить ефективність роботи банків крові та якість медичних послуг у випадках підвищеної потреби в донорській крові.

РОЗДІЛ 2. ПРОЕКТУВАННЯ ТА РЕАЛІЗАЦІЯ МОДЕЛІ

2.1 Опис обраного підходу та структури даних

Дані, які використовуються, є даними переливання крові, взяті з донорської бази даних з платформи Kaggle.

Дані зберігаються у файлі Transfusion.data і структуровані відповідно до маркетингової моделі RFMTC (варіант RFM). RFM означає Recency, Frequency і Monetary Value і зазвичай використовується в маркетингу для визначення ваших найкращих клієнтів. У нашому випадку клієнтами є донори крові.

RFMTC є різновидом моделі RFM. Нижче наведено опис значення кожного стовпця в нашому наборі даних:

- R (Останній час - місяці з моменту останньої пожертви)
- F (Частота - загальна кількість пожертвувань)
- M (грошовий – загальна кількість донорської крові в к.к.)
- T (Час - місяці з моменту першої донації)
- двійкова змінна, яка показує, чи здавав він/вона кров у березні 2007 року

Набір даних містить 5 стовпців, як зазначено вище, відповідно, і 748 записів даних, починаючи з 0 до 747. Дані було імпортовано в блокнот Python за допомогою функції pd.read_csv у бібліотеці Pandas у змінній під назвою transfusion. Останній стовпець даних містить двійковий розподіл, у якому 1 означає, що особа здавала кров після березня 2007 року, а 0 означає, що особа не здавала кров після березня 2007 року.

Ми можемо отримати базовий огляд набору даних, свого роду інформацію про нього, за допомогою функції transfusion.info(). Це дає наступний результат, який описує стовпці та повідомляє про нульові значення та їхній тип даних. У нашій базі даних немає нульових значень, а тип даних кожного стовпця – int64.

Результат можна побачити нижче на рисунку:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 748 entries, 0 to 747
Data columns (total 5 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Recency (months)                        748 non-null    int64
 1   Frequency (times)                       748 non-null    int64
 2   Monetary (c.c. blood)                   748 non-null    int64
 3   Time (months)                           748 non-null    int64
 4   whether he/she donated blood in March 2007  748 non-null    int64
dtypes: int64(5)
memory usage: 29.3 KB
```

Рисунок 2.1.1 – База даних без нульових значень

Дані можна переглянути за допомогою transfusion.head(), це дасть перші п’ять рядків набору даних, можна переглянути на рисунку нижче :

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0

Рисунок 2.1.2 – Перші рядки з бази даних

2.2 Підготовка даних

Перед побудовою моделі необхідно було провести попередню обробку даних, яка включала такі етапи:

- Нормалізація даних (рисунок 2.2.1) — приведення значень до однакового діапазону для забезпечення стабільного та швидкого навчання нейронної мережі. Ми проведемо нормалізацію для стовпця Monetary.
- Розподіл на тренувальну і тестову вибірки — виділення приблизно 70% даних для навчання та 30% для тестування з метою оцінки точності моделі. В нашому наборі даних 0 з'являється в 76% випадках. Ми хочемо зберегти однакову структуру в наборах даних навчання та

тестування, тобто обидва набори даних повинні мати 0 цільової частоти 76%. Це дуже легко зробити за допомогою методу train_test_split() з бібліотеки scikit learn — все, що нам потрібно зробити, це вказати параметр stratify. У нашому випадку ми стратифікуємо за стовпцем target. Ми зможемо це побачити на рисунку 2.2.2 що наведено нижче.

- Аугментація та балансування даних — у разі необхідності, особливо якщо кількість позитивних і негативних зразків у наборі даних була нерівною, для забезпечення кращої точності прогнозування.

```
# Import numpy
import numpy as np

# Скопіюйте X_train і X_test в X_train_normed і X_test_normed
X_train_normed, X_test_normed = X_train.copy(), X_test.copy()

# Вкажіть, який стовпець нормалізувати
col_to_normalize = 'Monetary (c.c. blood)'

# Нормалізація журналу
for df_ in [X_train_normed, X_test_normed]:
    # Додати нормалізований стовпець журналу
    df_['monetary_log'] = np.log(df_[col_to_normalize])
    # Видалити оригінальний стовпець
    df_.drop(columns=col_to_normalize, inplace=True)

# Перевірити дисперсію для X_train_normed
X_train_normed.var().round(3)
```

Рисунок 2.2.1 – Нормалізація даних

```
# Import train_test_split method
from sklearn.model_selection import train_test_split

# Розділити переливання DataFrame на
# Набори даних X_train, X_test, y_train і y_test,
# розшарування на стовпець `target`
X_train, X_test, y_train, y_test = train_test_split(
    transfusion.drop(columns='target'),
    transfusion.target,
    test_size=0.25,
    random_state=42,
    stratify=transfusion.target
)

# Роздрукуйте перші 2 рядки X_train
X_train.head(2)
```

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)
334	16	2	500	16
99	5	7	1750	26

Рисунок 2.2.2 – Поділ даних на тренувальну та тестову вибірки

2.3 Пояснення процесу

Отже після завантаження даних , їх нормалізації та розподілу на вибірки, використовуємо TPOT.

Було встановлено та імпортовано бібліотеки TPOT для подальшої обробки набору даних тестування.

TPOT — це інструмент автоматизованого машинного навчання Python, який оптимізує конвеєри машинного навчання за допомогою генетичного програмування. На рисунку 2.3.1 що знаходиться нижче, можна побачити роботу цього інструменту :

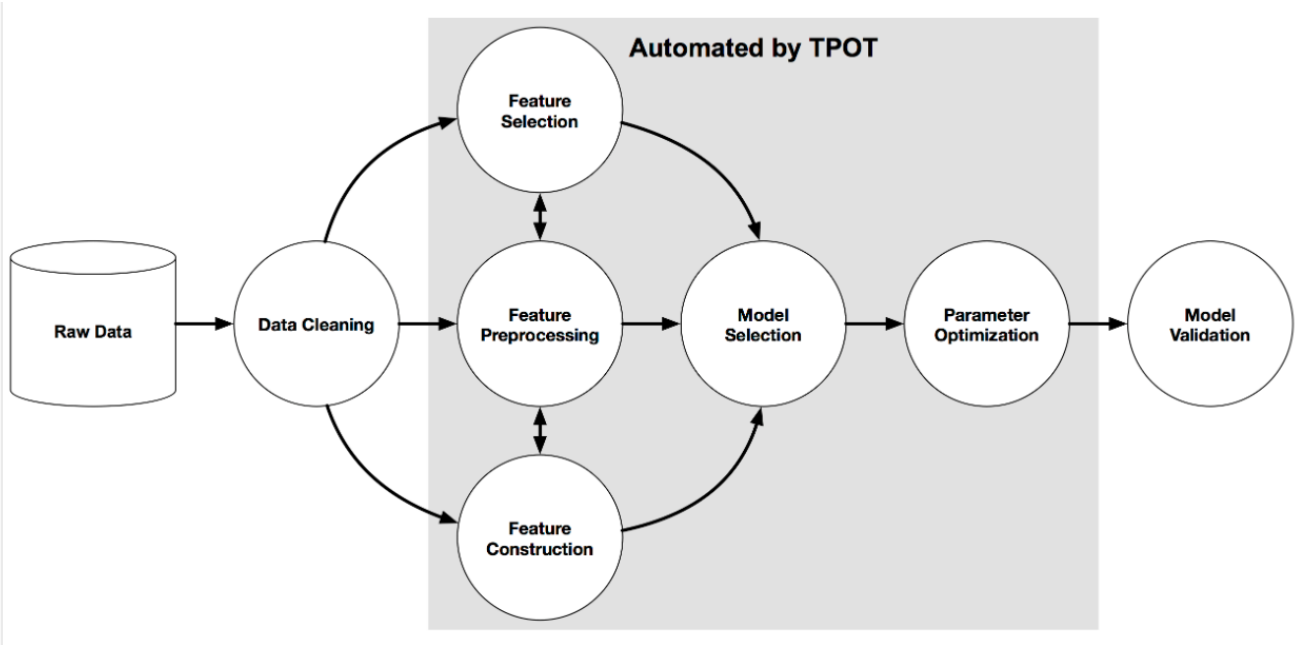


Рисунок 2.3.1 – TPOT

TPOT автоматично досліджуватиме сотні можливих конвеєрів, щоб знайти найкращий для нашого набору даних. Зауважте, що результатом цього пошуку буде конвеєр scikit-learn, тобто він включатиме будь-які етапи попередньої обробки, а також модель.

Використовуємо TPOT, для того щоб зосередитися на одній моделі, яку потім можемо досліджувати та оптимізувати далі.

Наступним кроком була ініціалізація TPOTClassifier, визначивши основні атрибути функції класифікатора, зберегли метод оцінки як roc_auc, визначили випадковий стан, генерацію та розмір популяції

На рисунку 2.3.2 можна побачити код реалізації.

```
# Створення екземпляра TPOTClassifier
tpot = TPOTClassifier(
    generations=5,
    population_size=20,
    verbosity=2,
    scoring='roc_auc',
    random_state=42,
    disable_update_check=True,
    config_dict='TPOT light'
)
```

Рисунок 2.3.2 – Ініціалізація TPOTClassifier

Далі ми підганяємо дані в модель за допомогою tpot.fit(X_train, y_train), після підгонки даних маємо згенерувати оцінку AUC за допомогою порівняння результатів tpot.predict_proba з y_test. Ми також друкуємо найкращі етапи. Результат показаний нижче на рисунку 2.3.3.

```
Error displaying widget

Generation 1 - Current best internal CV score: 0.7422459184429089

Generation 2 - Current best internal CV score: 0.7422459184429089

Generation 3 - Current best internal CV score: 0.7422459184429089

Generation 4 - Current best internal CV score: 0.7422459184429089

Generation 5 - Current best internal CV score: 0.7423330644124078

Best pipeline: LogisticRegression(RobustScaler(input_matrix), C=25.0, dual=False, penalty=l2)

AUC score: 0.7858

Best pipeline steps:
1. RobustScaler()
2. LogisticRegression(C=25.0, random_state=42)
```

Рисунок 2.3.3 – Фітінг даних на TPOT Classifier

TPOT вибрав LogisticRegression як найкращу модель для нашого набору даних без етапів попередньої обробки, що дало нам показник AUC 0,7858. Це досить високий показник.

Одне з припущень для моделей лінійної регресії полягає в тому, що дані та функції, які ми їм надаємо, пов’язані лінійним чином або можуть бути виміряні за допомогою лінійної метрики відстані. Якщо функція в нашому наборі даних має високу дисперсію, яка на порядок або більше перевищує інші функції, це може вплинути на здатність моделі навчатися на основі інших функцій у наборі даних.

Коригування високої дисперсії називається нормалізацією, яку ми виконали на початку навчання моделі для стовпця Monetary.

Тепер підганяємо наш набір даних до моделі логістичної регресії та навчаємо модель, а потім знову отримуємо оцінку AUC, щоб порівняти вихідний прогноз моделі.

2.4 Побудова моделі та прогнозування

Перевага використання моделі логістичної регресії полягає в тому, що її можна інтерпретувати. Ми можемо проаналізувати, яку частину дисперсії у змінній відповіді (цілі) можна пояснити іншими змінними в нашому наборі даних. Так як у нас вже була готова модель TPOT, тут ми створюємо екземпляр моделі логістичної регресії для вже нормалізованих даних.

```
# Importing modules
from sklearn import linear_model

# Створення екземпляра логістичної регресії
logreg = linear_model.LogisticRegression(
    solver='liblinear',
    random_state=42
)

# Тренування моделі
logreg.fit(X_train_normed, y_train)

# Оцінка AUC для моделі tpot
logreg_auc_score = roc_auc_score(y_test, logreg.predict_proba(X_test_normed)[:, 1])
print(f'\nAUC score: {logreg_auc_score:.4f}')
```

```
AUC score: 0.7891
```

Рисунок 2.4.1 – Тренування моделі

І ось ми може мо побачити , що на нормалізованих даних , точність моделі підвищилась на 0.5%. А в машинному навчанні навіть невеличкі покращення це вже добре.

Ми можемо перевірити результат та порівняти його для бібліотеки TPOT і моделі логістичної регресії що зображено на рисунку 2.4.2.

```python
# Importing itemgetter
from operator import itemgetter

# Сортувати моделі на основі їх показника AUC від найвищого до найнижчого
sorted(
    [('tpot', tpot_auc_score), ('logreg', logreg_auc_score)],
    key=itemgetter(1),
    reverse=True)
```

```
[('logreg', 0.7890972663699937), ('tpot', 0.7857596948506039)]
```

Рисунок 2.4.2 – Порівняння результатів AUC

Також в Додатку А можна буде переглянути повний код реалізація даної програми.

Після проведення попередніх етапів підготовки, можна прогнозувати можливість здачі крові для донорів Ось код реалізації для цього і результат його виконання :

```python
# Прогноз для кожного донора в тестовому наборі за допомогою TPOT
tpot_predictions = tpot.predict_proba(X_test)[:, 1]

# Прогноз для кожного донора в тестовому наборі за допомогою логістичної регресії
logreg_predictions = logreg.predict_proba(X_test_normed)[:, 1]

# Створення таблиці з номерами донорів і їх прогнозами
predictions_df = pd.DataFrame({
    'Donor Number': range(1, len(X_test) + 1),
    'TPOT Probability of Donation': tpot_predictions,
    'Logistic Regression Probability of Donation': logreg_predictions
})

# Виведення таблиці
print(predictions_df)
```

Рисунок 2.4.3 – Прогнозування для кожного донора

	Donor Number	TPOT Probability of Donation	\
0	1	0.406869	
1	2	0.145362	
2	3	0.392716	
3	4	0.382725	
4	5	0.358446	
..	...	...	
182	183	0.308726	
183	184	0.114557	
184	185	0.050511	
185	186	0.172380	
186	187	0.273333	

	Logistic Regression Probability of Donation
0	0.452876
1	0.142197
2	0.434499
3	0.414502
4	0.339023
..	...
182	0.288582
183	0.117598
184	0.041835
185	0.157384
186	0.299046

Рисунок 2.4.4 – Результат прогнозування для кожного донора

Також щоб отримати загальний прогноз відсотка донорів, які, ймовірно, прийдуть здавати кров, можна обчислити середнє значення прогнозованих ймовірностей для тестового набору даних, що зображено на рисунку нижче.

```
# Обчислення середньої ймовірності здачі крові для TPOT та логістичної регресії
tpot_mean_probability = tpot_predictions.mean() * 100
logreg_mean_probability = logreg_predictions.mean() * 100

# Виведення загального прогнозу/
print(f"Загальний прогноз (TPOT): {tpot_mean_probability:.2f}% донорів")
print(f"Загальний прогноз (Logistic Regression): {logreg_mean_probability:.2f}% донорів")
```

```
Загальний прогноз (TPOT): 24.06% донорів
Загальний прогноз (Logistic Regression): 24.58% донорів
```

Рисунок 2.4.5 – Загальний прогноз

ВИСНОВКИ ДО РОЗДІЛУ

У цьому розділі було розглянуто основні етапи проєктування та реалізації моделі прогнозування активності донорів крові. Початковий аналіз набору даних дозволив визначити його структуру та ключові показники (RFMTC-модель), які відображають поведінку донорів. Ці дані були підготовлені до моделювання шляхом нормалізації, що забезпечило стабільність і точність подальшого навчання, а також шляхом стратифікованого розподілу на тренувальну та тестову вибірки для збереження балансу класів.

Використання інструменту TPOT дозволило автоматизувати пошук оптимальної моделі, забезпечивши прискорення розробки та тестування різних підходів. Визначення логістичної регресії як найкращої моделі на початковому етапі, а також подальше її покращення через нормалізацію підтвердили правильність обраного підходу. Порівняння результатів AUC також вказало на суттєве зростання точності після обробки даних, що свідчить про важливість нормалізації і збалансованості вибірок для якісного прогнозування.

Отже, реалізована модель є достатньо точною та інтерпретованою, що робить її придатною для прогнозування активності донорів у майбутньому. Результати підтверджують ефективність автоматизованого підходу TPOT і засвідчують доцільність застосування логістичної регресії для розв'язання задачі, пов’язаної з прогнозуванням на основі поведінкових даних.

РОЗДІЛ 3. ВІЗУАЛІЗАЦІЯ ТА РЕЗУЛЬТАТИ

3.1 Візуалізація даних

Візуалізація та статистичний аналіз є основною частиною будь-якого проекту аналізу даних. Першим кроком є розуміння набору даних і проблеми, з якою ми маємо справу. Візуалізація даних і результатів на кожному кроці дає нам гарне уявлення про рішення.

На першому кроці ми використали метод countplot бібліотеки seaborn, ми підрахували кількість 1 і 0 у стовпці «погода, коли він/вона здавав кров у березні 2007 року», потім цей стовпець змінився на «target». Сюжет показує нам розподіл людей, які здали кров і не здали в березні. Ми чітко бачимо на рисунку 3.1, що найбільше людей не здали кров у березні.

```
transfusion.target.value_counts()
```

```
target
0    570
1    178
Name: count, dtype: int64
```

Рисунок 3.1.1 – Підрахунок кількості людей які здали кров

Наведений вище фрагмент коду представляє кількість значень у цільовому стовпці, він дає кількість 0 і 1. Те ж саме представлено нижче за допомогою методу countplot. Цей розподіл показує, що існує величезна різниця між цінностями людей, які здали кров у березні, та людей, які цього не зробили. Це свідчить про те, що більшість людей не здавали кров у березні 2007 року.

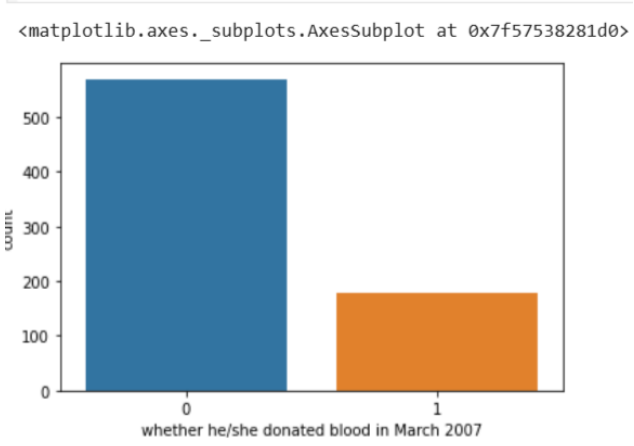


Рисунок 3.1.2 – Підрахунок кількості зданої крові з методом countplot

Також отримали базовий підсумок набору даних на початковому етапі аналізу даних за допомогою команди describe transfusion.describe(), вона дає підрахунок даних у кожному рядку, середнє значення даних і багато таких функцій, як стандартне відхилення , мінімум і максимум тощо.

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
count	748.000000	748.000000	748.000000	748.000000	748.000000
mean	9.506684	5.514706	1378.676471	34.282086	0.237968
std	8.095396	5.839307	1459.826781	24.376714	0.426124
min	0.000000	1.000000	250.000000	2.000000	0.000000
25%	2.750000	2.000000	500.000000	16.000000	0.000000
50%	7.000000	4.000000	1000.000000	28.000000	0.000000
75%	14.000000	7.000000	1750.000000	50.000000	0.000000
max	74.000000	50.000000	12500.000000	98.000000	1.000000

Рисунок 3.1.3 – Підсумок набору даних

Тут бачимо, що кількість рядків становить 748, що означає, що в наборі даних немає пропущених значень. Також бачимо, що в середньому кожен здавав кров 5 разів протягом 10 місяців.

Кореляція даних: це спосіб зрозуміти зв’язок між кількома змінними та атрибутами в нашому наборі даних. Використовуючи кореляцію, ми можемо отримати певну інформацію, наприклад:

- Один або кілька атрибутів залежать від іншого атрибута або причини іншого атрибута.
- Один або декілька атрибутів пов’язані з іншими атрибутами.

Таблиця нижче є кореляційною таблицею нашого набору даних, її генерує функція corr(). Як ми бачимо в наведеній нище таблиці, ми маємо як позитивні, так і негативні значення. Позитивна кореляція: означає, що якщо ознака A збільшується, то ознака B також збільшується, або якщо ознака A зменшується, то функція B також зменшується. Обидві функції рухаються в тандемі та мають лінійний зв’язок.

Також можна побачити додаткові графіки у додатку Б.

Негативна кореляція: означає, що якщо ознака A збільшується, то ознака B зменшується, і навпаки.

```
transfusion.corr()
```

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	target
Recency (months)	1.000000	-0.182745	-0.182745	0.160618	-0.279869
Frequency (times)	-0.182745	1.000000	1.000000	0.634940	0.218633
Monetary (c.c. blood)	-0.182745	1.000000	1.000000	0.634940	0.218633
Time (months)	0.160618	0.634940	0.634940	1.000000	-0.035854
target	-0.279869	0.218633	0.218633	-0.035854	1.000000

Рисунок 3.1.4 – Кореляційна таблиця

Ще можемо представити кореляцію через теплову карту в морській бібліотеці. Вона представляє кореляцію у форматі кольорового кодування та дає легенду кольорів збоку від графіка. Можна побачити по рисунку 3.1.5 що ми маємо позитивну кореляцію між частотою та часом. Також бачимо значення кореляції в кожній клітинці, поставивши значення annot = True.

```
sns.heatmap(transfusion.corr(), annot=True, linewidths=0.3, cmap="YlGnBu")
```

```
<Axes: >
```

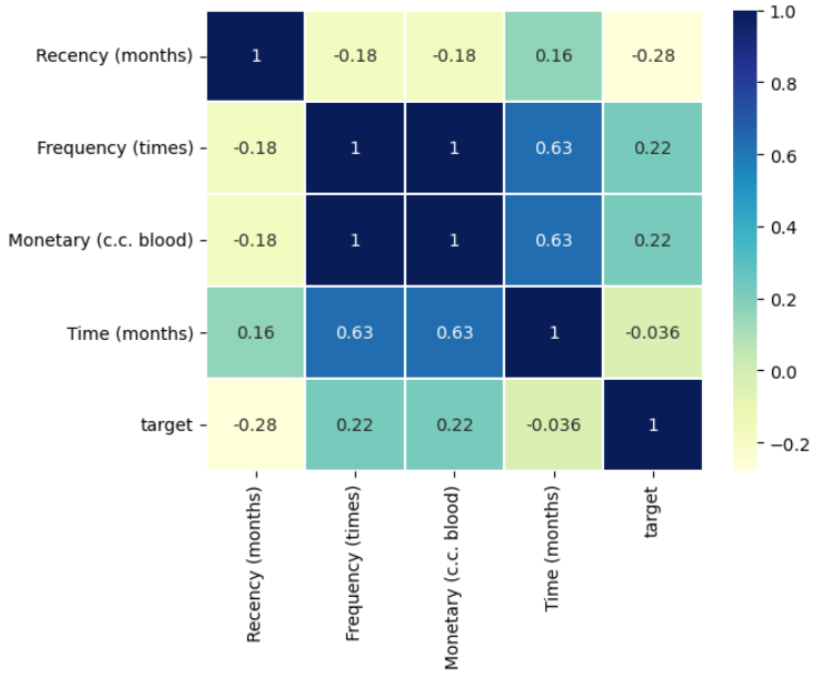


Рисунок 3.1.5 – Кореляція через теплову карту

3.2 Результати прогнозування

Після виконання попередніх пунктів, можна за рахунок створених моделей прогнозувати кількість донорів які можливо здадуть кров у наступному місяці. На рисунку нижче зображено результати прогнозування, та можливість здати кров для кожного донора:

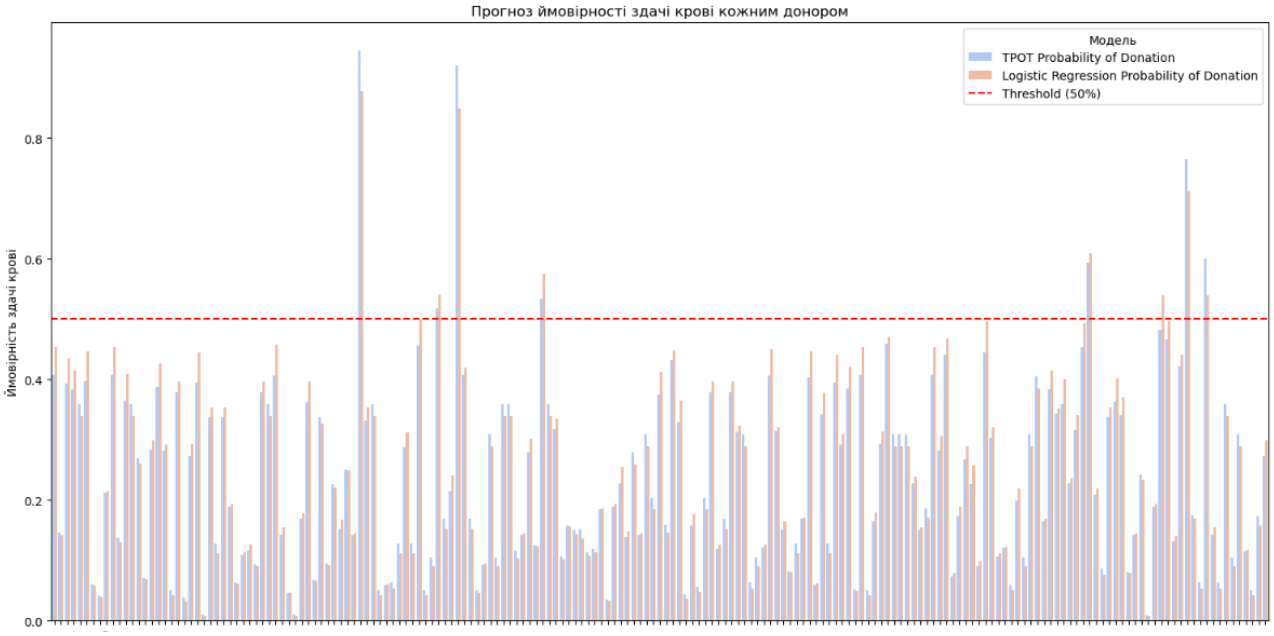


Рисунок 3.2.1 – Прогнозування можливості здачі крові для кожного донора

Ще одним корисним графіком для аналізу може бути ROC-крива для порівняння продуктивності двох моделей — TPOT та Logistic Regression. ROC-крива допомагає оцінити здатність моделі відрізняти донорів, які здадуть кров, від тих, хто не здасть, при різних порогах класифікації.

Модель з більшою AUC краще відрізняє донорів, які здадуть кров, від тих, хто не здасть. ROC-крива, що знаходиться ближче до верхнього лівого кута, показує вищу ефективність моделі (рисунок 3.2.2).

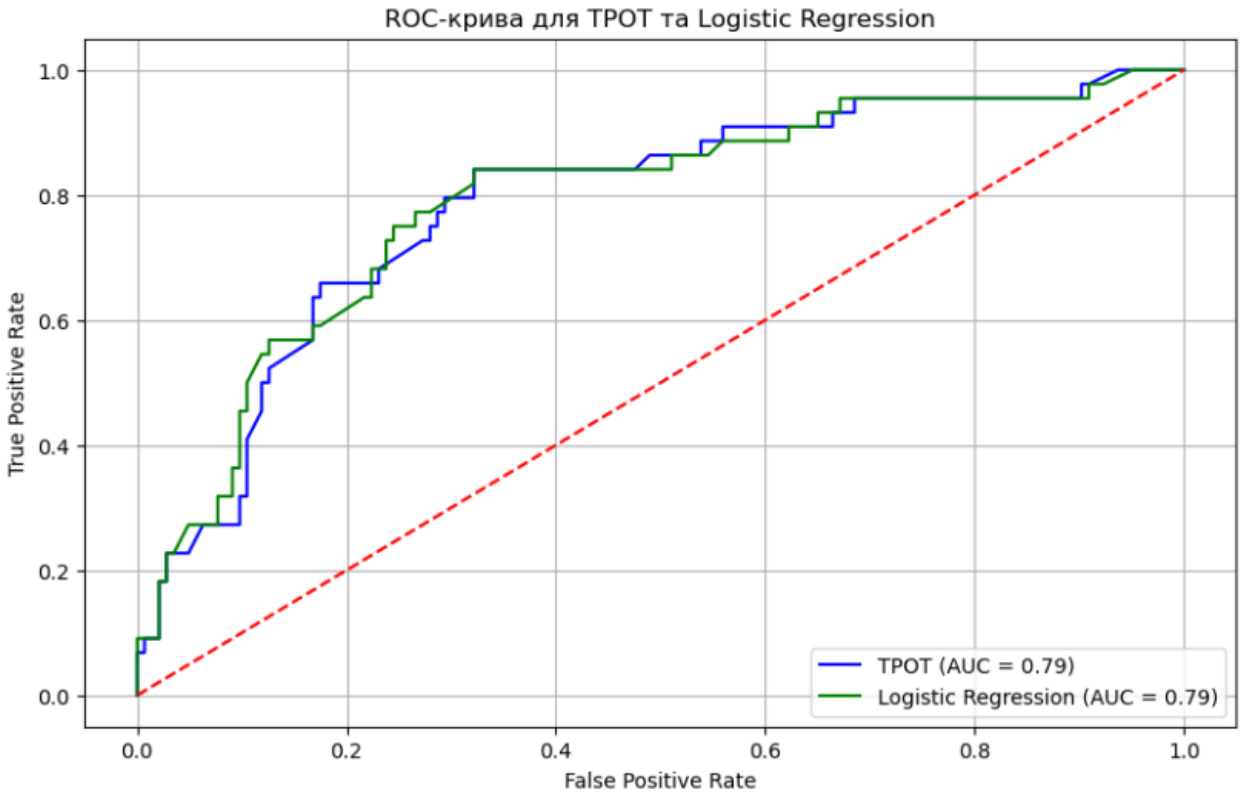


Рисунок 3.2.2 – Прогнозування можливості здачі крові для кожного донора

True Positive Rate (TPR) і False Positive Rate (FPR): ROC-крива показує співвідношення між TPR і FPR при різних порогах класифікації.

AUC (Area Under the Curve): Значення AUC допомагає оцінити загальну ефективність моделі. AUC, наближене до 1, свідчить про кращу якість класифікації.

Діагональна лінія: Червона пунктирна лінія представляє випадкову модель (AUC = 0.5).

ВИСНОВКИ ДО РОЗДІЛУ

У цьому розділі було проведено візуалізацію та статистичний аналіз, що допомогли глибше зрозуміти набір даних про донорів крові та його основні закономірності. Використання методів візуалізації, таких як countplot, дозволило чітко побачити дисбаланс у кількості донорів, які здали кров у березні 2007 року, і тих, хто не здавав. Це дозволяє зробити висновок про нерівномірний розподіл вибірки, що важливо враховувати для подальшого аналізу і моделювання.

Дослідження статистичних характеристик та використання методів кореляційного аналізу виявили важливі взаємозв’язки між змінними, зокрема, позитивну кореляцію між частотою донорства та часом від першої донації. Візуалізація у вигляді теплової карти також допомогла наочно показати ці зв'язки та їхню силу, що є цінним інструментом для виявлення лінійних залежностей.

Отримані результати вказують на те, що деякі показники в наборі даних можуть відігравати ключову роль у прогнозуванні поведінки донорів, що є основою для подальшого вдосконалення моделі. Статистичні і візуальні методи аналізу даних продемонстрували свою ефективність і підкреслили необхідність обробки та врахування зв’язків між атрибутами для точнішого прогнозування.

РОЗДІЛ 4. ПЕРСПЕКТИВИ ВДОСКОНАЛЕННЯ МОДЕЛІ

4.1 Переваги обраного підходу та інтерпретація результатів

Логістична регресія, як і обраний підхід із використанням TPOT, має кілька вагомих переваг. По-перше, автоматизований підхід за допомогою TPOT спростив пошук оптимальних гіперпараметрів і методів попередньої обробки, що прискорило процес розробки. Модель логістичної регресії, крім того, є інтерпретованою, тобто дозволяє визначити, як саме кожна змінна впливає на прогноз ймовірності донації крові. Такий інтерпретований результат особливо корисний у сфері охорони здоров'я, адже дозволяє бачити вагомість різних характеристик донорів у процесі ухвалення рішень.

Виконана нормалізація даних позитивно вплинула на продуктивність моделі, знизивши можливий негативний вплив великої дисперсії у змінних, таких як кількість зібраної крові.

Однак, модель все ще залишилася чутливою до дисбалансу у вибірці — більшість донорів не здавали кров у березні 2007 року, що створює перекіс, відображений у прогнозах.

4.2 Перспективи вдосконалення моделі

Тут запропоновані кілька ідей для покращення та вдосконалення моделі в майбутньому :

- Балансування вибірки: Можна розглянути використання методів аугментації даних або синтетичного створення нових позитивних зразків, наприклад, за допомогою алгоритму SMOTE (Synthetic Minority Over-sampling Technique). Це дозволить збільшити частку позитивних зразків у тренувальному наборі, що може підвищити точність прогнозування для менш представленого класу.
- Використання більш складних моделей: Хоча TPOT обрав логістичну регресію як оптимальну модель, варто перевірити, як такі методи, як випадковий ліс або градієнтний бустинг, могли б вплинути на точність, особливо в умовах значного дисбалансу вибірки. Ці методи є менш

чутливими до дисбалансу даних і можуть дати вищу продуктивність, хоч і з більшою складністю.

- Оптимізація гіперпараметрів: Поглиблене налаштування гіперпараметрів, особливо за допомогою ручного налаштування або використання більш продовженого часу навчання в TPOT, може допомогти підвищити продуктивність моделі. Збільшення кількості поколінь та розміру популяції під час оптимізації за допомогою генетичних алгоритмів TPOT може покращити результати, забезпечуючи більш детальне дослідження простору параметрів.
- Інтеграція додаткових ознак: Оскільки набір даних обмежений кількома параметрами, було б доцільно додати більше атрибутів, таких як дані про інші акції для донорів крові, інформацію про інтервали між донаціями, демографічні дані тощо. Це могло б розширити модель, зробити її більш інформативною та покращити її прогностичні можливості.

4.3 Практичне застосування моделі

Розроблена модель прогнозування, що враховує основні фактори поведінки донорів, може суттєво покращити діяльність Центру служби переливання крові. Її застосування дозволяє ефективніше планувати залучення донорів та акції, допомагаючи запобігти дефіциту або надлишку крові, що особливо важливо в умовах критичних ситуацій. Зокрема, можливості покращеної моделі охоплюють кілька ключових напрямків практичного застосування:

– Планування та оптимізація донаційних акцій. На основі точних прогнозів про ймовірність повторної донації крові Центр може краще планувати кампанії, спрямовані на залучення саме тих донорів, які з більшою ймовірністю відгукнуться на заклики до здачі крові. Модель дозволяє передбачити, коли донори з найбільшою ймовірністю зможуть зробити наступний внесок, що є корисним для управління

донаційними акціями та уникнення випадкових пікових навантажень на Центр.

- Індивідуалізація підходів до комунікації з донорами. Інтерпретована модель дозволяє визначити основні фактори, що впливають на бажання або здатність донорів здавати кров. Це відкриває можливість для створення персоналізованих пропозицій для кожного донора, що можуть включати спеціальні запрошення, нагадування про можливість донації або інформацію про особливі акції. Така персоналізація здатна зміцнити зв’язок із донорами, підвищуючи їх лояльність і мотивацію до регулярного внеску.
- Розподіл ресурсів та управління запасами крові. Прогнозуючи поведінку донорів, Центр може краще розподіляти ресурси на підготовку та організацію збору крові. Наприклад, дані моделі можуть вказувати на дати з очікуваним високим потоком донорів, що дозволить ефективніше підготувати запаси, забезпечити наявність необхідного персоналу та обладнання. Це допоможе уникнути нестачі певних груп крові або, навпаки, надмірного запасу, що може призвести до втрати частини заготовленої крові.
- Ідентифікація важливих груп донорів для цільових кампаній. Модель може допомогти виявити групи донорів, які мають високий потенціал до регулярного донорства або можуть потребувати спеціальних стимулів. Наприклад, групи, що раніше здавали кров рідко, але мають значний потенціал, можуть бути включені в цільові кампанії, які пропонують їм додаткову інформацію або заохочення, орієнтовані на конкретні потреби.
- Оцінка ефективності маркетингових та соціальних кампаній. Модель прогнозування може бути використана для оцінки результатів акцій та маркетингових зусиль, спрямованих на залучення донорів. Порівняння прогнозованих даних з фактичними результатами дозволяє оцінити ефективність різних видів заходів, виявляючи найбільш дієві практики

та кампанії. Такий аналіз допоможе Центру коригувати свої стратегії для досягнення кращих результатів у залученні донорів.

- Попередження кризових ситуацій та підтримка системи охорони здоров’я. Прогнозуючи тенденції у донорстві, модель здатна своєчасно виявити потенційні періоди дефіциту, що дозволяє Центру заздалегідь вжити заходів для його запобігання. Це особливо важливо в умовах надзвичайних ситуацій, коли попит на донорську кров може різко зрости. Своєчасне залучення донорів, підтримка належних запасів крові та забезпечення ефективного обслуговування пацієнтів можуть значно підвищити готовність системи охорони здоров’я до таких викликів.
- Сприяння суспільній обізнаності про донорство крові. Інформація, отримана з використанням прогнозної моделі, може бути корисною для розробки програм суспільної обізнаності та просвітництва. Визначаючи ключові мотивуючі чинники та бар'єри для донорів, Центр може створити інформаційні матеріали, які більшою мірою відповідають інтересам і потребам цільової аудиторії. Це сприятиме формуванню позитивного ставлення до донорства та залученню нових донорів.

Таким чином, модель прогнозування має широкий потенціал для підвищення ефективності діяльності Центру служби переливання крові. Вона дозволяє глибше розуміти поведінку донорів, оптимізувати операційні процеси та забезпечувати більш надійне задоволення потреб системи охорони здоров’я.

ВИСНОВКИ ДО РОЗДІЛУ

У цьому розділі було проведено детальне обговорення результатів моделювання, розглянуто переваги обраного підходу та визначено можливі напрямки вдосконалення моделі прогнозування поведінки донорів. Використання TPOT для автоматизованої оптимізації параметрів дозволило ефективно підібрати модель та підвищити її продуктивність. Завдяки інтерпретованій логістичній регресії ми змогли встановити найбільш значущі

чинники, що впливають на ймовірність повторної донації, що особливо цінно для практичних цілей у сфері охорони здоров’я.

Запропоновані перспективи покращення моделі, включаючи балансування вибірки, використання складніших моделей, оптимізацію гіперпараметрів та інтеграцію додаткових ознак, відкривають можливості для підвищення точності прогнозування та забезпечення більш точного управління донаціями. Практичне застосування цієї моделі в роботі Центру служби переливання крові надає інструменти для вдосконалення планування та організації збору крові, підтримки достатніх запасів та ефективного залучення донорів.

Таким чином, представлена модель має вагоме значення для системи переливання крові, зокрема для покращення обслуговування пацієнтів та підвищення надійності системи охорони здоров’я загалом. Подальше вдосконалення моделі може розширити її можливості, що забезпечить кращу адаптацію до змінюваних потреб, оптимізацію ресурсів та сприятиме популяризації донорства в суспільстві.

ВИСНОВКИ

У цій роботі розглянуто процес створення, аналізу та вдосконалення моделі прогнозування поведінки донорів крові з метою оптимізації роботи Центру служби переливання крові.

Протягом дослідження виконано кілька етапів аналізу та моделювання, що дозволило не тільки підвищити ефективність процесу залучення донорів, але й створити умови для більш надійного задоволення потреб системи охорони здоров’я у донорській крові.

Підсумовуючи основні результати, можна виділити кілька ключових аспектів:

- Аналіз та підготовка даних: Проведено очищення, нормалізацію та обробку вихідного набору даних, що забезпечило якість та стабільність даних для подальшого моделювання. Використання візуалізацій дозволило глибше зрозуміти структуру даних, зокрема розподіл донорів за часом та характеристиками, що стало основою для прийняття рішень на наступних етапах.
- Розробка моделі та автоматизований підхід до оптимізації: Логістична регресія, обрана як основна модель у поєднанні з автоматизованим підходом TPOT, дозволила створити інтерпретовану та ефективну модель прогнозування. Використання TPOT для підбору параметрів значно прискорило процес розробки та дозволило досягти оптимальної продуктивності.
- Аналіз результатів і перспективи вдосконалення: Розглянуто ключові фактори, що впливають на точність моделі, такі як дисбаланс класів та вибір гіперпараметрів. Запропоновані подальші кроки вдосконалення, включаючи балансування вибірки, використання більш складних моделей та інтеграцію додаткових атрибутів, що можуть суттєво покращити точність прогнозування в майбутньому.
- Практичне застосування моделі: Виявлені можливості практичного використання моделі для Центру служби переливання крові. Модель

може слугувати інструментом для планування акцій, оптимізації комунікації з донорами, управління ресурсами та прогнозування потреб у крові. Вона також відкриває можливість для покращення суспільної обізнаності про донорство та персоналізації підходів до донорів, що може підвищити лояльність донорів і мотивацію до регулярної участі.

Таким чином, результати дослідження вказують на значний потенціал застосування прогнозної моделі у роботі Центру служби переливання крові.

Завдяки виявленню чинників, що впливають на поведінку донорів, і розробці відповідних рекомендацій, дана модель може допомогти Центру ефективніше управляти процесом збору крові, забезпечуючи своєчасне задоволення попиту та стабільне функціонування системи охорони здоров’я.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Smart Platform for Data Blood Bank Management: Forecasting Demand in Blood Supply Chain Using Machine Learning. MDPI. URL: https://www.mdpi.com/2078-2489/14/1/31
2. scikit-learn: machine learning in Python. scikit-learn: machine learning in Python – scikit-learn 0.16.1 documentation. URL: https://scikit-learn.org/stable/
3. TPOT. URL: https://automl.info/tpot/
4. Здай кров - врятуй життя!. ДонорUA. URL: https://www.donor.ua/
5. Blood Transfusion: What to Know If You Get One. WebMD. URL: https://www.webmd.com/a-to-z-guides/blood-transfusion-what-to-know#1
6. American Red Cross - Blood Donation Process. URL : https://www.redcrossblood.org/donate-blood/how-to-donate/how-blood-donations-help.html
7. UCI Machine Learning Repository. UCI Machine Learning Repository. URL: https://archive.ics.uci.edu/dataset/176/blood+transfusion+service+center
8. Narkhede S. Understanding AUC - ROC Curve. Medium. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
9. Machine learning tutorial: Model selection. Neural Designer. URL: https://www.neuraldesigner.com/learning/tutorials/model-selection/
10. Target blood donors using machine learning. Neural Designer. URL: https://www.neuraldesigner.com/learning/examples/blood-donors-targeting/

ДОДАТКИ

ДОДАТОК А

Код програми

```
# Import pandas
import pandas as pd
# Зчитування датасету
transfusion = pd.read_csv('D:/курсач/transfusion (2).csv')
# Роздрукуйте перші рядки нашого набору даних
transfusion.head()

# Надрукуйте стислий підсумок переливання DataFrame
transfusion.info()
transfusion.describe()

import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(20,6))
sns.boxplot(x = 'Frequency (times)', y = 'Recency (months)',hue = 'target', data =
transfusion)

plt.figure(figsize=(20,6))
sns.lineplot(x='Frequency (times)', y='Recency (months)', hue='target', data =
transfusion)

plt.figure(figsize=(10,5))
sns.boxplot(x = 'target', y = 'Recency (months)', data = transfusion)

sns.heatmap(transfusion.corr(), annot=True, linewidths=0.3, cmap="YlGnBu")
```

```
transfusion.corr()

# Перейменуйте цільовий стовпець на 'target' для стислості
transfusion.rename(
    columns={'whether he/she donated blood in March 2007': 'target'},
    inplace=True
)
# Виведемо 2 рядки
transfusion.head(2)

# Надрукувати пропорції падіння цілі, округляючи результат до 3 знаків після коми
transfusion.target.value_counts(normalize=True).round(3)

transfusion.target.value_counts()

# Import train_test_split method
from sklearn.model_selection import train_test_split
# Розділити переливання DataFrame на
# Набори даних X_train, X_test, y_train і y_test,
# розшарування на стовпець `target`
X_train, X_test, y_train, y_test = train_test_split(
    transfusion.drop(columns='target'),
    transfusion.target,
    test_size=0.25,
    random_state=42,
    stratify=transfusion.target
)
# Роздрукуйте перші 2 рядки X_train
X_train.head(2)
```

```
# Import TPOTClassifier and roc_auc_score
from tpot import TPOTClassifier
from sklearn.metrics import roc_auc_score
# Створення екземпляра TPOTClassifier
tpot = TPOTClassifier(
    generations=5,
    population_size=20,
    verbosity=2,
    scoring='roc_auc',
    random_state=42,
    disable_update_check=True,
    config_dict='TPOT light'
)
tpot.fit(X_train, y_train)

# Оцінка AUC для моделі tpot
tpot_auc_score = roc_auc_score(y_test, tpot.predict_proba(X_test)[:, 1])
print(f'\nAUC score: {tpot_auc_score:.4f}')

# Друк найкращих кроків конвеєра
print('\nBest pipeline steps:', end='\n')
for idx, (name, transform) in enumerate(tpot.fitted_pipeline_.steps, start=1):
    # Вивести idx і перетворити
    print(f'{idx}. {transform}')

# Дисперсія X_train, округлення результату до 3 знаків після коми
X_train.var().round(3)

# Import numpy
```

```
import numpy as np
# Скопіюйте X_train і X_test в X_train_normed і X_test_normed
X_train_normed, X_test_normed = X_train.copy(), X_test.copy()
# Вкажіть, який стовпець нормалізувати
col_to_normalize = 'Monetary (c.c. blood)'

# Нормалізація журналу
for df_ in [X_train_normed, X_test_normed]:
    # Додати нормалізований стовпець журналу
    df_['monetary_log'] = np.log(df_[col_to_normalize])
    # Видалити оригінальний стовпець
    df_.drop(columns=col_to_normalize, inplace=True)

# Перевірити дисперсію для X_train_normed
X_train_normed.var().round(3)

# Importing modules
from sklearn import linear_model
# Створення екземпляра логістичної регресії
logreg = linear_model.LogisticRegression(
    solver='liblinear',
    random_state=42
)

# Тренування моделі
logreg.fit(X_train_normed, y_train)

# Оцінка AUC для моделі tpot
logreg_auc_score = roc_auc_score(y_test, logreg.predict_proba(X_test_normed)[:,
1])
```

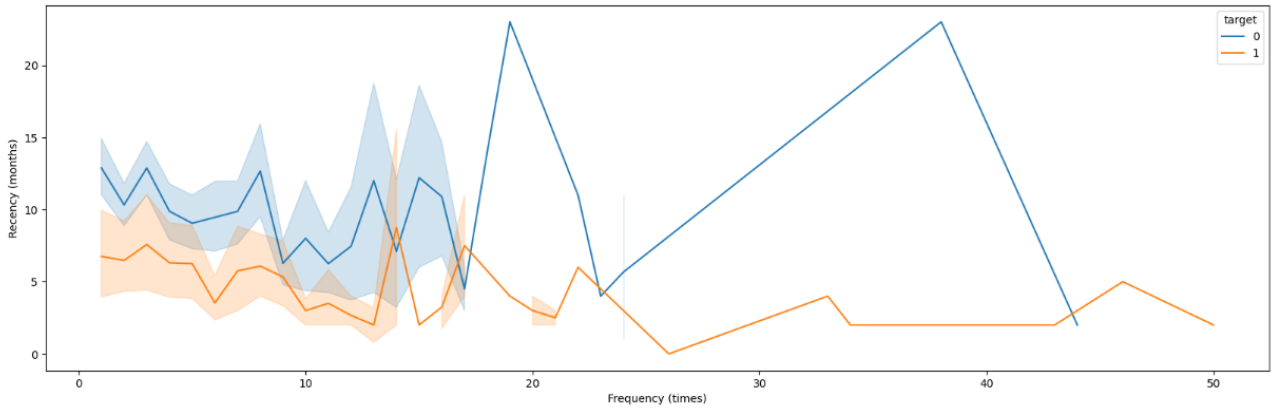
```
print(f'\nAUC score: {logreg_auc_score:.4f}')

# Importing itemgetter
from operator import itemgetter
# Сортувати моделі на основі їх показника AUC від найвищого до
найнижчого
sorted(
    [('tpot', tpot_auc_score), ('logreg', logreg_auc_score)],
    key=itemgetter(1),
    reverse=True)
```

ДОДАТОК Б

Додаткові графіки

<Axes: xlabel='Frequency (times)', ylabel='Recency (months)'>



<Axes: xlabel='target', ylabel='Recency (months)'>