Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Факультет інформатики
Кафедра інформатики

Застосування нейронних мереж для виявлення та аналізу захворювань на цукровий діабет

Текстова частина до кваліфікаційної роботи
за спеціальністю 121 «Інженерія програмного забезпечення»

Керівник кваліфікаційної роботи
Бучко О. А.
(*прізвище та ініціали*)

_____
(*підпис*)
“____”_____ 2024 р.

Виконав студент
Смовженко А.О.
(*прізвище та ініціали*)

_____
(*підпис*)
“____” _____ 2024 р.

Київ-2024

Тема: Застосування нейронних мереж для виявлення та аналізу захворювань на цукровий діабет

Календарний план виконання роботи:

№	Назва етапу курсової роботи	Термін виконання	Примітка
1.	Отримання завдання на кваліфікаційну роботу	вересень	
2.	Огляд літератури за темою роботи	жовтень-грудень	
3.	Створення практичної частини роботи	січень-квітень	
4.	Написання текстової частини роботи	березень-квітень	
6.	Надання роботи керівнику для перевірки	кінець квітня	
8.	Коригування роботи за результатами перевірки	кінець квітня-початок травня	
9	Подання роботи на кафедру для перевірки на плагіат	тиждень до захисту	

10.	Захист кваліфікаційної роботи	кінець травня	

Студент Смовженко А. О.

Керівник Бучко О. А.

“____” _____

Зміст

Вступ 5
Розділ 1 Теоретичні основи нейронних мереж 7
Розділ 2 Аналітика й перспективи нейроних мереж в медицині 14
Розділ 2.1 Перспективи нейроних мереж в медицині 14
Розділ 2.2 Аналітика і статистика датасету.....15
Розділ 3 Практична реалізація 22
Розділ 3.1 Перетворення даних для моделі 22
Розділ 3.2 Побудова моделей, їх оцінка, вибір найточнішої.....25
Розділ 4 Побудова фінальної моделі.....27
Розділ 4.1 Збереження отриманої моделі 29
Висновок 30
Джерела:.....32

Вступ

Штучний інтелект стрімко розвивається в сучасному світі й вважається найновішим прогресом людства. Тема медицини, лікування як такого – завжди була, є й буде актуальною, бо це необхідний аспект життя людини.

А що як штучний інтелект натренувати на основі алгоритмів ставити діагнози на основі аналізів людини, маючи історію пацієнта вираховувати вірогідність того чи іншого захворювання?

Якщо мати справді таку працюючу нейронну мережу в лікарнях, то можна збільшити якість й швидкість самого лікування – при наданні даних аналізів, історії пацієнта, його способу життя - можна буде поставити діагноз, прописати саме лікування, надати рекомендації для зниження вірогідності виникнення ускладнень чи інших хвороб.

З цього можна сказати що за цим майбутнє медицини, в даній роботі за мету поставлено зробити міні-модель функціоналу штучного інтелекту для виявлення та аналізу захворювань на основі показників людини.

Побудова й тренування моделі починається з аналізу даних, а аналіз даних включає в себе:

розуміння джерела походження даних (яким чином вони збирались та за яких умов);

розбір атрибутів що містить датасет, їх специфіка;

перевірка якості даних, тобто чи є пусті значення, чи є дані що занадто сильно виділяються з датасету (Outliers).

Для дослідження було обрано датасет CDC Diabetes Health Indicators[i]

Його опис:

Набір даних про показники здоров’я діабету містить статистику охорони здоров’я та інформацію про спосіб життя людей загалом разом із діагнозом діабету. 21 показник складаються з деяких демографічних даних, результатів лабораторних тестів і відповідей на запитання опитування для кожного пацієнта. Цільовою змінною для класифікації є наявність у пацієнта цукрового діабету, переддіабетичного стану або здоровий.

Дані були зібрані CDC (Центри з контролю та профілактики захворювань у США)

Розділ 1 Теоретичні основи нейронних мереж

Розуміння даних, їх розбір

Аби чітко й послідовно розробити план дій та побачити картину в цілому – треба розібрати усі метрики що є в датасеті. Після дослідження метрик можна ідейно зрозуміти що саме підлягає дослідженню та на які групи та сегменти варто розділити нашу базу даних для подальшого якісного проведення аналітики.

Надалі респондентів даної бази даних також будемо найменувати як пацієнтів, в цьому випадку будемо вважати це як за синоніми.

1. ID – це id пацієнта, тобто порядковий номер.

2. Diabetes_binary – бінарне значення, що вказує чи було пацієнту діагностовано прediабет або діабет, чи ні. 0 – нема діабету, 1 – є. (Варто зазначити що прediабет не тягне за собою яскраво виражених симптомів, тобто в більшості випадків люди не знають про наявність цього ступеня хвороби. Зазвичай дізнаються постфактум що у них був прediабет – вже коли набули сам діабет. Це і є одна з причин збору даних пацієнтів для визначення або прогнозування наявності прediабету в людини.)

3. HighBP – фактор наявності високого кров'яного тиску (гіпертонія). Дані записоні також бінарним чином: 0 – нема високого тиску, 1 – є.

4. HighChol – високий холестерин, 0 – не є високим, 1 – є.

5. CholCheck - вказує, чи перевіряли людину рівень холестерину протягом останніх п’яти років. 0 – не було перевірки, 1 – була.

6. BMI - Індекс маси тіла (Body Mass Index), вираховується за формулою: $BMI=kg/m^2$
   Тобто маса людини в кілограмах ділена на квадрат зросту в метрах.
   Прийнято виділяти такі категорії:
   16 і менше - Виражений дефіцит маси тіла
   16—18,5 - Недостатня (дефіцит) маса тіла
   18,5—25 - Норма
   25—30 - Зайва маса тіла (предожиріння)
   30—35 - Ожиріння 1 ступеня
   35—40 - Ожиріння 2 ступеня
   40 і більше - Ожиріння 3 ступеня

7. Smoker – бінарний показник що вказує на відповідь на питання: “Чи викурили ви принаймні 100 цигарок за все своє життя?”. 0 – відповідь ні, 1 – так.

8. Stroke - бінарний показник що вказує чи коли-небудь був у пацієнта інсульт, 0 – не було ,1 – був.

9. HeartDiseaseorAttack – вказує чи була у пацієнта коронарна хвороба серця (coronary heart disease) або інфаркт міокарда (myocardial infarction). 0 – не була ,1 – була.

10. PhysActivity - бінарний показник що вказує на відповідь на питання: “Чи була фізична активність протягом останніх 30 днів, не враховуючи роботу?”. 0 – відповідь ні, 1 – так.

11. Fruits - бінарний показник що вказує на відповідь на питання: “Чи ви споживаєте фрукти один або більше разів на день?”. 0 – відповідь ні, 1 – так.

12. Veggies - бінарний показник що вказує на відповідь на питання: “Чи ви споживаєте овочі один або більше разів на день?”. 0 – відповідь ні, 1 – так.

13. HvyAlcoholConsump – бінарний показник на питання про вживання великої кількості алкоголю (дорослі чоловіки, які вживають більше 14 напоїв на тиждень, і дорослі жінки, які вживають більше 7 напоїв на тиждень) 0 - ні, 1 = так.

14. AnyHealthcare – бінарний показник що вказує на наявність будь-якого медичного страхування, передплачені плани, такі як HMO (Health maintenance organization), тощо. 0 = ні 1 = так

15. NoDocbcCost – бінарний показник що вказує на відповідь на питання: “Чи був час за останні 12 місяців, коли вам потрібно було звернутися до лікаря, але ви не змогли через дорогу вартість?”. 0 – відповідь ні, 1 – так.

16. GenHlth – показник що вказує на відповідь на питання: “Як би ви оцінили свій загальний стан здоров'я від 1 до 5?”.
    Шкала оцінювання:
    1 - відмінний,
    2 - дуже добрий,
    3 - добрий,

4 - задовільний,

5 - поганий

17. MentHlth – показник відповіді на питання: "Задумайтесь про своє психічне здоров'я, яке включає стрес, депресію та емоційні проблеми. Скільки днів протягом останніх 30 днів ваше психічне здоров'я було не на висоті?"
    Шкала значень: від 1 до 30 днів.

18. PhysHlth – показник відповіді на питання: "Зараз думаючи про ваше фізичне здоров'я, яке включає фізичні захворювання та травми, скільки днів протягом останніх 30 днів ваше фізичне здоров'я було не на висоті?".
    Шкала значень: від 1 до 30 днів.

19. DiffWalk – бінарний показник що вказує на відповідь на питання: "Чи маєте ви серйозні труднощі з ходінням або підйомом по сходах?". 0 – відповідь ні, 1 – так.

20. Sex – бінарний показник гендеру пацієнта, 0 – жінка, 1 – чоловік.

21. Age – значення віку пацієнта розподілене на категорії від 1 до 13 включно на таких умовах:

    1 - Вік 18 до 24: Респонденти з віком від 18 до 24 років (18 <= ВІК <= 24)

    2 - Вік 25 до 29: Респонденти з віком від 25 до 29 років (25 <= ВІК <= 29)

3 - Вік 30 до 34: Респонденти з віком від 30 до 34 років (30 <= ВІК <= 34)

4 - Вік 35 до 39: Респонденти з віком від 35 до 39 років (35 <= ВІК <= 39)

5 - Вік 40 до 44: Респонденти з віком від 40 до 44 років (40 <= ВІК <= 44)

6 - Вік 45 до 49: Респонденти з віком від 45 до 49 років (45 <= ВІК <= 49)

7 - Вік 50 до 54: Респонденти з віком від 50 до 54 років (50 <= ВІК <= 54)

8 - Вік 55 до 59: Респонденти з віком від 55 до 59 років (55 <= ВІК <= 59)

9 - Вік 60 до 64: Респонденти з віком від 60 до 64 років (60 <= ВІК <= 64)

10 - Вік 65 до 69: Респонденти з віком від 65 до 69 років (65 <= ВІК <= 69)

11 - Вік 70 до 74: Респонденти з віком від 70 до 74 років (70 <= ВІК <= 74)

12 - Вік 75 до 79: Респонденти з віком від 75 до 79 років (75 <= ВІК <= 79)

13 - Вік 80 років або старше: Респонденти з віком від 75 до 79 років (80 <= ВІК<= 99)

22. Education – показник що є оцінкою рівня освіти від 1 до 6 (цілі числа), що був визначений таким чином:

1 - Ніколи не відвідували школу або тільки дитячий садок

2 - 1-8 класи (початкова школа)

3 - 9-11 класи (неповна середня школа)

4 - 12 клас або GED (середня школа)

5 - Коледж від 1 до 3 років (неповна вища освіта або технічна школа)

6 - Коледж 4 роки або більше (вища освіта)

23. Income – показник що є оцінкою рівня доходу за рік від 1 до 8 (цілі числа), що був визначений таким чином:

1-2 - Менше $15,000

3-4 - Від $15,000 до менше $25,000

5 - Від $25,000 до $35,000

6 - Від $35,000 до $50,000

7-8 - $50,000 або більше

Одні фактори з наведених – фактичні та медичні показники, такі як: HighBP, Stroke, HeartDiseaseorAttack, BMI, Sex, Age.

Решта факторів – базуються на відповідях на питання про спосіб життя людини.

З показників можна будувати теорії й запитання, наведемо кілька прикладів.

Який показник найбільше впливає на результат прогнозування?

Ким є наші респонденти?

Чи є прямий зв'язок: Дохід → є медичне страхування → є можливість медичного обстеження → менша вірогідність набути прediабет чи діабет?

Чи є статистична різниця по хворобам між тими хто раз на день вживає овочі та тими хто фрукти?

Яка стать чи категорія віку більш схильна до діабету?

Який фактор є більш вирішальним: ментальне чи фізичне здоров’я? Чи є кореляція або пряме відношення цих факторів?

Що більше впливає на розвиток діабету: алкоголь чи куріння?

Наскільки суттєвими факторами є Індекс маси тіла та рівень холестерину?

Розділ 2 Аналітика й перспективи нейроних мереж в медицині

Розділ 2.1 Перспективи нейроних мереж в медицині

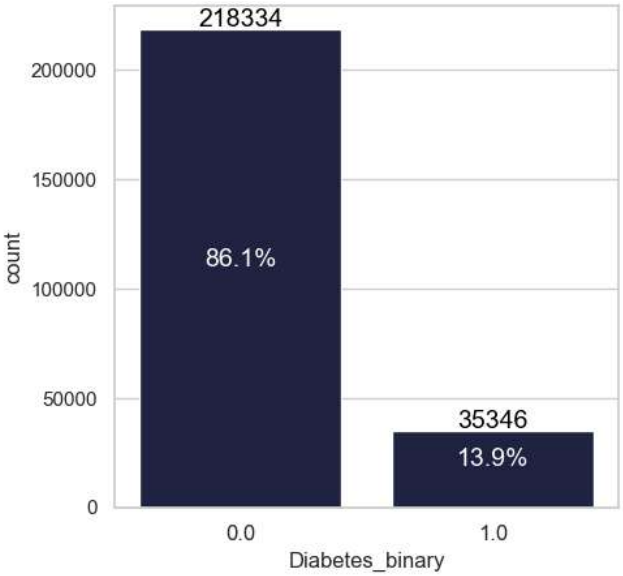
Маючи набагато ширший діапазон медичних показників, історію хвороб кожного із пацієнтів, історію хвороб їх сімей, а у випадку наявності симптомів ще й дані по ним та поточного стану пацієнта – в теорії можна було б створити справді точну програму для визначення та прогнозування хвороб. Саме такий набір даних міг би наблизити до інструменту майбутнього що міг надавати рекомендації, вірогідність виникнення певного захворювання, визначати термін наступного медичного обстеження та навіть прописувати направлення, ліки або вітаміни що потрібні пацієнту на теперішній момент часу.

Це може стати можливим якщо правильним чином зібрати величезну базу даних, але це вкрай важка робота через людське бажання не надавати дозвіл на доступ до чуттєвих показників що в більшості випадків повністю описують їх спосіб життя та їх близьких.

Одна з задач та цілей даної роботи – прописати загальний підхід обробки даних.

Треба розуміти що інноваційних інсайдів про взаємозв’язок рівня певних гормонів чи стан тих чи інших органів тіла людини до виникнення предіабету чи діабету – не буде через брак такої інформації у публічному доступі. Це дуже гарна загальна ідея для співпраці багатьох лікарень по всьому світу.

Розділ 2.2 Аналітика і статистика датасету



Перше, з чого треба почати – подивитись який у нас розподіл по основній метриці даних: скільки з пацієнтів хворі на діабет, а скільки ні.

З рисунку 1 бачимо що в датасеті 13,9% (35346) – пацієнтів мають прediабет чи діабет, а решта – ні.

Вже можемо зрозуміти що треба буде знайти шляхи покращення співмірності для побудови кращої моделі для визначення діагнозу.

Також перевіримо дані на їх тип запису й на незаповнені значення. Результат – рисунок 2.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
```

В кожній з наших 22 колонок значень бачимо кількість Non-Null співпадає з розміром нашого датасету і кожне значення має тип даних float64 – випливає висновок що дані ‘чисті’ та підготовлені.

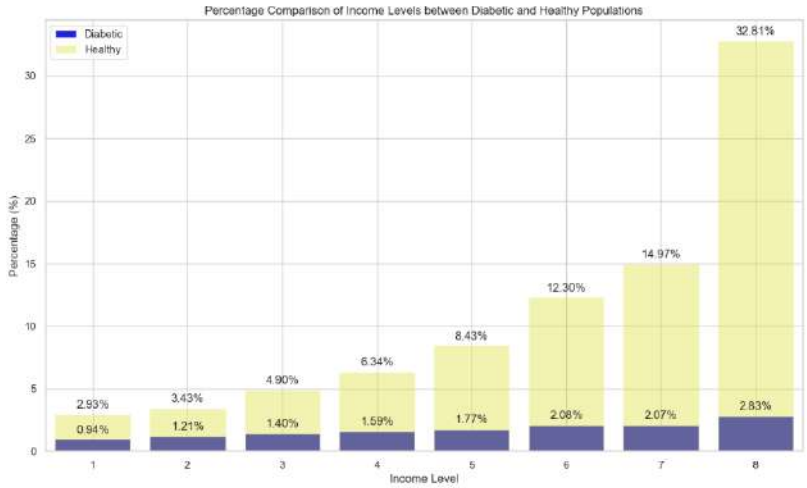
Під час аналітики по людям у будь якому випадку варто притримуватись послідовності у дослідженні, тобто варто вже з самого початку знати ким є пацієнти аби знати наскільки вибірка може співпадати чи ні з реальністю. З офіційних посилань звідки дані були

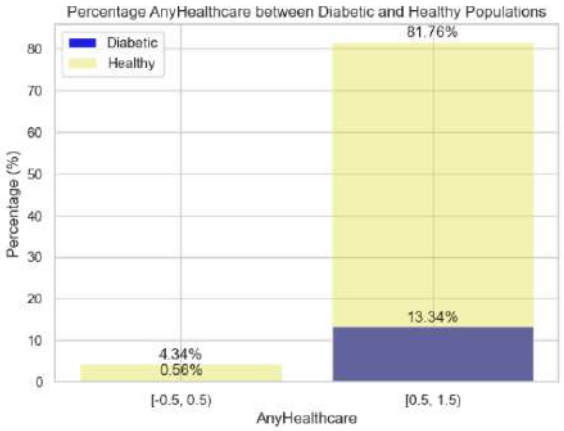
узяті – дані були збирані у Сполучених Штатах загалом в кожному штаті, окрузі Колумбія (DC) і Пуерто-Ріко.

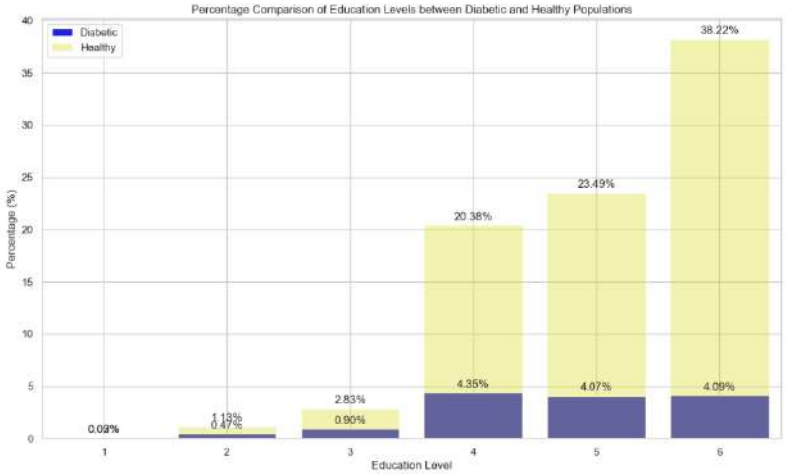
З джерела нам стає відомо що респонденти – американці, надалі розберемо статистичний розподіл вибірок «з діабетом» та «без діабету».

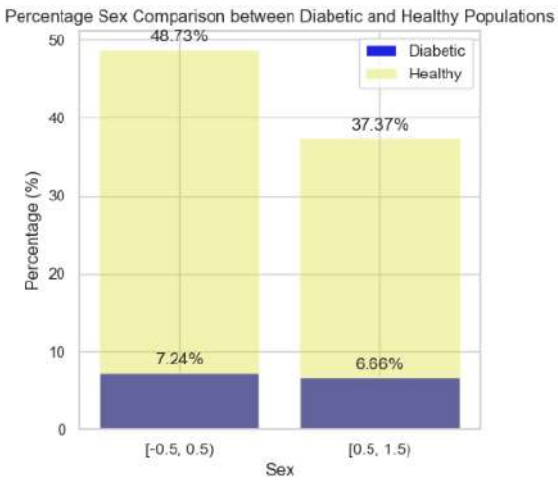
Один з найкращих способів показати одну й ту саму метрику в датасеті розділену за 1 категорією в котрій немає забагато значень – це гістограма.

На рисунку 3 бачимо що більша частина респондентів мають гарну заробітну плату – тобто за логікою у більшості має бути медичне страхування й скоріш за все гарна освіта. На рисунку 4 одразу бачимо підтвердження цього припущення – 95% в сумі мають медстрахування. Рисунок 5: більшість мають освіту 4 і вище – тобто майже всі мають повну середню освіту. Рисунок 6: графік показує що в нашій вибірці більше жінок ніж чоловіків.







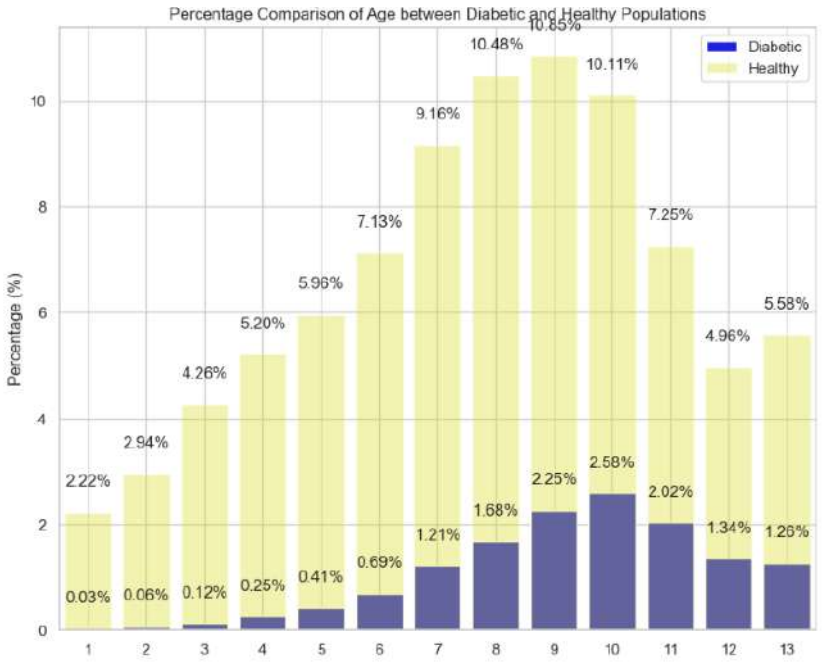


“Температурний” графік показує відсотковий розподіл людей за освітою та доходом на рисунку 7. Розподіл
показує собою логіку чим вища

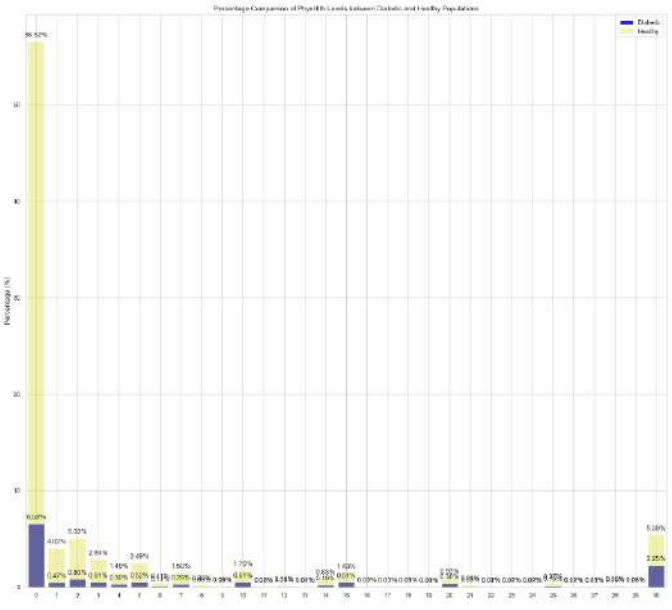
освіта – тим вищий прибуток.

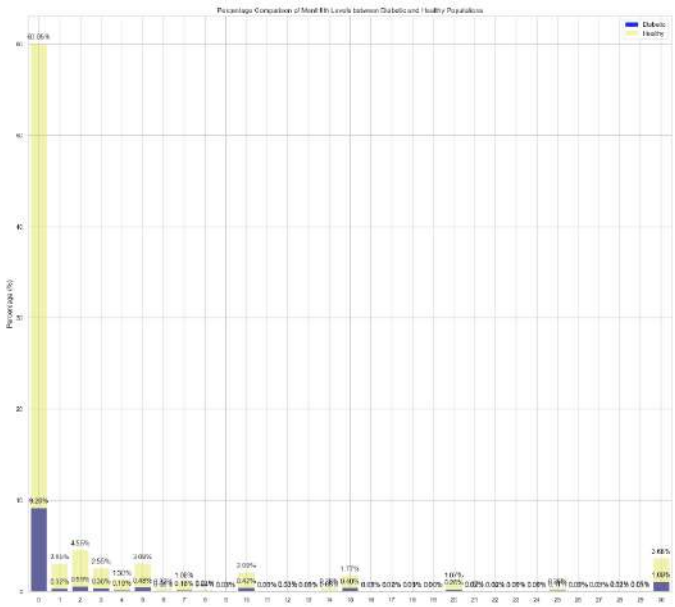
Рисунок 8: Розподіл віку показує нам що велика частина вибірки (50%+-) це люди віком від 50 до 70 років – є старшим поколінням. Респондентів від 18 до 44 років – лише 20% Також з графіка видно що вибірка хворих на діабет є трохи ще старшим ніж вибірка здорових.

З цієї статистики ми маємо загальний портрет респондентів і надалі будемо розуміти в контексті якої вибірки є наші записи даних й робити з цього більш конструктивні висновки.



З розбору кожної з метрик пам’ятаємо що серед них більшість – це бінарні, але є ще інші з більшим розподілом. Варто перевірити якість нефактичних метрик, такі як MentHlth – бо люди зазвичай схильні давати досить круглі числа у випадках усного питання з великим діапазоном відповідей.



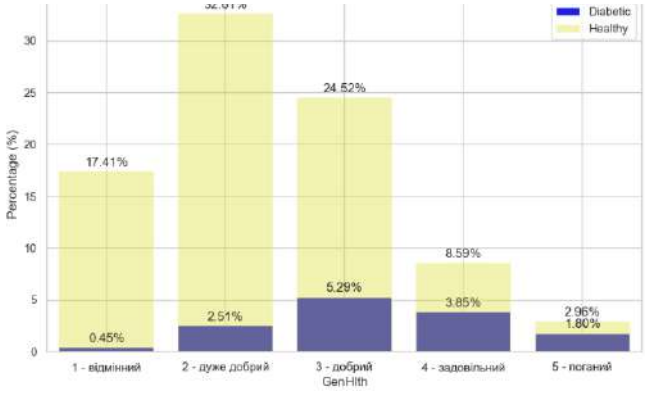


З рисунків 9 і 10 (розподіл оцінки фізичного і ментального стану самими респондентами) стає помітним людський фактор і манеру давати відповідь на питання з великим діапазоном. Бачимо дуже схожі картини – більшість надали 0 днів, що означає що не було жодних поганих днів за самопочуттям фізичним і ментальним за останні 30 днів. Далі звертаємо увагу на значення 5, 10, 15, 20, 25, 30 – вони мають більший відсоток ніж близькі до них значення, тому що люди при наданні відповіді навряд чи справді точно рахуватимуть кількість певних днів якщо це більше тижня, тому люди надають приблизне значення обираючи круглі, більш приємні для сприйняття значення. Цей фактор також ще помітний у поділці у 7 днів – скоріш за все респонденти надавали цю відповідь через факт що в

одному тижні рівно 7 днів або ж просто підсвідомо вважали що 7 – щасливе число.

Ці показники надають інформацію про загальну картину самопочуття респондентів, але ставити їх в ряд більш серйозних метрик – не варто.

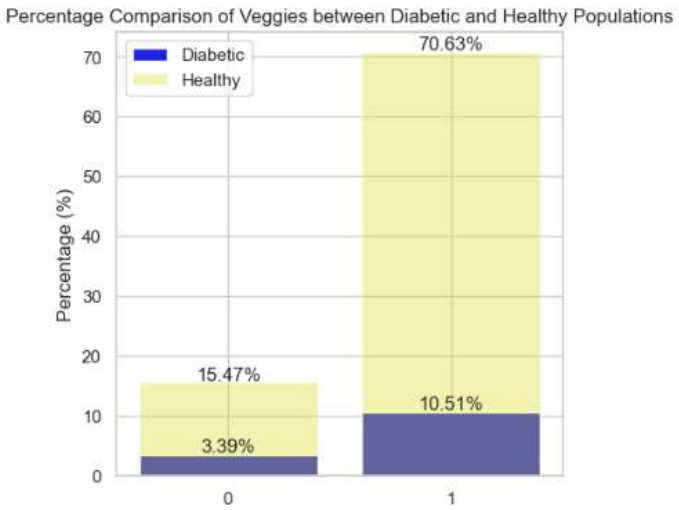
На рисунку 11 бачимо розподіл загального самопочуття респодентів.



Можна стверджувати що у вибірці у здорових від діабету людей оцінюють своє здоров`я більш в гарному стані в порівнянні з хворими на діабет (або предіабет)

Рисунок 12 – чи споживають респонденти кожного дня овочі. Рисунок 13 - чи споживають респонденти кожного дня фрукти.

Більше 80% вибірки споживають овочі кожного дня, а фруктів приблизно 63%.



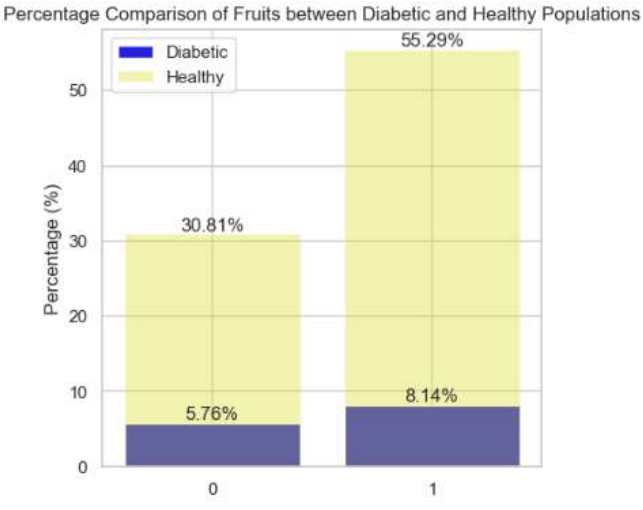
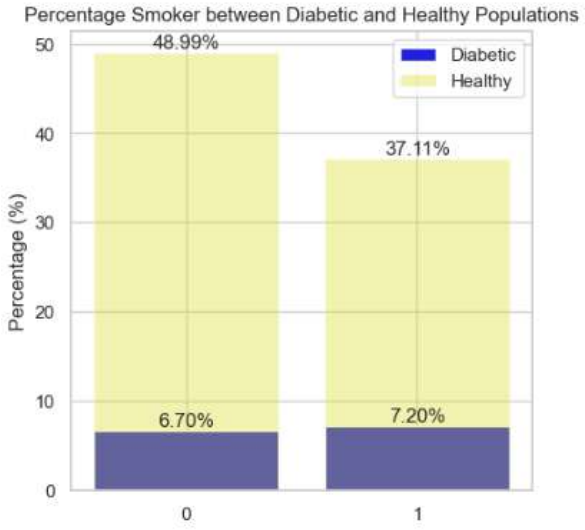
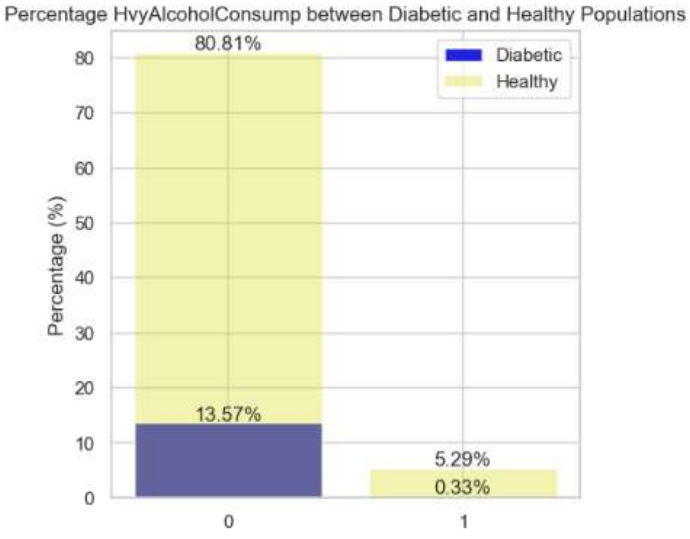


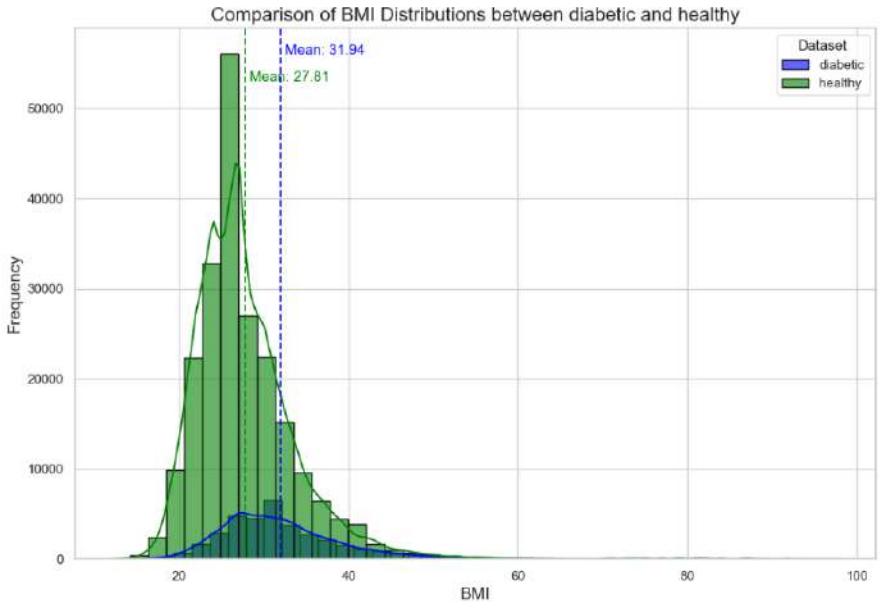
Рисунок 14 – розподіл на курців і не курців. До 45% з вибірки – курці, бачимо що серед хворих на діабет (або прediабет) – більше курців ніж не курців.

Рисунок 15 – розподіл на фактор великого споживання алкоголю. Менше 6% з резпондентів вживають алкоголь в надмірній кількості, тобто вибірка містить вкрай мало людей з проблемами на алкоголь.

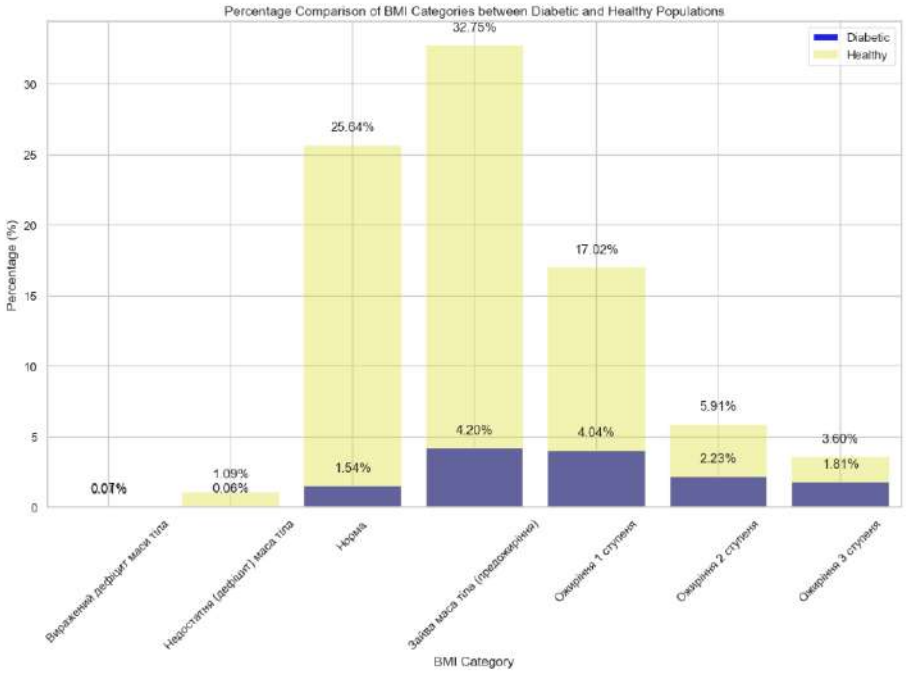




Перейдімо до медично-фактичних метрик: на рисунку 16 зображений графік з розподілом значень BMI респондентів. Середня вага на квадрат зросту респондентів без діабету – 27.81, а з – 31,94. З норм медичного розподілу виходить що середньостатистичний респондент, що не хворий на діабет, класифікується як людина з зайвою вагою, а середнє значення хворої на діабет чи прediабет вже класифікується як ожиріння першого ступеня.



Зробимо кастомний розподіл значень BMI на категорії – рисунок 17.



Стає більш чітко видно що частка респондентів з діабетом більша в кожній категорії чим більше значення BMI – вже можна зробити прогноз що значення BMI на визначення діабету чи прediабету в моделі буде суттєвим фактором.

Розділ 3 Практична реалізація

Розділ 3.1 Перетворення даних для моделі

Використаємо бібліотеку imbalanced-learn:

nm = NearMiss(version = 1 , n_neighbors = 10)

Однією з найкращих практик є використання алгоритму NearMiss що діє шляхом ресемплінгу, тобто робить повторну вибірку даних.

version=1 означає використання версії номер 1 алгоритму, а n_neighbors=10 вказує на кількість сусідів, які будуть використовуватися для визначення які спостереження з більшого класу треба видалити.

Взаємна інформація

Зробимо вибір метрик, які мають найбільшу взаємну інформацію з мітками класу в датасеті:

feature_scores = mutual_info_classif(df_features, df_target, random_state=0)

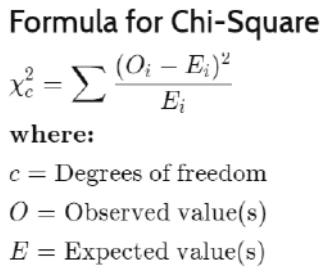
Виклик функції mutual_info_classif, яка обчислює взаємну інформацію між кожною ознакою в df_features та мітками класу в df_target. Параметр random_state=0 дозволяє повторити наш результат.

Отримуємо такий список:

['GenHlth', 'PhysHlth', 'Income', 'DiffWalk', 'MentHlth', 'BMI', 'Education', 'PhysActivity', 'HighBP', 'Veggies']

Далі проженемо код, який використовує метод вибору найкращих ознак SelectKBest із бібліотеки scikit-learn, використовуючи chi-squared статистику як критерій для вибору метрик, які найбільше корелюють з класовими мітками.

Статистика хі-квадрат



Статистика хі-квадрат (χ2) — це міра різниці між спостережуваними та очікуваними частотами результатів набору змінних.

χ2 використовується для аналізу різниць у категоріальних змінних, особливо для тих, що мають номінальне значення.

χ2 залежить від розміру різниці між фактичними та спостережуваними значеннями та розміру вибірки.

χ2 буде використаний для перевірки, чи пов'язані змінні між собою або вони є незалежними одна від одної.

Рисунок 18 – формула для обрахунку χ2.

Втілення через код:

skb = SelectKBest(score_func=chi2, k=threshold)

sel_skb = skb.fit(df_features, df_target)

score_func=chi2 – обирання хі-квадрату для оцінки значень метрик.

k=threshold означає, що буде вибрано 10 найкращих ознак за рейтингом т я threshold=10 в нашому випадку.

Отримуємо: ['HighBP', 'BMI', 'Stroke', 'HeartDiseaseorAttack', 'NoDocbcCost', 'GenHlth', 'MentHlth', 'PhysHlth', 'DiffWalk', 'Income']

Пірсон

Застосуємо кореляцію Пірсона для визначення сили лінійної залежності між кожною метрикою в базі даних та мітками класу, тобто Diabetes_binary.

Для цього використаємо бібліотеку scipy.stats:

p,_ = pearsonr(df_features[i], df_target)

Ця функція повертає кореляцію і p-значення для подальшого формування списку метрик.

Отримуємо:

['GenHlth', 'Income', 'DiffWalk', 'PhysHlth', 'Education', 'PhysActivity', 'BMI', 'MentHlth', 'HighBP', 'HeartDiseaseorAttack']

Сумарно

Зробимо список найкращих, тобто найвпливовіших метрик що є в усіх з вибірок:

['GenHlth', 'PhysHlth', 'Income', 'DiffWalk', 'MentHlth', 'BMI', 'HighBP']

Отримали список із 7-ми найбільш значущих метрик.

Розділ 3.2 Побудова моделей, їх оцінка, вибір найточнішої

Моделі включають логістичну регресію, випадковий ліс, дерево рішень, k-найближчих сусідів та метод опорних векторів.

```
models =[
    ['LR', LogisticRegression(), data],
    ['DT', DecisionTreeClassifier(max_depth=5), data],
    ['RF', RandomForestClassifier(max_depth=5,n_estimators=1000, class_weight='balanced'), data],
    ['KNN', KNeighborsClassifier(n_neighbors= 6), data],
    ['SVM', SVC(probability=True), data]
]
```

Логістична регресія — це статистична модель, яка використовується для прогнозування імовірності події і є статистичним регресійним методом, що застосовують у випадку, коли залежна змінна є бінарною (як і в нашому випадку).

Дерево рішень — алгоритм навчання, який розділяє дані на більш специфічні сегменти на основі значень конкретних метрик, представлених у вигляді графу-дерева.

Випадковий ліс (Random Forest) — метод машинного навчання, який усереднює багато дерев рішень для забезпечення більш стабільного і точного прогнозування шляхом зменшення варіативності.

k-найближчих сусідів (k-NN) — це алгоритм що сегментує об’єкти на основі найближчих метрик у наборі даних, а загальна схема зважування

полягає в тому, щоб надати кожному сусіду певну вагу 1/d, де d — відстань до сусіда.

Метод опорних векторів (SVM) — це алгоритм класифікації, який вираховує гіперплощину в багатовимірному просторі, яка собою і сегментує, тобто класифікує наші метрики шляхом визначення по яку сторону від неї попала та чи інша метрика.

Для оцінки якості моделі будемо використовувати таких як AUC, точність, F1-міра, відгук, та точність прогнозування.

```
roc_model_ls = []
accuracy_model_ls = []
recall_model_ls = []
precision_model_ls = []
f1_model_ls = []
```

AUC score — метрика, що несе за собою значення площі під кривою помилок, вона дає оцінку моделі про її спроможність розрізняти між класами. Чим буде вище значення AUC, тим краще модель буде розрізняти.

Точність (Accuracy) — відсоток вірно класифікованих випадків серед усіх випадків (Який % правильного прогнозування діабету).

Відгук (Recall) — відсоток вірно позитивних результатів, що були вірно ідентифіковані моделлю з усіх реальних позитивних результатів.

Точність прогнозування (Precision) — відсоток вірно позитивних результатів серед усіх результатів, котрі модель ідентифікувала як позитивні.

F1-міра — гармонічне середнє між Precision і відгуком Recall.

Після того як ми прогнали різні моделі й маємо їх оцінки ефективності по різним метрикам – складемо рейтинг для визначення найкращої моделі:

Рисунок 19.

```
evaluations = evaluations.sort_values(by='AUC Score', ascending=False, ignore_index=True)
evaluations
```

	classifiers	Feature selection method	AUC Score	Accuracy	Precision	Recall	F-measure
0	SVM	All	0.951643	0.900556	0.946844	0.849337	0.895444
1	LR	All	0.943192	0.880234	0.921642	0.831844	0.874444
2	RF	All	0.942746	0.874198	0.938940	0.801185	0.864610
3	SVM	Pearson - 10	0.940696	0.883016	0.944640	0.812328	0.873502
4	RF	Pearson - 10	0.940664	0.872171	0.942999	0.790706	0.860164
5	SVM	Mic - 10	0.939692	0.880375	0.941054	0.811201	0.871316
6	LR	Pearson - 10	0.939572	0.871747	0.913540	0.819630	0.864041
7	RF	Mic - 10	0.939464	0.870143	0.940508	0.789856	0.858624
8	RF	Skb - 10	0.937499	0.866135	0.942822	0.781071	0.854358
9	LR	Mic - 10	0.937163	0.867880	0.912307	0.813563	0.860110
10	LR	Skb - 10	0.935327	0.866890	0.920764	0.804427	0.858673
11	RF	Mic - 7	0.934767	0.864391	0.940749	0.778801	0.852149
12	RF	Best - 7	0.934282	0.864391	0.921458	0.794548	0.853310
13	LR	Mic - 7	0.931167	0.860713	0.912889	0.798628	0.851945
14	LR	Best - 7	0.931128	0.860336	0.909150	0.798442	0.850207
15	RF	Pearson - 7	0.929435	0.856988	0.917742	0.784449	0.845876
16	SVM	Skb - 10	0.928022	0.873963	0.948562	0.792233	0.863378
17	LR	Pearson - 7	0.924381	0.851094	0.896130	0.794439	0.842226
18	RF	Mic - 5	0.923474	0.858072	0.924522	0.779779	0.846004
19	RF	Choose - 7	0.922348	0.847510	0.944888	0.737826	0.828617
20	SVM	Best - 7	0.921539	0.870426	0.944978	0.784669	0.857395

Проганяючи по декілька разів майже завжди показував кращий результат класифікатор Random Forest за методом Пірсона з 10 метриками на розгляд.

Через це ми й створемо модель використовуючи саме це.

Розділ 4 Побудова фінальної моделі

Ділемо наші дані на train і test частини, де test частина – 30%:

X = df_features[high_score_features_mic]

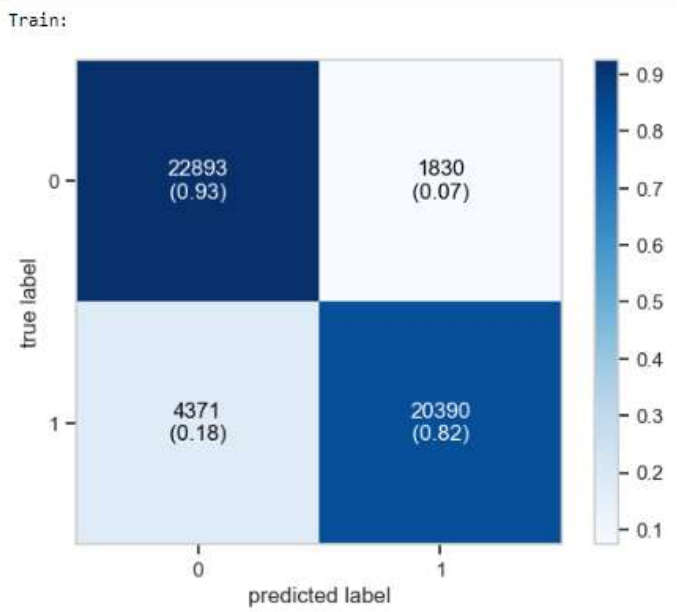
y = df_target

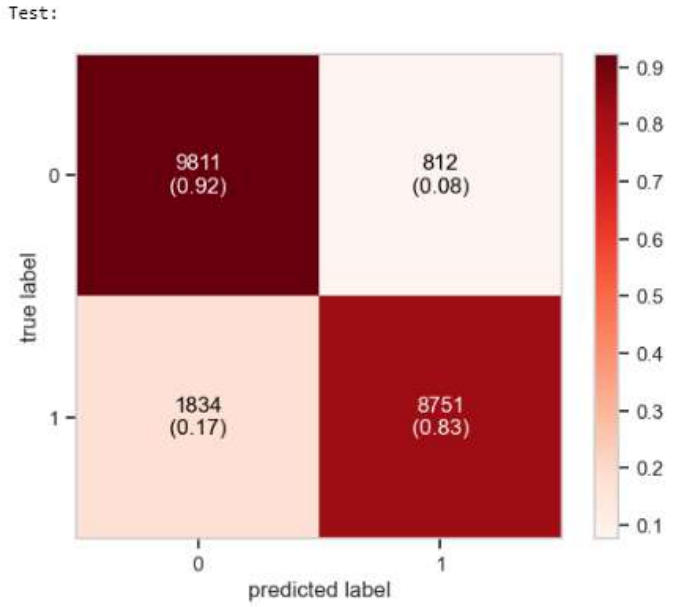
X_train , X_test , y_train , y_test = train_test_split(X, df_y, test_size=0.3)

model.fit(X_train, y_train) #тренуємо модель

Дивимось roc_auc_score моделі: 0.9387 – вийшов чудовий результат.

Візуалізуємо у матричному вигляді наскільки добре модель справляється з train і test, тобто покажемо з якою вірогідністю модель визначить або ні що респондент має діабет чи ні: Рисунок 20.

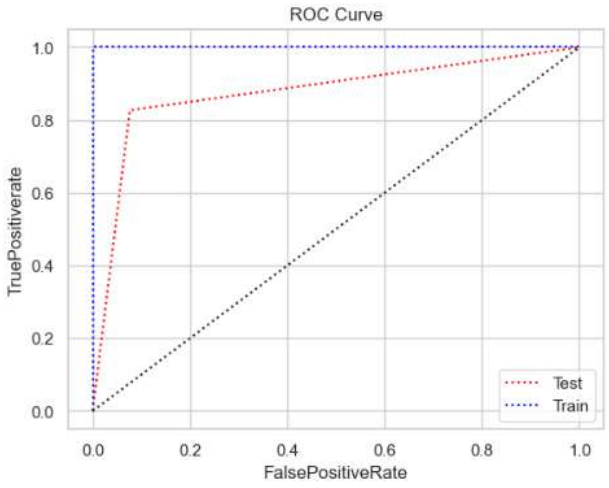




roc_curve -- функція, що використовується для обчислення відношення помилково позитивних (FPR) та істинно позитивних (TPR) результатів.

Побудуємо цю криву що відобразить 2 параметри: Істинний позитивний показник та хибнопозитивний рівень.

Рисунок 21 – наша крива.



Розділ 4.1 Збереження отриманої моделі

Після створення моделі, зберігаємо її:

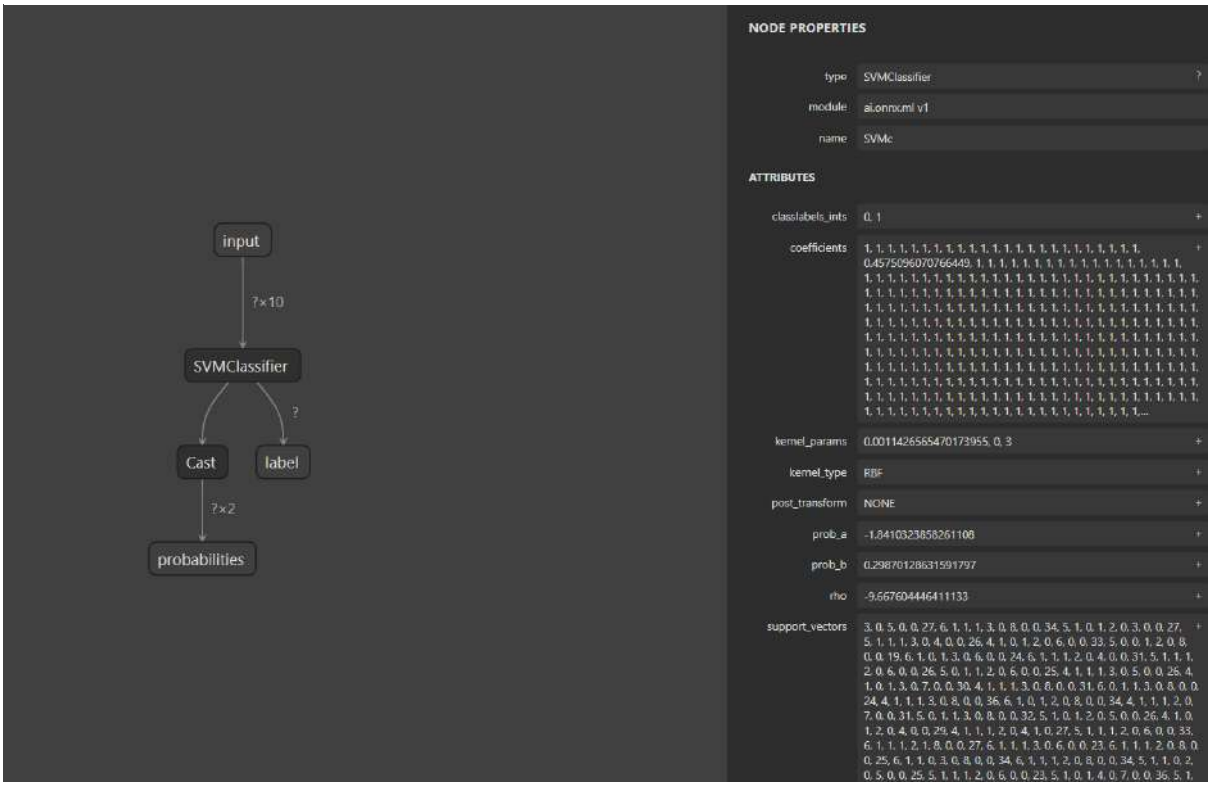
initial_type = [('input', FloatTensorType([None, 10]))]

onx = convert_sklearn(model, initial_types=initial_type, options={'zipmap': False})

with open("ModelForDiabetesPrediction.onnx", "wb") as f:

    f.write(onx.SerializeToString())

Візуалізація нашої моделі через сайт[ii]:



Висновок

У роботі було досліджено послідовність дій, найкращі практики, загальні проблематики й перспективи нейронних мереж для виявлення чи діагностування захворювань.[iii]

Робота несе інформацію про базові практики дослідження та оперування даних такі як:

1. Повний розбір природи даних.
2. Отримання портрету респондентів й взаємозв'язків між даними.
3. Майбутня перспектива нейронних мереж в сфері медицини.
4. Перетворення даних для подальшого використання їх у моделях машинного навчання.
5. Знаходження й оцінювання різних класифікаторів для отримання кращого результату ефективності моделі.
6. Тренування моделі й візуалізація результату.

При побудові графіків, на котрих базуються аналіз й висновки даної роботи, було практично реалізовано код мовою Python у середовищі Jupyther Notebook з використанням бібліотек: Seaborn, Matplotlib, mlxtend.

Також для роботи з даними та для побудови моделі використовувались бібліотеки: numpy, pandas, sklearn, scipy, imblearn, tqdm, skl2onnx.[iv]

Результати дослідження показали, що нейронні мережі є і будуть прикладними для багатьох напрямків людського життя, особливо у сфері медицини через пряму залежність між фактичних наборами даних і діагностуванням захворювань. Важливу роль відіграє належний вибір метрик, їх збір та перенесення у цифровий вигляд для подальшої обробки даних і побудові на їх основі моделей.[v]

Маючи гарно відточені моделі відкривається провідна перспектива створення програми діагностування хвороб та прогнозування їх можливого виникнення.

Зібрати коректно абсолютно усі можливі дані про спосіб життя людини та її особливості – неможливо. Саме тому для розвитку такого продукту будуть потрібні люди, що будуть визначати суттєві для результату моделі метрики та шляхи їх отримання для подальшої імплементації в модель.

Робота на практиці демонструє повний практичний підхід до даних та весь шлях створення нейронної мережі – саме тому структура й послідовність роботи надає собою повноцінну схему, що можна відтворити на інших даних у майбутніх дослідженнях.[vi]

Було описано, на практиці зроблено і оцінено ефективність такі види класифікаторів як логістична регресія, дерево рішень, випадковий ліс, k-найближчих сусідів та метод опорних векторів.[vii]

Було отримано й збережено модель визначення діабету та прediабету у вигляді .onnx файлу.[viii]

Джерела:

[i] Джерело даних:
https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

[ii] Сайт для візуалізації моделі
https://netron.app/

[iii] Стаття:
https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-021-05489-x

[iv] Документація бібліотеки Python:
https://scikit-learn.org/stable/auto_examples/index.html#classification

[v] Сайт про дані:
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

[vi] Сайт на тему метрик що оцінюють якість моделі:
https://arize.com/blog-course/f1-score/

[vii] Стаття на тему бінарних класифікаторів:
https://medium.com/@andrii.gozhulovskyi/choosing-a-model-for-binary-classification-problem-f211f7a4e263

[viii] Посилання на гітхаб практичної частини
https://github.com/AntonSmovzhenko/Diploma.git