

Київський столичний університет імені Бориса Грінченка
Факультет інформаційних технологій та математики
Кафедра комп'ютерних наук

«Допущено до захисту»

Завідувач кафедри комп'ютерних
наук, доцент, канд техн. наук

_____ Ірина МАШКІНА

« ____ » _____ 2025 р.

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня «бакалавр»

Спеціальність 122 «Комп'ютерні науки»

Тема: Розробка веб-додатку для оцінки ризику серцевого нападу

Виконав

Студент ІНб-2-21-4.0д.
Дорошенко О. О.

(підпис)

Науковий керівник

Карпенко А. С.

(підпис)

Київ – 2025

Київський столичний університет імені Бориса Грінченка
Факультет інформаційних технологій та математики
Кафедра комп'ютерних наук

Завідувач кафедри комп'ютерних
наук, доцент, канд техн. наук
_____ Ірина МАШКІНА

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

«Розробка веб-додатку для оцінки ризику серцевого нападу»
Виконавець – студент Дорошенко Олексій Олександрович
групи ІНБ-2-21-4.0д.

1. Вихідні дані: Публічні медичні дані та дані соціальних опитувань щодо теми, наукові роботи та публікації за напрямом дослідження, зокрема з інформатики.
2. Основні завдання: Здійснити огляд публікацій за темою роботи. Обґрунтувати мету, об'єкт, предмет та наукові основи дослідження. Розробити завдання, структуру та зміст бакалаврської роботи. Обрати та обґрунтувати методи і засоби дослідження. Підготувати матеріали до розділів роботи відповідно до індивідуального завдання. Проаналізувати науково-методичні та джерела з теми дослідження. Визначити складові, скласти висновки та оформити бакалаврську роботу відповідно до затверджених на кафедрі вимог. Підготувати презентацію за темою роботи.
3. Пояснювальна записка: Обсяг – 48 стор. формату А4 комп'ютерного набору з дотриманням вимог стандарту і методичних рекомендацій кафедри.
4. Графічні матеріали: презентація.
5. Строк подання роботи на кафедру: «__» _____ 2025р.

Науковий керівник

Кандидат педагогічних наук

Карпенко Анастасія Сергіївна

Дата:

Виконавець:

Дорошенко Олексій Олександрович

Дата:

Календарний план

№	Назва етапу дипломної роботи	Строк виконання етапу роботи	Примітка
1	Формування теми дипломної роботи бакалавра	до 18 жовтня	Виконано
2	Ознайомлення з методичними рекомендаціями до дипломної роботи	до 15 листопада	Виконано
3	Складання змісту та календарного плану роботи	до 20 листопада	Виконано
4	Підбір літератури, складання завдання	до 5 грудня	Виконано
5	Написання реферату та вступу	до 9 грудня	Виконано
6	Написання першого розділу	до 8 лютого	Виконано
7	Написання другого розділу	до 25 березня	Виконано
8	Написання третього розділу	до 2 квітня	Виконано
9	Написання висновків	до 8 квітня	Виконано
10	Виправлення зауважень	до 12 травня	Виконано
11	Створення презентації	до 20 травня	Виконано
12	Захист матеріалів дипломної роботи на засіданні кафедри	за 1 місяць до захисту	Виконано
13	Офіційний захист матеріалів дипломної роботи на засіданні екзаменаційної комісії	згідно графіку захистів	

Студент _____

Керівник роботи _____

ЗМІСТ	
ВСТУП	1
РОЗДІЛ 1. ПРОБЛЕМА КАРДІО-ВАСКУЛЯРНИХ ХВОРОБ: МАСШТАБИ, НАСЛІДКИ ТА НЕДООЦІНКА	3
1.1. ПОШИРЕНІСТЬ ТА АКТУАЛЬНІСТЬ ПРОБЛЕМИ СЕРЦЕВО-СУДИННИХ ЗАХВОРЮВАНЬ	3
1.2. НАСЛІДКИ СЕРЦЕВИХ НАПАДІВ	7
1.3. НЕДООЦІНКА ПРОБЛЕМИ В СУСПІЛЬСТВІ	11
1.4. ФОРМУВАННЯ ВЛАСНИХ ПОГЛЯДІВ НА ДОСЛІДЖУВАНУ ПРОБЛЕМУ	14
РОЗДІЛ 2. ДОСЛІДЖЕННЯ ДАНИХ ТА ПОБУДОВА МОДЕЛІ МАШИННОГО НАВЧАННЯ	17
2.1. ОГЛЯД ДЖЕРЕЛА ТА ЕКСПЛОРАТИВНИЙ АНАЛІЗ	17
2.2. ТРАНСФОРМАЦІЯ, ЕНКОДИНГ, СТВОРЕННЯ НОВИХ ОЗНАК ТА ІМПУТАЦІЯ ДАНИХ	31
2.3. ПОБУДОВА МОДЕЛЕЙ ТА ВИБІР МЕТРИК ТА МЕТОДІВ ВАЛІДАЦІЇ	35
РОЗДІЛ 3. СТВОРЕННЯ, ТЕСТУВАННЯ ТА ДЕПЛОЙ ВЕБ-ЗАСТОСУНКУ	43
3.1. ОБҐРУНТУВАННЯ ВИБОРУ ТЕХНОЛОГІЙ ТА АРХІТЕКТУРА ДОДАТКУ	43
3.2. Пояснення логіки функціонування додатку	46
3.3. ТЕСТУВАННЯ ТА РОЗГОРТАННЯ ДОДАТКУ	49
ВИСНОВКИ	52
СПИСОК ДЖЕРЕЛ	54

ДОДАТКИ	59
ДОДАТОК А	59
ДОДАТОК Б	61
ДОДАТОК В	67
ДОДАТОК Г	69

ВСТУП

Медицина була та є однією з провідних галузей людської діяльності, завдяки недосконалості людського організму. Від лікування народною медициною ми прийшли до тисяч лікувальних препаратів, що мають науково обґрунтовану ефективність та лікують ті хвороби, що ще кілька сторічч тому вважались невиліковними. Людство зробило величезний прорив у медицині від часів появи перших поселень і з розвитком інформаційних технологій він продовжує набирати обертів. Тим не менш, все ще існує багато хвороб, які ми вилікувати не в силах, незважаючи на всі наші здобутки. Туберкульоз, рак, СНІД, аутоімунні захворювання, альцгеймер та інші продовжують щороку вбивати мільйони людей. І таким чином кожна людина знає про їх небезпеку, просвітництво про запобігання даним хворобам активно проводиться по всьому світу. Але, якими б ці хвороби не були страшними, не вони є головним смертей у всьому світі. Насправді, на них приходить разом приходить лише четверта частка від загальної маси усіх смертей, спричинених хворобами. Більше смертей викликають інші хвороби, про які всі, насправді, всі добре знають. Хвороби, які є виліковними та добре запобігаються за допомогою профілактики. І все ж таки, можливо, найбільш недооцінені хвороби зі всіх. Саме хвороби серця складають третину від усіх ненасильних смертей у світі.

Об'єктом дослідження є моніторинг та оцінка ризиків у людей з потенційною загрозою серцево-судинних захворювань. Предметом дослідження є цифрові технології та методи розробки веб-додатків, спрямованих на моніторинг та оцінку ризику серцевого нападу на основі поведінкових даних користувача. Метою дослідження є розробка функціонального веб-додатку, що дозволяє оцінювати ризик серцевих хвороб за допомогою методів машинного навчання, забезпечуючи зручний інтерфейс для користувачів та швидкий доступ до результатів оцінки. Для досягнення мети необхідно виконати наступні завдання:

1. Аналіз трендів, проблем та можливостей в області лікування та профілактики кардіо-васкулярних хвороб.
2. Дослідити доступні публічні дані, на яких буде будуватись предиктивна модель.
3. Побудова та тестування моделі оцінки ризику хвороб серця.
4. Створення веб-додатку у форматі опитувальника в якості інтерфейсу між користувачем та моделлю.
5. Тестування та розгортання застосунку у мережі інтернет

РОЗДІЛ 1. ПРОБЛЕМА КАРДІО-ВАСКУЛЯРНИХ ХВОРОБ: МАСШТАБИ, НАСЛІДКИ ТА НЕДООЦІНКА

1.1. ПОШИРЕНІСТЬ ТА АКТУАЛЬНІСТЬ ПРОБЛЕМИ СЕРЦЕВО-СУДИННИХ ЗАХВОРЮВАНЬ

Серцево-судинні захворювання (ССЗ) залишаються пріоритетною медико-соціальною проблемою в усьому світі. За даними Всесвітньої організації охорони здоров'я, ССЗ вже багато років є головною причиною смерті населення, як видно на рисунку 1.1. Зокрема, «у 2019 році від ССЗ померло близько 17,9 млн людей (32% всіх смертей у світі), і 85% цих випадків були викликані серцевими нападами або інсультами»[1].

Таким чином, інфаркт міокарда (серцевий напад) робить найбільший внесок у глобальну смертність від неінфекційних хвороб. Для порівняння, ще «у 2000 році кількість смертей від ССЗ становила приблизно 14 млн, тобто за два десятиліття цей показник зріс до майже 18 млн»[2]

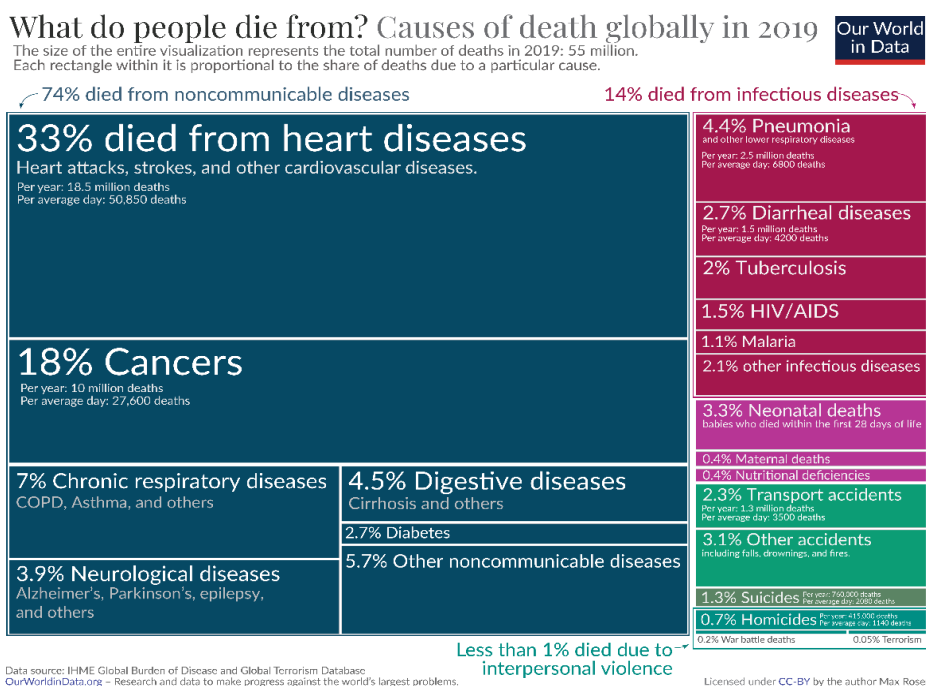


Рис. 1.1 - Від чого люди помирають? Причини смерті по всьому світі 2019

Зростання частки ССЗ як причини смерті особливо помітне з переходом суспільств до індустріальної та постіндустріальної стадій розвитку, коли відбулося зниження смертності від інфекційних хвороб та одночасне зростання поширеності хронічних недуг (включно з хворобами серця)

Наприклад, «у період 2000–2019 років кількість смертей від ішемічної хвороби серця зросла більш ніж на 2 млн і сягнула 9 млн на рік».[3]

Натомість смертність від ряду інфекцій (як-от туберкульоз, ВІЛ/СНІД) за той самий час суттєво знизилася

Ці дані підкреслюють актуальність проблеми ССЗ у сучасному світі та необхідність зміщення акцентів медицини з лікування інфекційних хвороб на профілактику й контроль неінфекційних.

Вразливість до серцевих нападів нерівномірно розподілена серед різних груп населення. Ризик різко зростає з віком: більшість інфарктів трапляються у людей старше 45–55 років, особливо після 65 років. [4]

Хоча традиційно вважалося, що чоловіки більш схильні до серцевих нападів, серцево-судинні захворювання є провідною причиною смерті і серед жінок. [5]

Серцеві захворювання займають перше місце за смертністю і серед чоловіків, і серед жінок у всьому світі [5][6]

Примітно, що у жінок інфаркт міокарда часто виникає в середньому на 9–10 років пізніше, ніж у чоловіків, але наслідки можуть бути тяжчими. [7]

До груп підвищеного ризику також належать люди з обтяженою спадковістю, особи з низьким соціально-економічним статусом та ті, хто піддається впливу кількох поведінкових факторів ризику одночасно.

Сучасний спосіб життя істотно сприяє епідемії серцево-судинних хвороб. Основні поведінкові фактори ризику розвитку інфаркту міокарда включають нездорове харчування, недостатню фізичну активність, вживання тютюну та зловживання алкоголем [1].

Раціони багатьох людей все більше складаються з промислово оброблених продуктів, багатих на сіль, цукор і транс-жири. За результатами великого дослідження, підвищене споживання ультра-обробленої їжі асоціюється зі зростанням ризику серцево-судинних захворювань і смертності.

Зокрема, кожна додаткова порція таких продуктів на день пов'язана зі збільшенням ризику важких серцевих подій на 7% та смертності від ССЗ на 9%. [8]

Одночасно у світі спостерігається пандемія малорухомості: за оцінками ВООЗ, понад 80% підлітків і близько 27% дорослих не виконують мінімальні рекомендації з фізичної активності. [9]

Такий сидячий спосіб життя разом із висококалорійним харчуванням призводить до зростання ожиріння і діабету – важливих факторів ризику інфаркту. Сучасне урбанізоване середовище також приносить хронічний стрес, забруднення повітря та інші фактори, що додатково сприяють атеросклерозу. Внаслідок цього переважна частка випадків інфарктів міокарда є наслідком модифікованих чинників. Глобальне дослідження INTERHEART показало, що дев'ять піддаваних впливу факторів (куріння, нездоровий раціон, гіподинамія, ожиріння, підвищений артеріальний тиск, діабет, психосоціальний стрес, надмірне вживання алкоголю та дисліпідемія) сумарно пояснюють понад 90% ризику першого інфаркту міокарда. [7]

Інакше кажучи, до 90% серцевих нападів можна було б попередити усуненням факторів ризику, пов'язаних із способом життя.

Попри очевидну роль профілактики, історично системи охорони здоров'я приділяли більше уваги лікуванню вже наявних хвороб, ніж запобіганню їх виникненню. Багато дослідників відзначають, що сучасна медицина часто націлена на «ремонт» здоров'я, а не на його збереження. Профілактика йде за залишковим принципом, а лікування вже наявної

хвороби є головним фокусом сучасної охорони здоров'я, що веде до перебільшених та неконтрольованих витрат. Люди покладаються на ліки та лікарів, втрачають здатність самотійно слідкувати за здоров'ям та вимушені ще більше покладатись на безпосереднє лікування хвороб. [11]

Це підтверджують і статистичні дані: приміром, у США на профілактичні послуги витрачається лише близько 3–5% загального обсягу видатків на охорону здоров'я [11], тоді як переважна частка ресурсів йде на лікування уже наявних хронічних захворювань. У результаті профілактичний потенціал реалізується не повністю. ВООЗ наголошує, що сучасні системи охорони здоров'я мають змістити акцент на первинну ланку та профілактичні заходи, оскільки раннє виявлення і контроль факторів ризику можуть запобігти передчасній смерті від ССЗ. [1]

Іншими словами, вкладення в профілактику серцево-судинних хвороб сьогодні – це збережені життя і здоров'я завтра.

1.2. НАСЛІДКИ СЕРЦЕВИХ НАПАДІВ

Інфаркт міокарда, інсульт та інші хвороби серця мають важкі наслідки як для фізичного, так і для психічного здоров'я людини. Гостре ураження серцевого м'яза нерідко призводить до розвитку ускладнень: гострої серцевої недостатності, аритмій, кардіогенного шоку. Навіть при успішному лікуванні в гострий період приблизно у 20% хворих розвивається хронічна серцева недостатність, що обмежує їхню фізичну працездатність та якість життя. [12]

Багато пацієнтів після інфаркту міокарда скаржаться на стійку задишку, швидку втомлюваність, ангінозні болі при навантаженнях та зниження толерантності до фізичної активності. За даними досліджень, відсоток виживання після перенесеного інфаркту залишається невисоким –

п'ятирічна смертність серед пацієнтів, що перенесли гострий коронарний синдром, становить близько 50% (з урахуванням повторних подій). [13]

Особливо несприятливий прогноз у хворих, що перенесли повторний серцевий напад: якщо повторний інфаркт стається впродовж кількох місяців після першого, ймовірність смерті в наступні 5 років сягає близько 50%. [14]

Таким чином, перенесений серцевий напад накладає на пацієнта довічний «тягар» – потребу у прийомі ліків, медичному спостереженні, обмеженнях у способі життя і постійному контролі факторів ризику, аби запобігти рецидиву.

Вплив серцевого нападу на психічне здоров'я також дуже значний. Пацієнти часто переживають депресію, тривожні розлади, посттравматичний стрес після такої події. «Депресія в три рази більше поширена серед пацієнтів після серцевої, порівняно з усією популяцією, і складає від 15% до 20% усіх жертв» [15]

Тобто до п'ятої частини осіб, що перенесли інфаркт, страждають на депресію – це у рази більше, ніж у загальній популяції. Крім того, майже кожен третій хворий після інфаркту відчуває клінічно значиму тривожність або страх повторного нападу

Психологічні наслідки можуть погіршувати відновлення: депресивні та тривожні стани асоціюються з гіршою реабілітацією, нижчою прихильністю до лікування та навіть підвищеним ризиком повторних серцевих нападів.

Але, окрім усього вище перерахованого, серцевий напад – це не тільки медичний, а й часто значний фінансовий удар для людини. Лікування гострого інфаркту потребує дороговартісних втручань (коронарна ангіографія, стентування або шунтування, інтенсивна терапія), а подальша реабілітація – тривалого прийому ліків (антиагреганти, бета-блокатори, статини тощо), спостереження лікарів та часто повторних госпіталізацій. У

країнах з розвинутими страховими системами значну частину витрат бере на себе страхова чи держава, але пацієнт все одно може втрачати доходи через непрацездатність. У разі відсутності страхування гострий коронарний синдром може призвести до катастрофічних витрат для родини. За оцінками, у 2022 році в США хвороби серця та інсульти коштували економіці ~\$320 млрд, з них \$252,2 млрд – прямі медичні витрати, а решта – втрати продуктивності. [5]

Це означає, що кожні 33 секунди тільки американська система охорони здоров'я, яка з поміж усього є однією з найрозвинутіших у світі, втрачає життя однієї людини від серцево-судинної хвороби, і кожен такий випадок тягне значні видатки

Серцево-судинні захворювання обходяться США приблизно в 1 млрд доларів щодня з урахуванням вартості лікування і непрямих збитків

В Європейському Союзі ситуація подібна – сумарні витрати на ССЗ оцінюються у €282 млрд на рік (станом на 2021 рік)

Глобально фінансовий тягар теж колосальний: за прогнозами Всесвітньої федерації серця, до 2030 року сумарні прямі та непрямі витрати, пов'язані з ССЗ, перевищать 1 трильйон доларів США на рік. [16]

Я приводжу дані з США та ЄС, так як ці економічні зони мають досить розвинену систему моніторингу економічних тенденцій та добробуту громадян. Хоча подібні дані для України здебільшого відсутні, але було б логічно припустити, що для неї ситуація в цілому схожа. З меншою вірогідністю вона краща, так як Україна має слабкішу економіку та інституції, які тісно кореллюють з добробутом громадян. [17]

Висока захворюваність і смертність від інфарктів означає, що величезні кошти витрачаються, які могли б бути зекономлені за умови ефективної профілактики. Профілактичні заходи – це інвестиція, що здатна дати суттєвий економічний зиск як для окремих людей, так і для системи охорони здоров'я в цілому. По-перше, попередження одного випадку

інфаркту означає збереження десятків тисяч доларів, які знадобилися б на його лікування, не кажучи вже про збережене життя та працездатність людини. По-друге, на рівні суспільства профілактика масово скорочує витрати: за розрахунками, 100% охоплення населення основними профілактичними втручаннями могло б зменшити витрати на лікування ССЗ приблизно на \$904 млрд щорічно у світі (близько 10%). [18]

Дослідження, проведене CDC, показало, що навіть часткове впровадження профілактичних програм дає результат. Згідно з моделюванням, досягнення 90% виконання рекомендацій з профілактики дозволило б скоротити річні медичні витрати в США на ~\$54 млрд. [19]

Також це дослідження з'ясувало, що реалізація популяційних стратегій (боротьба з тютюнопалінням, поліпшення харчування тощо) може запобігти до 70% нових випадків хронічних хвороб і таким чином суттєво знизити витрати на їх лікування. Враховуючи, що серцеві напади становлять значну частку цих хвороб, економія коштів через їх попередження буде пропорційно значною. На глобальному рівні ВООЗ підрахувала, що відсутність активних дій проти гіподинамії призведе до 500 мільйонів нових випадків захворювань серця, діабету та інших НІЗ до 2030 року, що обійдеться у \$300 млрд додаткових витрат на лікування.

Тож, профілактика – це не лише про здоров'я, а й про економічну доцільність: кожен долар, вкладений сьогодні у боротьбу з факторами ризику, повертається кратно у вигляді зменшення майбутніх витрат.

1.3. НЕДООЦІНКА ПРОБЛЕМИ В СУСПІЛЬСТВІ

Незважаючи на поширеність серцево-судинних захворювань, проблема серцевих нападів часто недооцінюється суспільством. Багато людей мають низький рівень обізнаності щодо факторів ризику і перших ознак інфаркту, а отже – не усвідомлюють повною мірою власну уразливість. Дослідження

показують, що населення недооцінює загрозу ССЗ, особливо серед жінок та молоді. Наприклад, опитування у Сингапурі виявило, що обізнаність жінок про те, що хвороби серця є причиною номер один їхньої смерті, знизилася з 65% у 2009 році до лише 44% у 2019 році. [20]

У дослідженні з Берліна лише 41,35% учасниць правильно оцінили свій ризик розвитку ССЗ, а 48,65% – його недооцінили. [20]

Існує хибне уявлення, ніби інфаркт – це «чоловіча» хвороба або проблема виключно літніх людей, через що молодші жінки та чоловіки середнього віку можуть ігнорувати симптоми або не вважати себе групою ризику. [21]

У реальності ж серцевий напад може статися у будь-якому віці, особливо за наявності факторів ризику, і нерідко вражає зовні здорових на вигляд людей.

Низький рівень знань про симптоми інфаркту міокарда призводить до того, що люди не завжди можуть вчасно розпізнати напад і надати собі або оточуючим належну допомогу. Більшість знає про класичний симптом – біль у грудях, але мало хто обізнаний про інші прояви. За даними опитування, лише 26% респондентів були обізнані, що біль у щелепі, шиї чи спині може бути ознакою серцевого нападу, тоді як про біль у грудях знали 72%. [22]

При цьому тільки 11,5% опитаних змогли правильно назвати всі основні симптоми інфаркту.

У США, де проводяться масштабні інформаційні кампанії, не більше половини дорослого населення здатні перелічити п'ять ключових симптомів серцевого нападу (біль або дискомфорт у грудях, слабкість чи запаморочення, біль у руках чи плечі, біль у шиї/щелепі/спині, ускладнене дихання). [23]

Особливо тривожно, що люди часто не знають, як правильно діяти при появі симптомів інфаркту. Зволікання з викликом невідкладної допомоги

значно підвищує ризик смерті чи тяжких ускладнень, адже «час – це м'яз серця». Проте опитування свідчать, що велика частка населення не усвідомлює необхідності негайно телефонувати до швидкої допомоги. Лише 35,6% респондентів у згаданому дослідженні відповіли, що викликать швидку при підозрі на інфаркт у оточуючого. [22]

Решта або вагалися з відповіддю, або обирали неправильні дії (наприклад, спробують самотійно відвезти хворого до лікарні чи чекатимуть, «поки минеться»).

Низька усвідомленість ризиків відображається і в тому, скільки людей реально займаються профілактикою. Попри доступність інформації про здоровий спосіб життя, лише незначна частка населення дотримується всіх ключових рекомендацій щодо профілактики ССЗ. Приміром, дослідження у США показало, що лише близько 3% дорослих мають так званий «ідеальний» серцево-судинний профіль (нормальна вага, здорова дієта, достатня фізична активність, відсутність паління та нормальний рівень холестерину/глюкози). [24]

Навіть якщо брати менш суворі критерії, ситуація все одно невтішна: тільки ~11% дорослих виконують хоча б п'ять з п'яти основних правил здорового способу життя (раціональне харчування, фізична активність, відмова від тютюну, помірне споживання алкоголю, підтримання нормально ваги).

За даними ВООЗ, понад 1,4 млрд людей у світі практично не займаються фізкультурою, а близько 27% дорослих ведуть недостатньо активний спосіб життя всупереч рекомендаціям (150 хв помірної активності на тиждень). [9]

Аналогічно, нездорове харчування з надлишком солі, цукру та жирів залишається дуже поширеним, попри інформування про його шкоду. Недостатня профілактична активність населення означає збереження високих ризиків: якщо людина продовжує палити, нехтує дієтою і

рухливістю, то її шанси отримати інфаркт у рази вищі, ніж у того, хто змінив поведінку на краще. [8][7]

Отже, у суспільстві спостерігається розрив між теоретичною можливістю запобігти серцевим нападам і практичною реалізацією профілактики. Брак усвідомлення масштабів загрози та важливості превентивних заходів призводить до того, що серцево-судинні захворювання продовжують забирати життя тисяч людей щодня. Кардіологи наголошують, що підвищення обізнаності населення – критично важливе завдання. Необхідно, аби кожна людина розуміла: інфаркт може торкнутися кожного, але значною мірою йому можна запобігти.

Хвороби серця часто недооцінюються на особистому рівні. Подолати цю проблему можна через просвіту, скринінгові програми та мотивацію до здорового способу життя. Досвід показує, що коли люди краще поінформовані про ризики та симптоми, вони більш схильні змінювати поведінку і швидше звертаються по допомогу при насторожуючих ознаках

Таким чином, підвищення рівня знань і відповідальності населення за своє здоров'я є невід'ємною складовою стратегії зниження смертності від серцевих нападів.

1.4. ФОРМУВАННЯ ВЛАСНИХ ПОГЛЯДІВ НА ДОСЛІДЖУВАНУ ПРОБЛЕМУ

Бувають моменти, коли людина відчуває, що щось з її організмом не так. Це може бути просто інтуїтивне відчуття або більш об'єктивне відчуття як біль чи ущільнення в незвичному місці. Тіло зазнає негативних змін, але дії приймаються лише у випадку, коли потенційні ризики перевищують втрати. Консультація з лікарем може здаватись не вартою

часу та коштів, але усвідомлення неправильності таких міркувань настає лише тоді, коли наслідки розділяють життя на до та після.

Я вважаю що великою мірою проблема лежить саме у недооцінці ризику та помилкових упереджень щодо ССЗ. Молода людина буде покладатись на знання про те, що молодість є її захистом від кардіо-васкулярних хвороб. Атлетична особа буде вважати що регулярні зайняття спортом гарантують їй безпеку від хвороб серця. Або людина, що не палить та не вживає алкоголь матиме помилкові уявлення про свою ризик групу. Але як багато людей колись замислювалось та цікавилось як виглядають реальні жертви кардіо-васкулярних хвороб, у випадках коли особистий досвід знайомих та рідних відсутній?

Думаю, що більшість людей була б досить здивованою, якби побачила як виглядають реальні жертви ССЗ. Не рідко вони виглядають як звичайні, здорові люди за півроку до хвороби. Перші симптоми виявляються, але рідко, частота їх прояву все ще низька. Ці симптоми списуються на стрес, випадкові процеси організму чи просто як даність.

Молоді та спортивні, жінки та чоловіки, ніхто не застрахований від хвороб серця. Але поки існують помилкові упередження – консультація з лікарем буде залишатись «не вартою часу та грошей». Тим не менш, можливо, аби процедура оцінки цього ризику була дещо простішою, то люди були б більш заохочені до його моніторингу.

Ті, хто не збирався йти до лікаря аби дізнались про те, що знаходяться у групі ризику, можливо, ще один раз обдумали б цю ідею. А якби людина отримала низький рівень ризику це не мало б сенсу, так як вона все одно не збиралась користуватись медичною консультацією. Рівень надійності такого онлайн-інструменту недостатньо високий щоб визначити рішення цілком, але і недостатньо низький щоб повністю ігнорувати.

Таким чином, я доводжу висновку, що веб-додаток для оцінки ризику ССЗ є доступним та дієвим методом привернути увагу до проблем хвороб

серця та зробити внесок у руйнування помилкових упереджень про їх фактори ризику.

Висновки до розділу 1

Серцево-судинні захворювання, і зокрема інфаркт міокарда, нині становлять надзвичайно поширену та актуальну проблему, обумовлену значною мірою поведінковими факторами. Наслідки перенесеного серцевого нападу тяжкі – як фізично, так і психологічно, – і тягнуть значні медико-економічні витрати. Попри можливість профілактики більшості таких подій, проблема часто недооцінена: багато людей не знають свої ризики, не розпізнають симптоми і не вживають превентивних заходів. Це підкреслює необхідність активізувати просвітницьку роботу та впроваджувати сучасні методи оцінки ризику (зокрема, на основі поведінкових даних із застосуванням методів машинного навчання) з метою раннього виявлення осіб груп ризику та запобігання серцевим нападам до їх виникнення. Врешті, зміна акценту з лікування наслідків на упередження причин – ключ до подолання епідемії серцево-судинних захворювань у XXI столітті.

РОЗДІЛ 2. ДОСЛІДЖЕННЯ ДАНИХ ТА ПОБУДОВА МОДЕЛІ МАШИННОГО НАВЧАННЯ

2.1. ОГЛЯД ДЖЕРЕЛА ТА ЕКСПЛОРАТИВНИЙ АНАЛІЗ

Медичні дані є досить конфіденційною та чутливою темою, тому їх вільне розповсюдження є скоріше виключенням з правила. Окрім цього, пошук ускладнюється вузькою специфікою цих медичних даних – вони повинні бути актуальними, містити інформацію про факт серцевої хвороби людини та її поведінкові дані.

Загалом мені вдалось знайти лише один набір даних, що походить від авторитетного джерела та підходить під усі критерії. А саме [дані з опитування CDC](#), що є частиною Behavioral Risk Factor Surveillance System (BRFSS). Це щорічне телефонне опитування, яке проводиться для збору даних про стан здоров'я громадян США. База даних BRFSS є найбільшою безперервно проведеною системою опитувань у світі, яка збирає дані у всіх 50 штатах, окрузі Колумбія та трьох територіях США. Щорічно опитуванням охоплюється понад 400 000 дорослих.

Було обрано набір за 2022 рік з пропущеними значеннями. Набір даних містить всього 246 022 записів. Усього в даних 40 змінних, включаючи цільові змінні про факт серцевого нападу, інсульту або болей в грудях.

Серед 37 незалежних змінних 6 мають неперервних характер та 31 є категоріальними, серед яких 6 є мультикласовими. До оригінальних змінних увійшли: стать, вікова категорія, стан загального здоров'я (від погано до дуже добре), кількість днів за останній місяць з психологічними та фізичними труднощами, час останнього відвідування лікаря, факт перебігу обструктивного хронічного бронхіту, астми, COVID19, діабету, раку шкіри, депресивного розладу, артрити, хвороби нирок, факт тютюнопаління, паління електронних цигарок, факт споживання алкоголю, факт фізичної активності, факт наявності скану грудей за останній рік, факт наявності щеплень, факт проблем з зором, слухом, концентрацією, виконанням повсякденних задач, вдяганням та ходьбою, категорія раси та штат, вага, зріст, кількість годин сну.

Було прийнято рішення одразу відкинути змінні штату та раси, так як вони є дуже чутливими до регіону та вводять зайвий шум при екстраполяції даних на людей поза США.

В межах наступних кроків я виконав повноцінний аналіз пропусків, дуплікатів, розподілів, мультикореляції та кореляції з цільовою змінною. Розуміння природи даних є дуже важливим у контексті побудови моделей машинного навчання. Це перший і найважливіший крок у процесі побудови моделей по причині того, що існує безліч різних методів побудови машинного навчання. Десятки різних моделей, способів кодування даних, формування підвибірок, відбору, очистки та інжинірингу ознак. Якщо кожен з цих етапів сильно звузити, хоча б до 10 методів на етап, то для всього 5 етапів все ще існує $5^{10} = 100\,000$ можливих комбінацій цих методів. Ця проблема є фундаментальною в галузі машинного навчання та носить назву «No Free Lunch Theorems for Search». Вона вказує на те, що не існує універсальної моделі, а підбір найбільш оптимальної залежить від природи даних та її розуміння дослідником. [25]

Аналіз дуплікатів виявив що лише 150 записів з більше ніж 400 тис. є дуплікованими по всім незалежним змінним. Враховуючи кількість змінних

є дуже мало ймовірним, що 150 людей дало однакові відповіді на усі запитання. Виходячи з припущення що це була помилка під час збору даних було прийнято рішення видалити дані дуплікати.

Отже, вдалось з'ясувати, що дані містять велику кількість пропусків, єдина змінна без пропусків – стать. Більша частина має частку пропусків від 5% до 18%, що є суттєвим та потребує окремого розгляду.

Важливою особливістю даних виявився значний дисбаланс класів – 1 до 20. Цільова змінна є позитивною лише в 5% випадків. Це суттєво змінює перелік підходів та метрик до задачі.

Одним з головних викликів під час побудови моделі машинного навчання є витік даних. [26] Це явище, коли модель під час тренування отримує інформацію про цільову змінну, на якій відбувається її валідація, при тому не узагальнюючи дані. Це, в свою чергу, веде до значно завищеної оцінки ефективності моделі або навіть до спотворення її внутрішніх представлень. Однією з частих причин даного явища є «витікаючі ознаки». Тобто, ті ознаки, що не мають реальної предиктивної здібності, але мають високий рівень кореляції з цільовою змінною. Як правило це стається внаслідок порушення каузальних відношень між залежними та незалежними змінними. Наприклад, людина пересувається з палицею через труднощі з ходьбою, але палиця не є причиною цих труднощів, лише наслідком. Якщо необхідно передбачити труднощі з ходьбою у майбутньому цей фактор є поганим предиктором, так як він з'являється вже після появи цих труднощів.

Ця проблема особливо актуальна для даних дослідження, адже респонденти під час опитування надавали інформацію актуальну вже після нападу. Уважно переглянувши усі змінні я висунув теорію про витікання деяких змінних. Як ось наприклад змінні загального стану здоров'я, труднощів з пересуванням чи виконанням повсякденних задач, кількість днів за місяць з психологічними чи фізичними проблемами, останній візит до лікаря, факт скану грудей, факт наявності щеплень та вакцин. Адже після

серцевого нападу здоров'я людини, як фізичне, так і психологічне, передбачувано погіршиться, зумовивши зміни в раніше перерахованих змінних.

Наприклад, здоров'я серця впливає на здоров'я нирок та навпаки. [26] Або вакцинація проти грипу асоціюється зі зниженням частоти серцевих нападів серед людей із хворобами серця, особливо серед тих, хто переніс серцевий напад в минулому році. [27]

Дані розділялись на самому початку, перед проведенням біваріантного та уніваріантного аналізу, з використанням стратифікації по цільовій змінній, що є важливим при дизбалансі класів.

У ході побудови теплової матриці (рис. 2.1) пропусків виявилось що велика частина пропусків відбувається одночасно, що може свідчити про MAR типологію пропусків, тобто у пропусках є певна закономірність і вони не є повністю випадковими (MCAR). [28] Використання теста Літла підтвердило гіпотезу та допомогло встановити, що лише у 6% випадків пропуску є незалежними один від одного. [29]

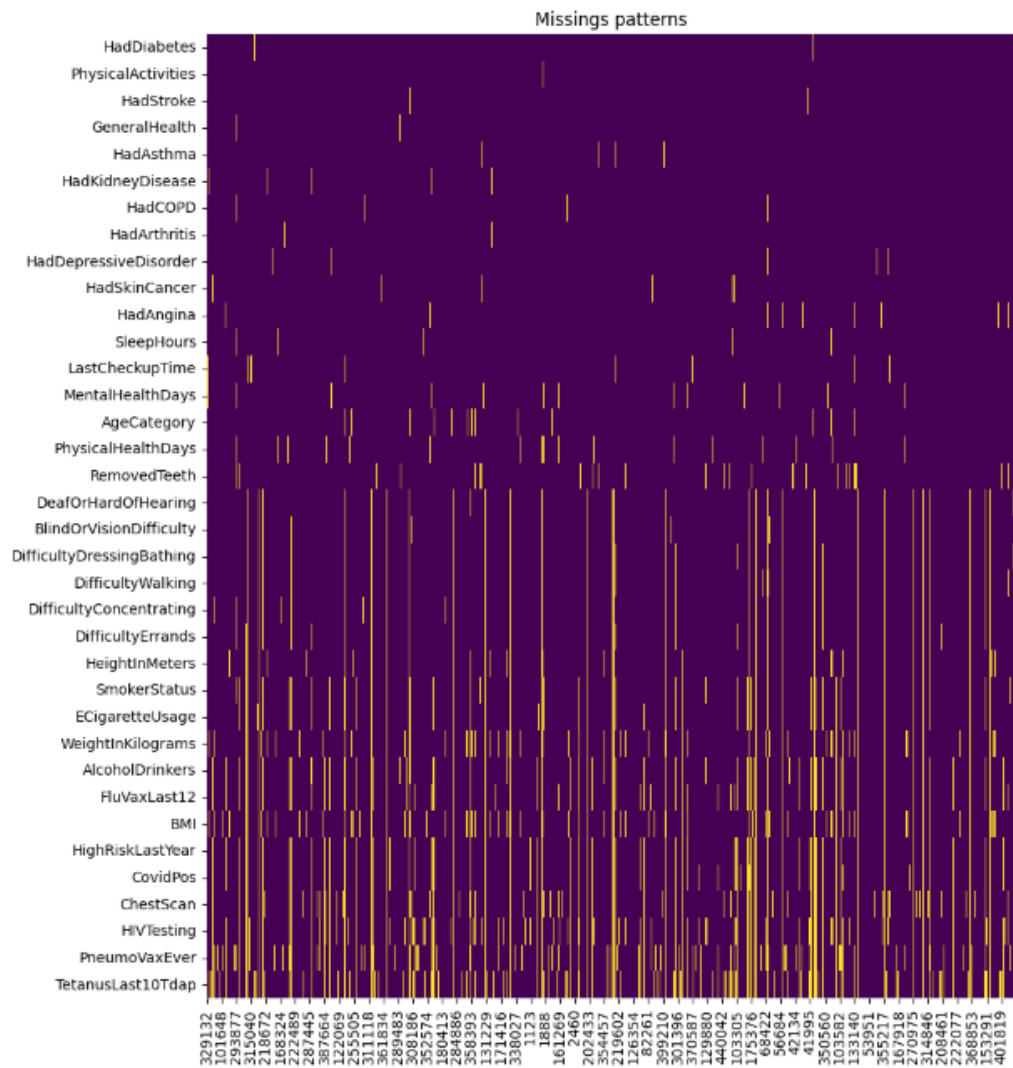


Рис. 2.1 - Теплова карта пропусків

Виходячи з даної знахідки та факту значної частини пропусків в даних я можу дійти висновку, що має сенс використовувати просунуті техніки імпутації типу RandomForestImputer та утриматись від стандартної імпутації середнім чи модою. [30]

В межах уніваріантного аналізу розподілів змінних (рис 2.2) вдалось з'ясувати що велика кількість категоріальних змінних має сильний дизбаланс класів.

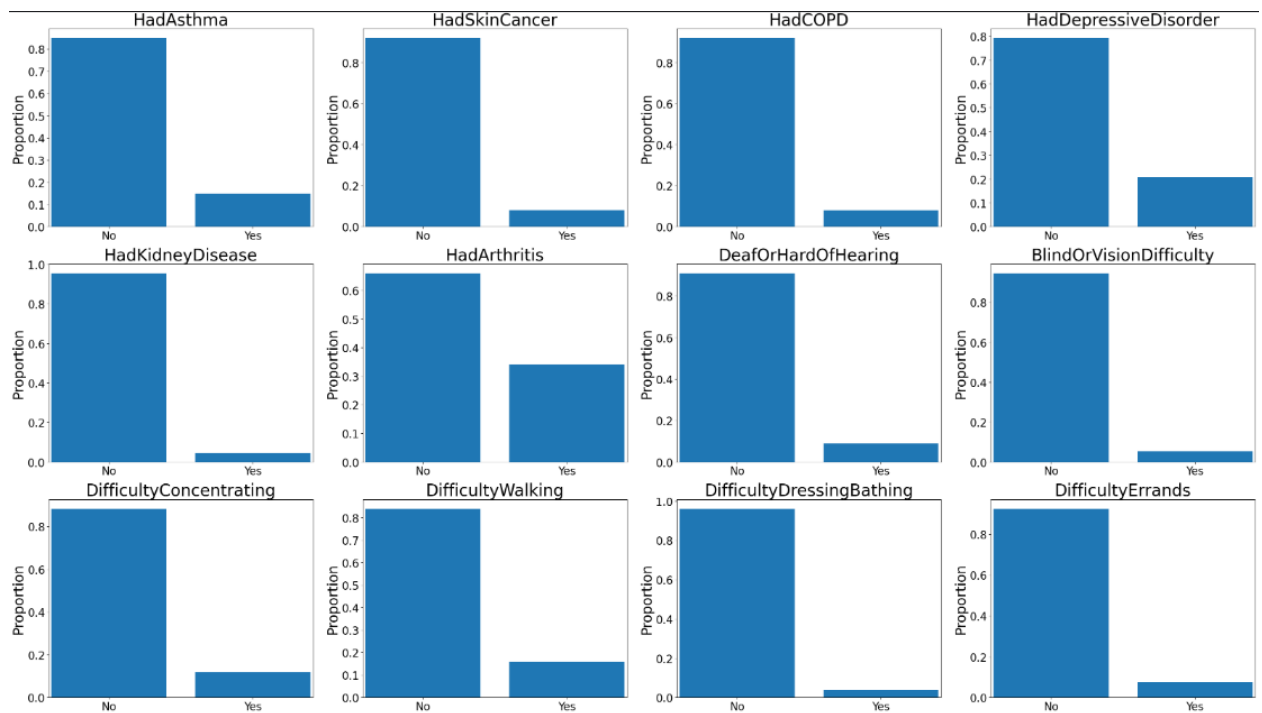


Рис. 2.2 - Розподіл класів в бінарних ознаках

І все ж таки, розподіл класів в деяких змінних викликає підозри у витоці даних. Наприклад, близько 20% позитивних випадків у змінної факту наявності депресії. Хоча за офіційною статистикою лише 6% дорослого населення США страждало на депресивний розлад станом на 2021 рік. [32]

Розподіл класів в мультикласових ознаках (рис 2.3) не викликає питань, але й не надає практично корисної інформації про дані.

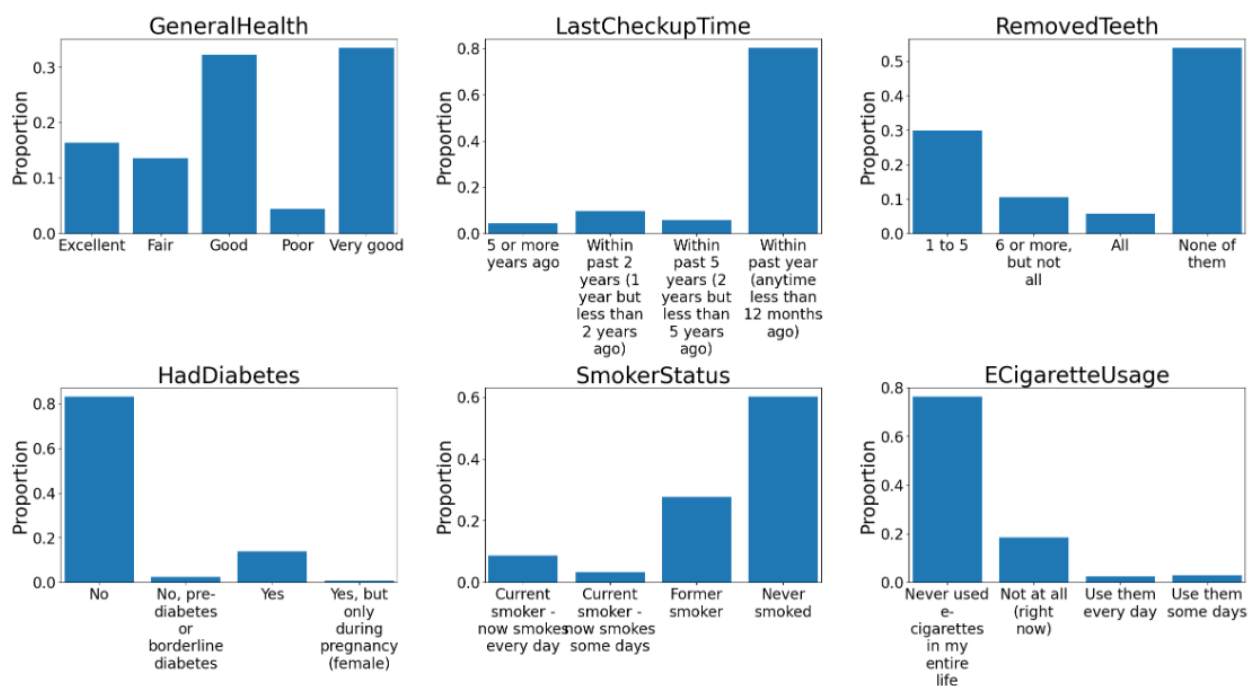


Рис. 2.3 - Розподіл класів мультикласових ознак

Розподіли в неперервних ознаках (рис 2.4) дещо далекі від гаусівських, зокрема розподіли ознак кількості днів за останній місяць з фізичними чи психологічними проблемами мають довгий хвіст направо.

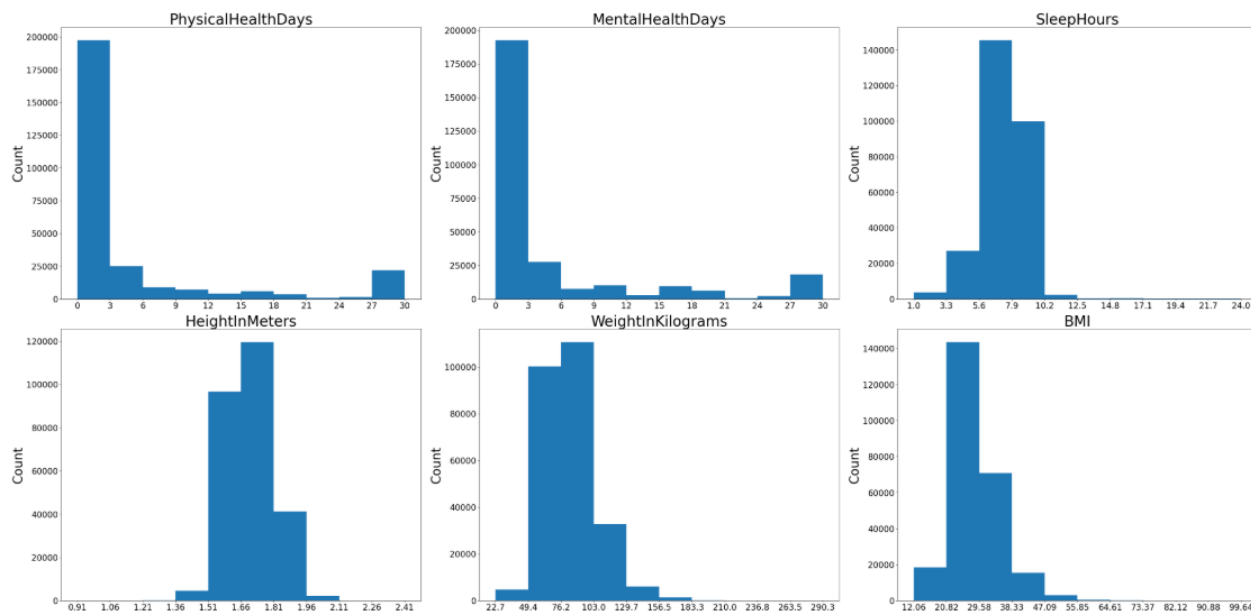


Рис. 2.4 - Розподіли неперервних ознак

Це є проблемою, так як нормальна форма розподілу в неперервних ознаках є дуже важливою для моделей, заснованих на градієнтному спуску.

[31]. Таким чином я доходжу висновку, що існує необхідність у трансформації неперервних ознак та приведенні їх до більш подібного до гаусівського розподілу.

Вікові категорії розподілені з перекосом у бік більшого віку (рис 2.5), зокрема спостерігається стрімке збільшення починаючи з категорії 55+.

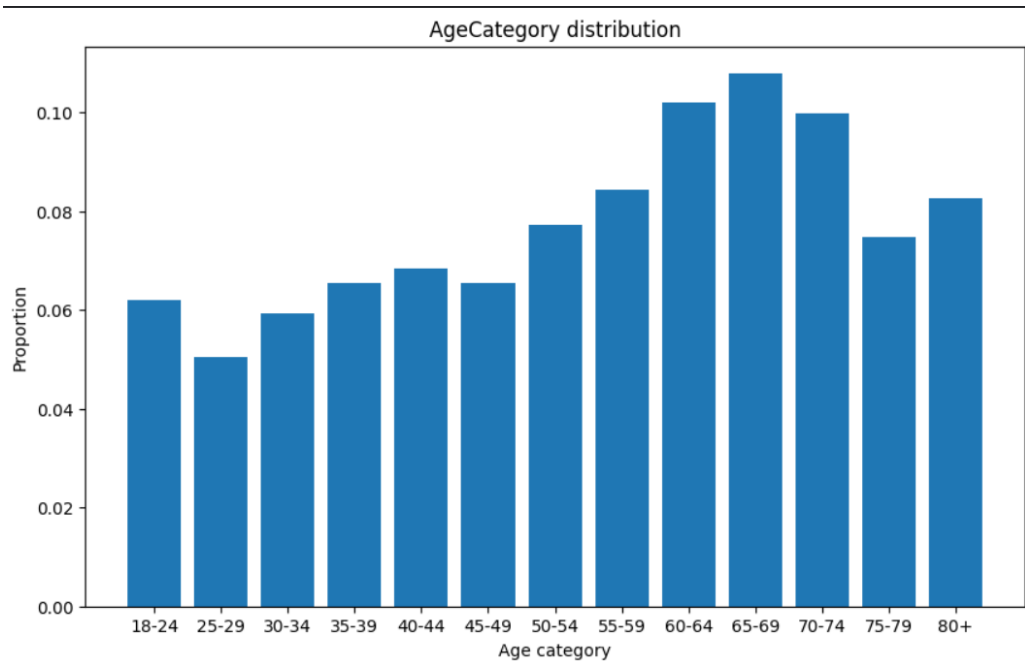


Рис. 2.5 - розподіл вікових категорій

В межах біваріантного аналізу я провів візуалізацію кореляції ознак до цільової змінної, де чорним виділено розподіл всієї змінної, а червоним – частка позитивних випадків цільової змінної в даній категорії.

Дані (рис 2.6.) підтверджують інформацію з раніше приведених досліджень, де вказувалось, що стать та рівень фізичної активності тісно пов'язані з частотою хвороб серця. Тим не менш, різниця не є колосальною, жертвами ССЗ є як чоловіки, так і жінки, як фізично активні, так і фізично менш активні люди.

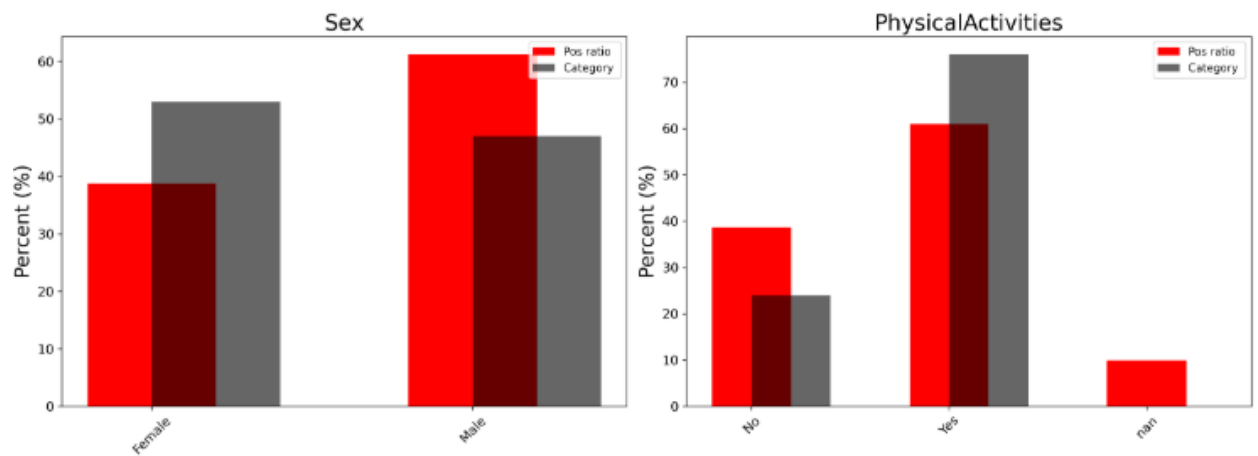


Рис. 2.6 - Взаємозв'язок статі та фізичної активності з ССЗ

Існує очевидна кореляція обструктивного бронхіту, хвороби нирок, артриту, проблем з слухом та зором з цільовою змінною (рис 2.7), хоча не до кінця зрозуміли чи є ці відношення каузальними. Тим не менш, диференціація по ознаці депресії є слабкою, що спростовую попередню підозру про витік даних з неї.

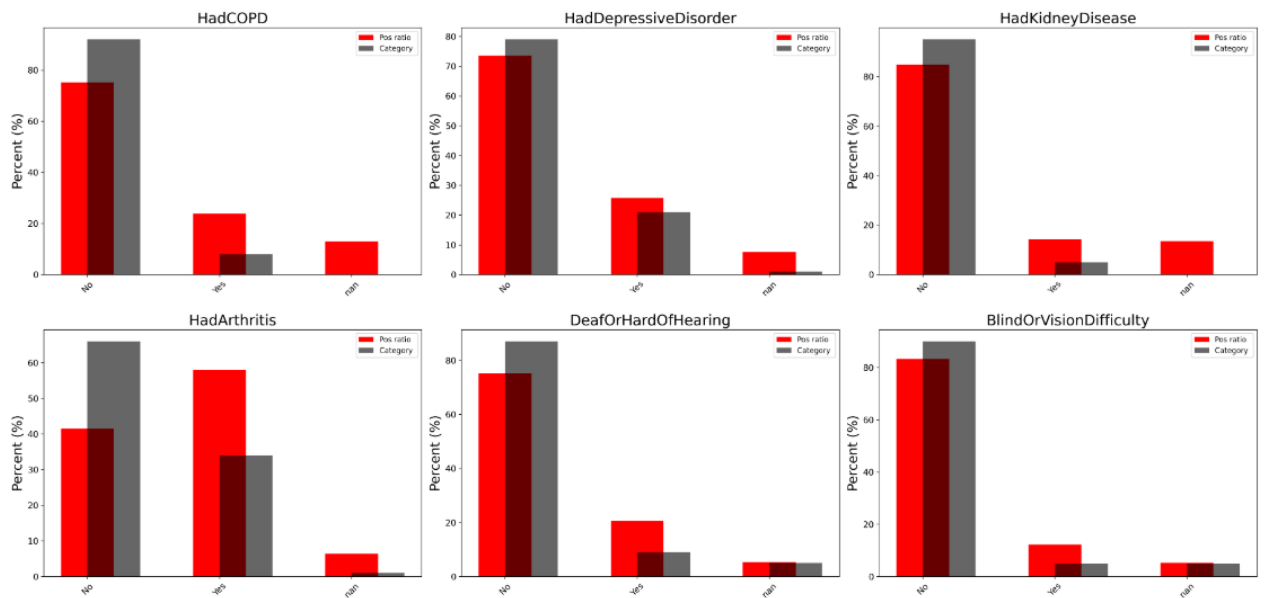


Рис. 2.7 - Кореляція з цільовою змінною незалежних ознак з підозрою на витік даних

Вдалось підтвердити сильний зв'язок між споживанням алкоголю та хворобами серця (рис. 2.8). Тим не менш кидається в очі дещо контрверсійний сильний зв'язок наявності скану грудей та ССЗ. Навряд чи скан грудей може сильно збільшити ризик хвороб серце, але може бути проведений після факту серцевого нападу чи інсульту. Що дає гарні підстави вважати дану ознаку витікаючою.

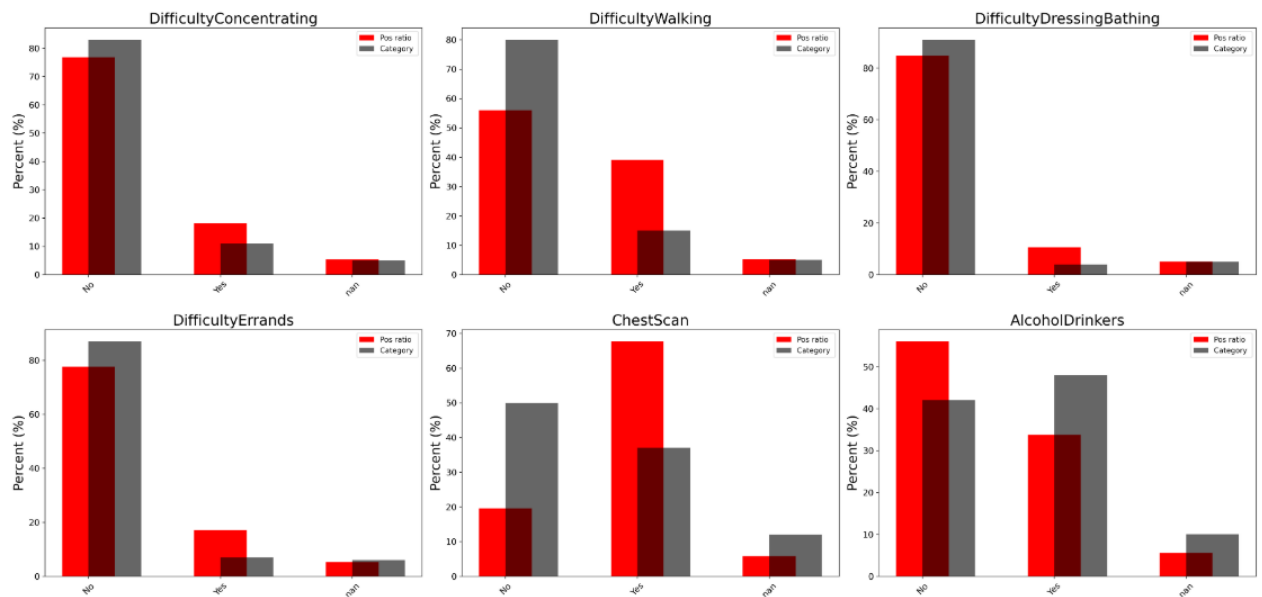


Рис. 2.8 - Корреляція з цільовою змінною ознак споживання алкоголю, скану грудей та інших

Цікаво виглядає незначний, але явний взаємозв'язок між тестуванням на ВІЛ та хворобами серця (рис 2.9), зокрема є дуже інформативною ознака пропуску в даній змінній. Сильний взаємозв'язок між щепленням від грипу та ССЗ підсилює підозру про витік даних з цієї ознаки.

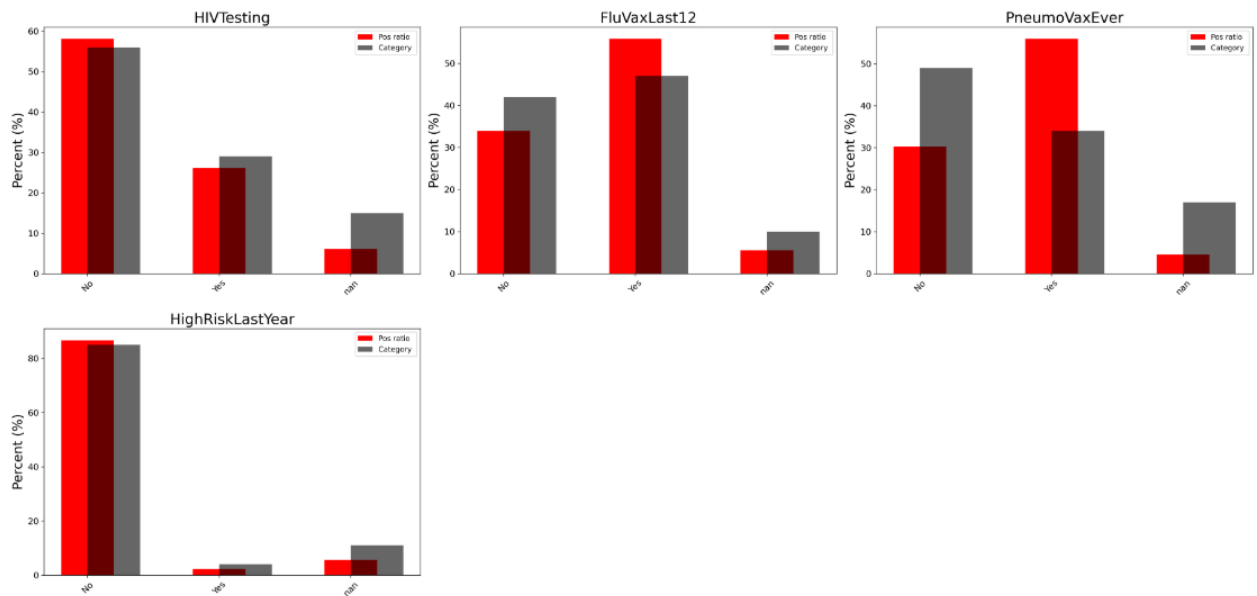


Рис. 2.9. - Кореляція з цільовою змінною ознак щеплення та вакцинації

Є очевидним сильний взаємозв'язок між станом загального здоров'я та цільовою змінною (рис. 2.10). Незважаючи на це, більшість випадків серцевих хвороб знаходяться саме у категорії гарного стану здоров'я, що робить інтерпритацію ознаки важкою та підтверджує попередні твердження про небезпеку ССЗ навіть для зовні здорових людей.

Неочікувано виглядає позитивна кореляція між цільовою змінною та ознакою останнього часу відвідування лікаря (рис. 2.10). Було б логічно припустити, що чим частіше людина буває у лікаря, тим краще вона слідує за здоров'ям та відповідно має менший ризик хвороб серця. Тим не менш, більше схоже що в даному випадку дані не можуть правильно захопити цю залежність. Можливо через упередження виживаючого, коли люди що відвідують лікаря рідше мають менше шанс взяти участь в опитуванні та пережити серцевий напад. [33] Так чи інакше ознаку, скоріш за все, доведеться видалити.

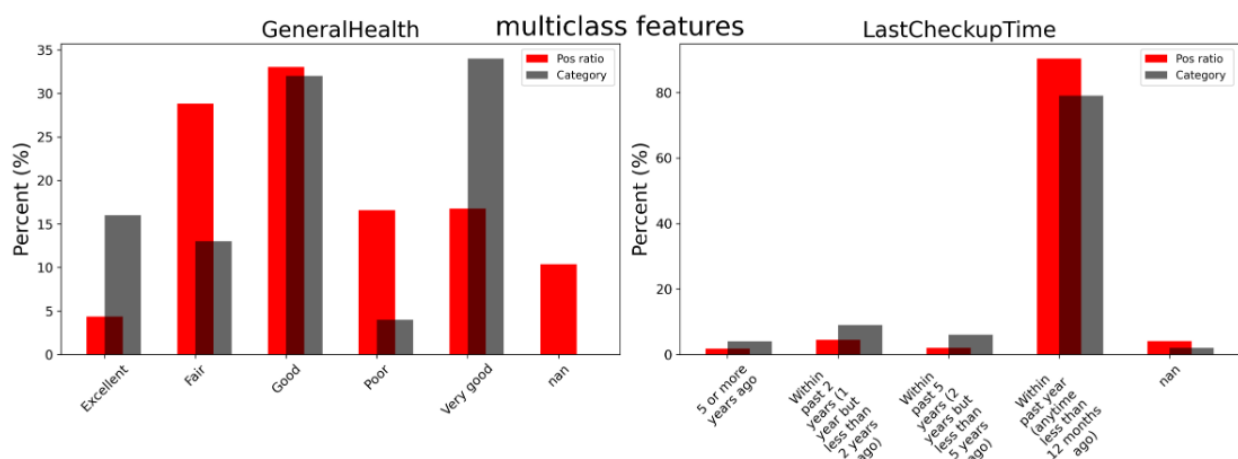


Рис. 2.10 - Кореляції цільової змінної з ознаками загального стану здоров'я та часу відвідування лікаря

Дані підтверджують твердження про сильний взаємозв'язок між палінням та хворобами серця (рис. 2.11). Але це справедливо лише для паління тютюну. Схоже, що диференціація по палінню електронних цигарок або майже відсутня, або дуже нелінійна, або сильно корельована з віком.

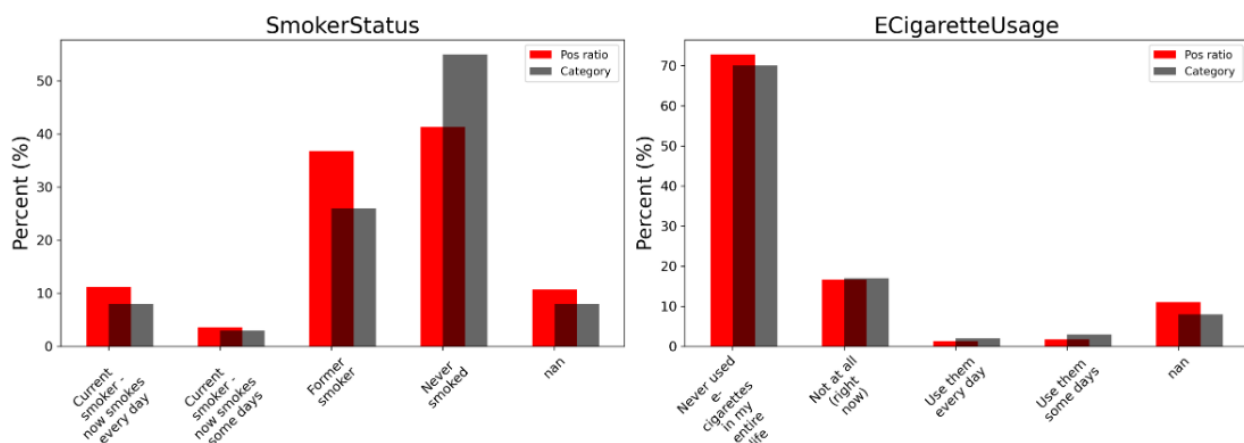


Рис. 2.11 - Кореляція між цільовою змінною та ознаками паління

Взаємозв'язок між віком та хворобами серця виглядає досить сильним та лінійним (рис 2.12). Що цікаво, пропуски в ознаці щеплення від правця є досить інформативними.

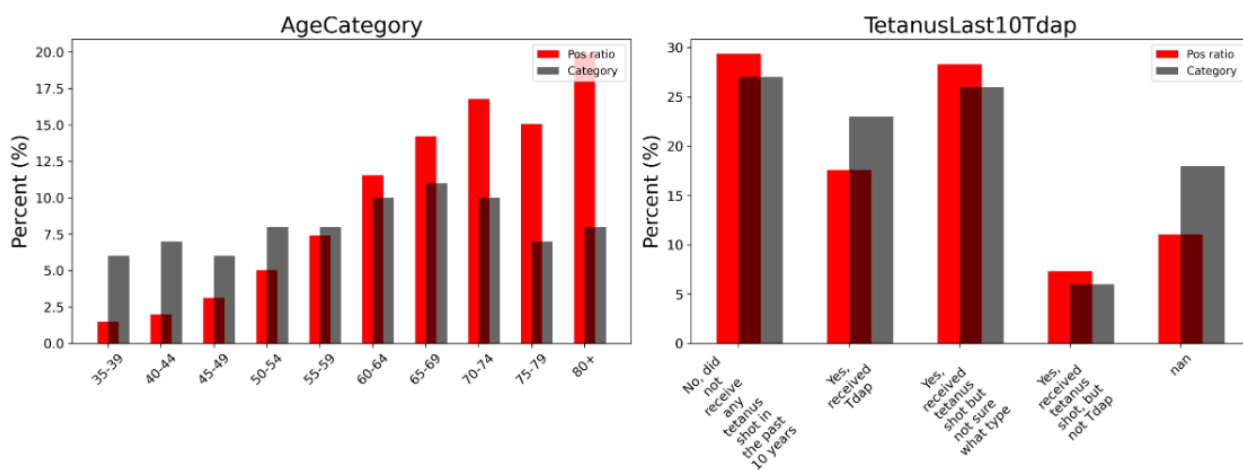


Рис. 2.12 - Кореляція між цільовою змінною та ознаками віку і щеплення від правця

Ознака кількості видалених зубів показує чітку кореляцію з цільовою змінною (рис. 2.13). Тим не менш, ця ознака може бути сильно мультикорельованою з віком або з статусом паління, це означає що варто приділити увагу визначенню мультикореляцій серед незалежних ознак. В той же час пропуски в даній ознаці виглядають досі інформативними.

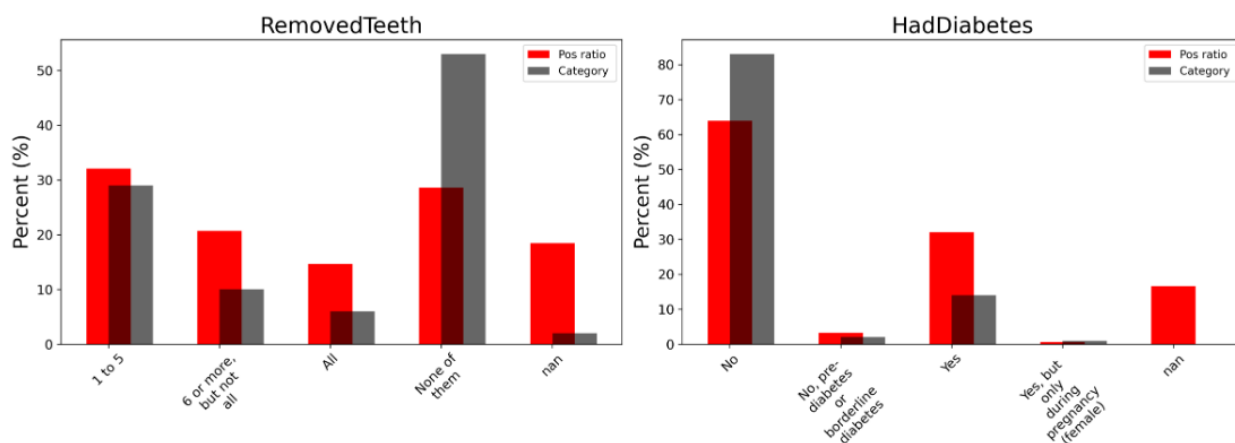


Рис. 2.13 - Кореляцій між цільовою змінною та ознакою діабету і видалених зубів

З поточних результатів схоже, що взаємозв'язок між COVID19 та хворобами серця практично відсутній, або він є нелінійним (рис. 2.14).

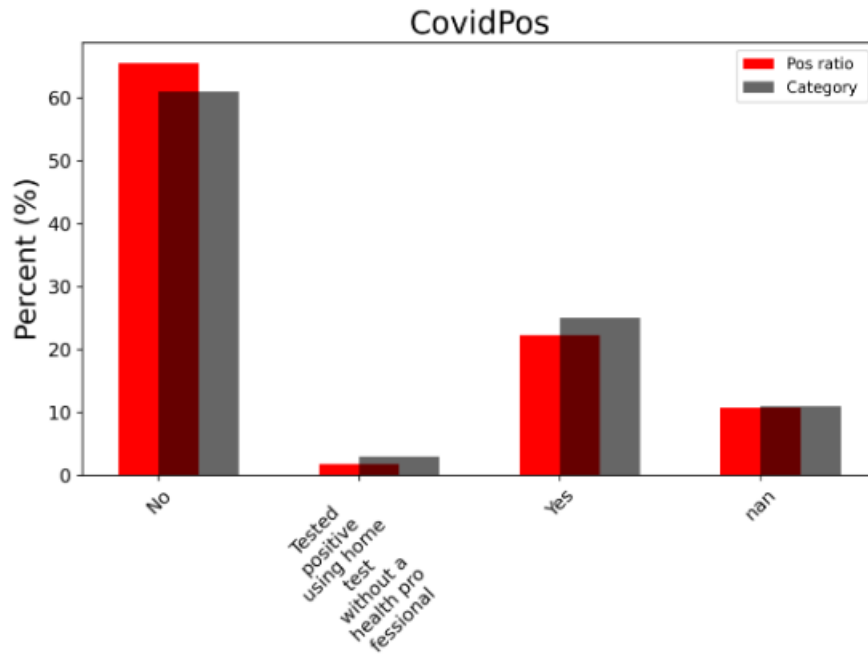


Рис. 2.14 - Кореляція між цільовою змінною та ознакою наявності COVID19

За результатами біваріантного аналізу вдалось зробити кілька ключових висновків:

1. Абсолютна більшість пропусків у даних належать MAR типології. Для імпутації варто застосувати моделі ітеративної імпутації
2. Розподіли більшості ознак є нерівномірними. Це ускладнює аналіз мультиколінеарності та корельованості ознак з цільовою змінною, відкидаючи алгоритми типу chi-square
3. Цільова змінна має значний дисбаланс класів, що вимагає застосування специфічних метрик та методів балансування класів
4. Неперервні ознаки вимагають трансформації та приведення до нормального розподілу
5. Пропуски багатьох змінних несуть корисну інформацію про цільову змінну
6. Деякі категоріальні змінні показують тенденції до мультиколінеарності

7. Більшість підозр про витік даних не підтвердилось. Але з'явилися вагомі підстави видалити ознаки наявності скану грудей та часу відвідування лікаря

2.2. ТРАНСФОРМАЦІЯ, ЕНКОДИНГ, СТВОРЕННЯ НОВИХ ОЗНАК ТА ІМПУТАЦІЯ ДАНИХ

Жодна модель машинного навчання не може працювати з текстовими представленнями напряму. Так чи інакше необхідно виконати «кодування» категоріальних змінних, їх приведення до числового представлення.

На сьогодні існує безліч різних методів кодування змінних, серед яких не існує універсальних. Тим не менш, існує набір базових технік кодування, що стабільно показують гарні результати. Серед них ordinal, onehot, frequency, target, woe (weight of evidence) кодування.

Ординальне кодування використовується для ознак, категорії яких мають природний порядок. Між такими категоріями можна однозначно виміряти відстань. Наприклад, категорія 55 років більше за категорію 50 років. Ця техніка привласнює цілочисельні мітки, які збільшуються пропорційно порядку категорії. Для цієї техніки є важливим правильний порядок категорій.

Техніка onehot на противагу ординальному кодуванню передбачає незалежність категорій одна від одної. Для кожної категорії ця техніка створює вектор, який набуває ненульового значення лише за факту наявності категорії. Головним недоліком даної техніки є стрімке зменшення розмірності, адже якщо у ознаки, наприклад, 20 класів, то на виході буде отримано 20 нових бінарних ознак. [34]

Таргетний, frequency та woe енкодер зазвичай знаходять своє застосування у випадках високої кардинальності категоріальних ознак та незалежності категорій. Зокрема таргетний енкодер кожній категорії

привласнює міру її асоціації з цільовою змінною. Ця міра може змінюватись в залежності від варіації таргетного енкодера, але цей метод має значний недолік – для запобігання витоку даних необхідно створювати додаткову тренувальну підвибірку. Це особливо критично у контексті дизбалансу класів, коли кожен позитивний випадок є дуже цінним. [35]

В умовах відсутності категоріальних ознак високої кардинальності та факту значного дизбалансу класів цільової змінної було прийнято рішення обмежитись стандартними методами кодування – ordinal та onehot енкодинг.

Для трансформації та наближення неперервних ознак до нормального розподілу було протестовано 3 різних підхода – через log трансформацію, sqrt трансформацію та трансформацію Йео-Джонсона. [36]

Виходячи з порівняльного аналізу (рис. 2.15 та рис. 2.16) можна помітити, що log трансформація є найбільш «м'якою», зберігаючи оригінальну пропорцію, при цьому наближаючи розподіл до нормального.

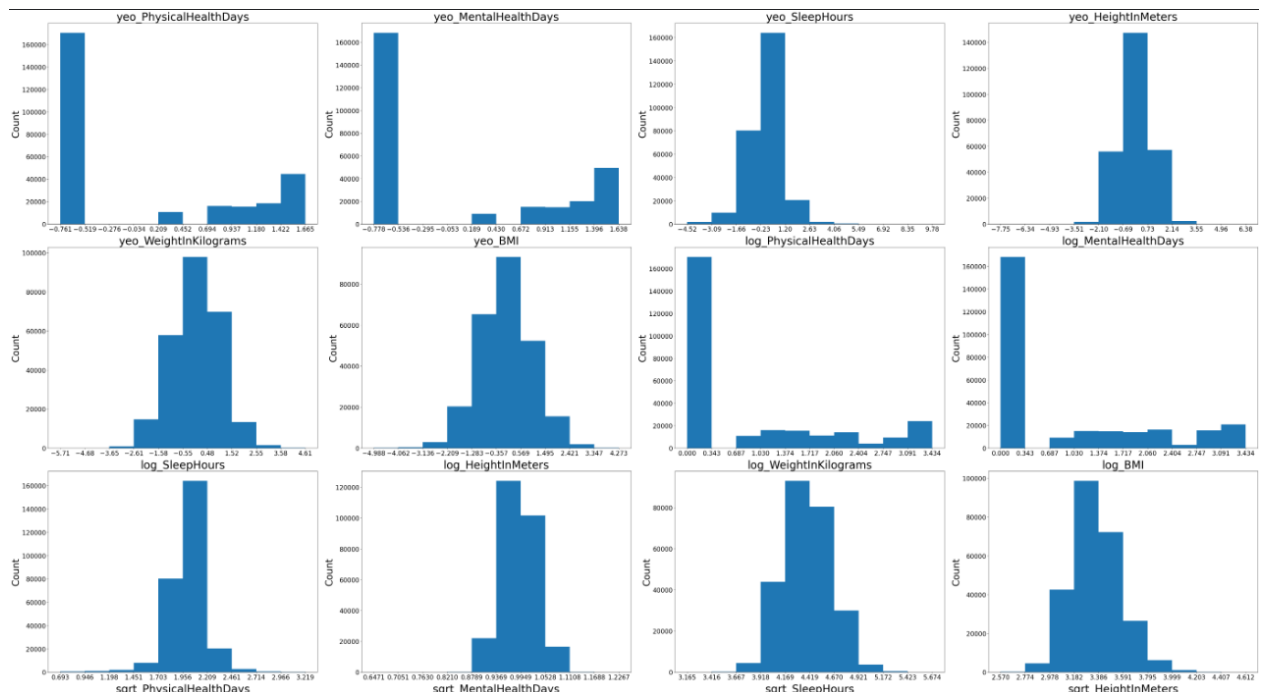


Рис. 2.15 - Порівняльний аналіз log трансформації та трансформації Йео-Джонсона

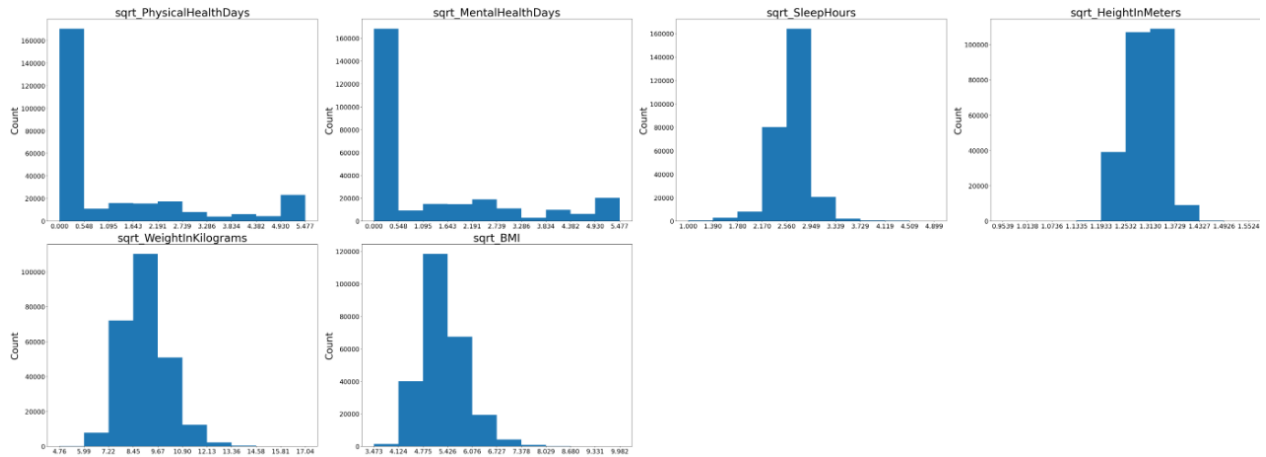


Рис. 2.16 - Порівняльний аналіз sqrt трансформації

Під час підбору імпутеру було з'ясовано, що популярні python бібліотеки не надавали просунутих методів імпутації, які могли б працювати на одиночному інференсі. Це важливо, адже весь цикл обробки даних необхідно буде повторити під час роботи застосунку. Користувач буде надсилати всього 1 запит, отже імпутер повинен вміти оброблювати всього 1 запис за раз. У ході подальшого аналізу з'ясувалось, що ці ж самі бібліотеки не надають подібних імпутерів в цілому. Єдиний вихід який залишався – писати власну реалізацію імпутера, яка без проблем працювала б для одного запису за раз.

Ситуацію погіршила необхідність створювати окрему підвибірку даних для тренування просунутих імпутерів. Окрім того, існує досить багато варіацій просунутих імпутерів. Отже, довелось піти на компроміс та скористуватись простим імпутером, жертвуючи точністю в обмін на надійність та збереження повного обсягу початкових тренувальних даних. Повний код реалізації імпутера знаходиться в Додаток 1.

У контексті створення нових ознак використовувався MissingIndicator трансформер, який за допомогою onehot методу кодує пропуски кожної категорії класу.

Для дослідження мультиколінеарності використовувалась AMI (Adjusted Mutual Information) метрика. Вона має ряд переваг перед іншими метриками виміру асоціації, так як є симетричною, стійкою до дизбалансу

класів та різниці кількості категорій. Її єдиним недоліком є проблема інтерпритації, так як масштаб визначає природою кожного набору даних індивідуально.

Повний код препроцессора даних знаходиться в Додаток Б.

2.3. ПОБУДОВА МОДЕЛЕЙ ТА ВИБІР МЕТРИК ТА МЕТОДІВ ВАЛІДАЦІЇ

Перш ніж перейти до побудови моделей необхідно вирішити питання об'єктивної оцінки їх ефективності. Тільки базова реалізація scikit-learn пропонує більше 20 готових різних метрик для задач бінарної класифікації. Завдяки попередньому дослідженню даних вдалось дізнатись, що необхідна метрика повинно бути чутливою до дизбалансу класів.

Проблема у тому, що більшість стандартних метрик можуть надати занадто оптимістичну оцінку моделі, так як вони оцінюють загальну точність моделі, замість оцінки кожного класу окремо.

Наприклад, класична точність. У випадку коли у цільовій змінній присутній баланс класів 1 до 99, то просто постійно передбачаючи негативний клас точність для позитивного класу буде складати 0%, в той час як для всієї цільової змінної точність буде 99%.

Ще однією популярною метрикою є AUC (Area Under Curve), яка на відміну від ассигасу, яка оцінює ефективність моделі по одному трешхолду (зазвичай 0.5), оцінює всю криву трешхолдів та дає більш повну оцінку узагальнюючій здібності моделі. Тим не менш, ця метрика теж погано підходить до оцінки моделі при значному дизбалансі класів, так як теж не враховує важливість позитивного класу. [38]

Одними з центральних метрик оцінки ефективності моделей в контексті дизбалансу класу є «precision» та «recall». Вони працюють на основі оцінки повної конфузійної матриці, враховуючи кількість

дійснопозитивних (tp), хибнопозитивних (fp), дійснонегативних (tn) та хибнонегативних (fn) передбачень кожного класу. [39]

- Accuracy (Точність): Відсоток правильно передбачених випадків серед усіх передбачень.

$$\frac{(TP + TN)}{(TP + FP + TN + FN)}$$

- Precision (Чіткість\Точність): Відсоток правильних передбачень серед усіх передбачень позитивного класу.

$$p = \frac{TP}{TP + FP}$$

- Recall (Повнота): Відсоток правильних передбачень позитивного класу серед усіх реальних випадків позитивного класу.

$$r = \frac{TP}{TP + FN}$$

Також вирізняють і F1 метрику, яка фактично визначає співвідношення між чіткістю та повнотою. Вона має кілька різних варіацій, в тому числі арифметичне середнє та гармонійне середнє.

Але, усі раніше перераховані метрики також мають значний недолік. Вони оцінюються узагальнюючи здатність моделі лише на одному конкретному трешхолді за раз. Тобто, у просторі трешхолдів моделі може існувати краще співвідношення, яке ці метрики не враховують.

Недоліку попередніх метрик полишена average precision, яка фактично інтегрує увесь простір трешхолдів, враховуючи важливість позитивного класу, обраховуючи precision на кожному кроці.

І врешті-решт останній спосіб оцінки моделей, який хоч і не інтегрує весь простір трешхолдів, але дозволяє отримати максимально повну інформацію по конкретному трешхолду – конфузійна матриця. За її допомогою можна не тільки зрозуміти точність та покриття, а ще й дізнатись які саме помилки модель зазвичай робить.

Для виконання інференсу використовувався метод кросс-валідації, який дозволяє отримати передбачення для усіх міток, як тренувальних, так і валідаційних, поступово зміщуючи тренувальний фолд на ті мітки, що до цього використовувались як валідаційні. [40]

У ході порівняльного аналізу було використано 3 моделі:

1. Логістична регресія newton-cholesky. На відміну від класичної логістичної регресії заснованої на градієнтному спуску, цей метод підганяє параболу до кривизни цільової функції (функції втрат) і мінімізує квадратичну функцію параболу, що сприяє швидшій збіжності та точному оновленню порівняно з градієнтним спуском (рис. 2.17). [41]

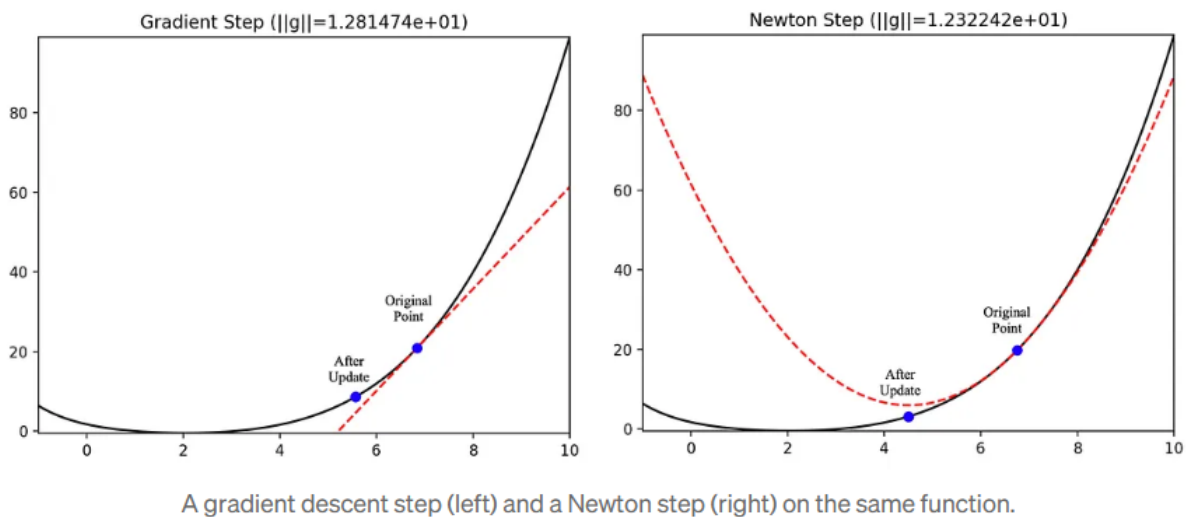


Рис. 2.17 - Принцип роботи newton-cholesky алгоритму оптимізації

2. RandomForest – класичний ансамбль дерев рішень, кожне з яких досліджує випадкову підвибірку ознак.
3. Lightgbm – бустинговий ансамбль дерев, що вирізняється своєю швидкістю роботи, розроблений Microsoft у 2017 році. [42]

За результатами тестування (рис. 2.18, 2.19 та 2.20) видно, що усі 3 моделі мають приблизно однаковий рівень точності.

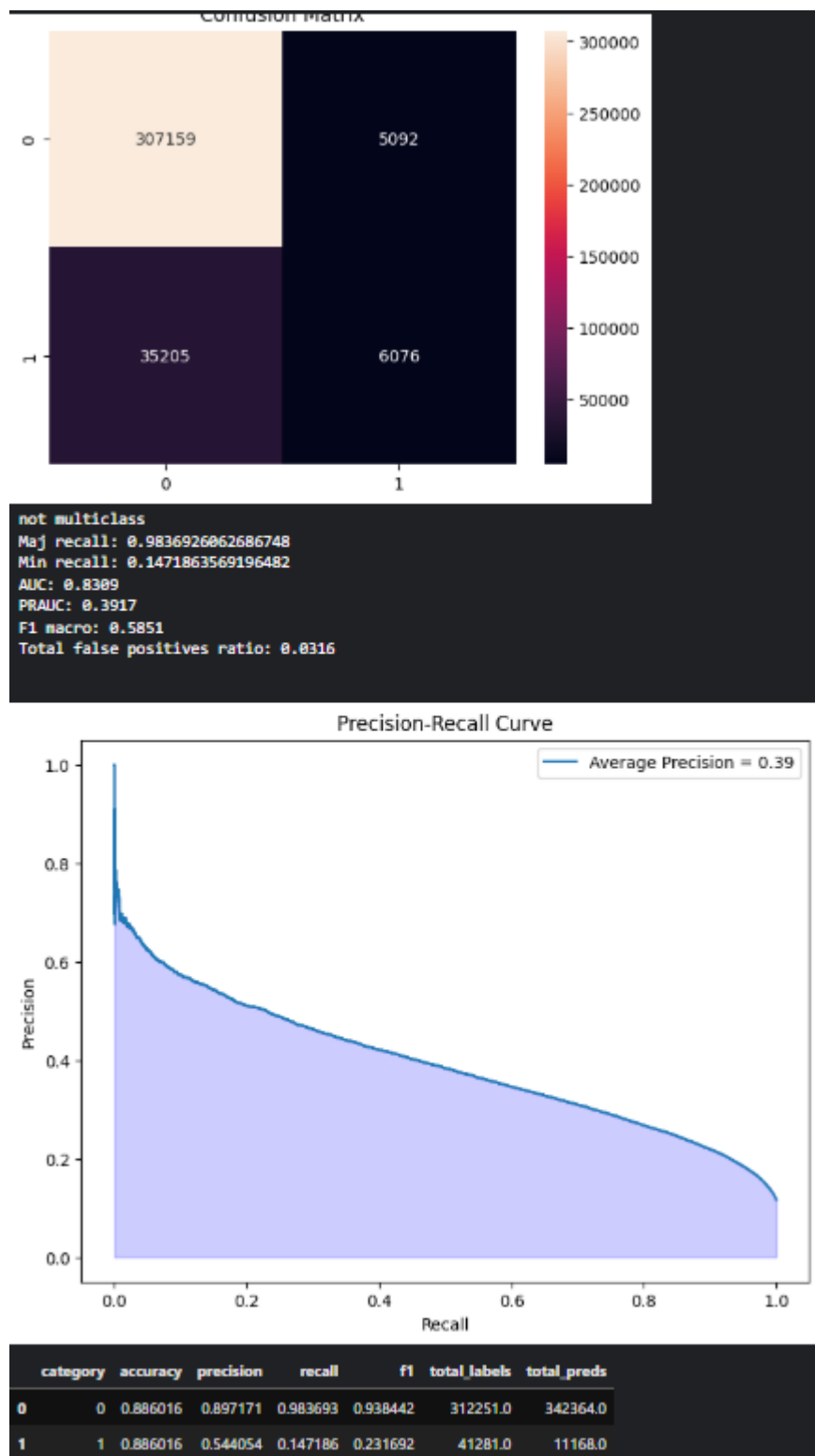


Рис. 2.18 - Результати тестування newton-cholesky логістичної регресії

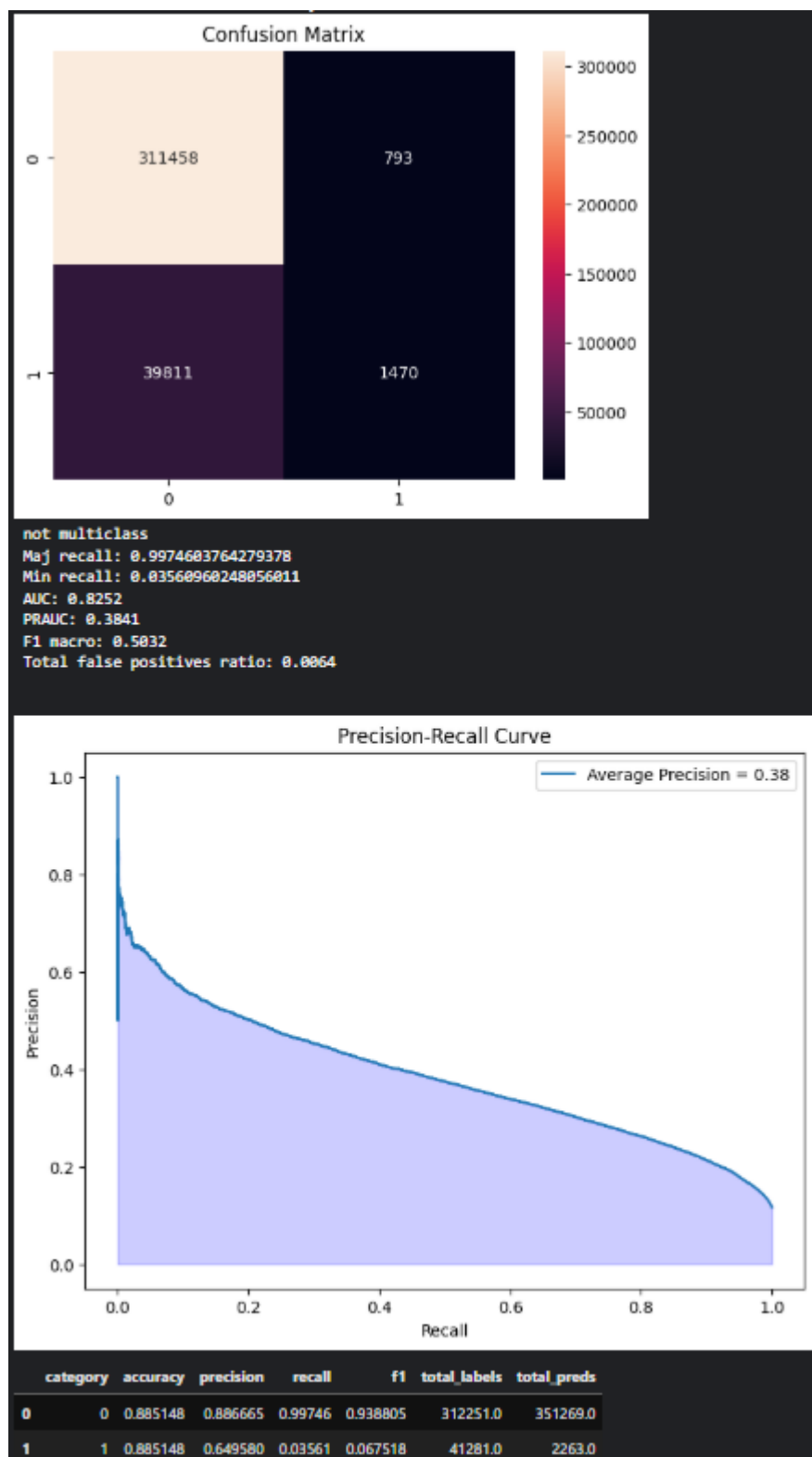


Рис. 2.19 - Результати тестування Random Forest

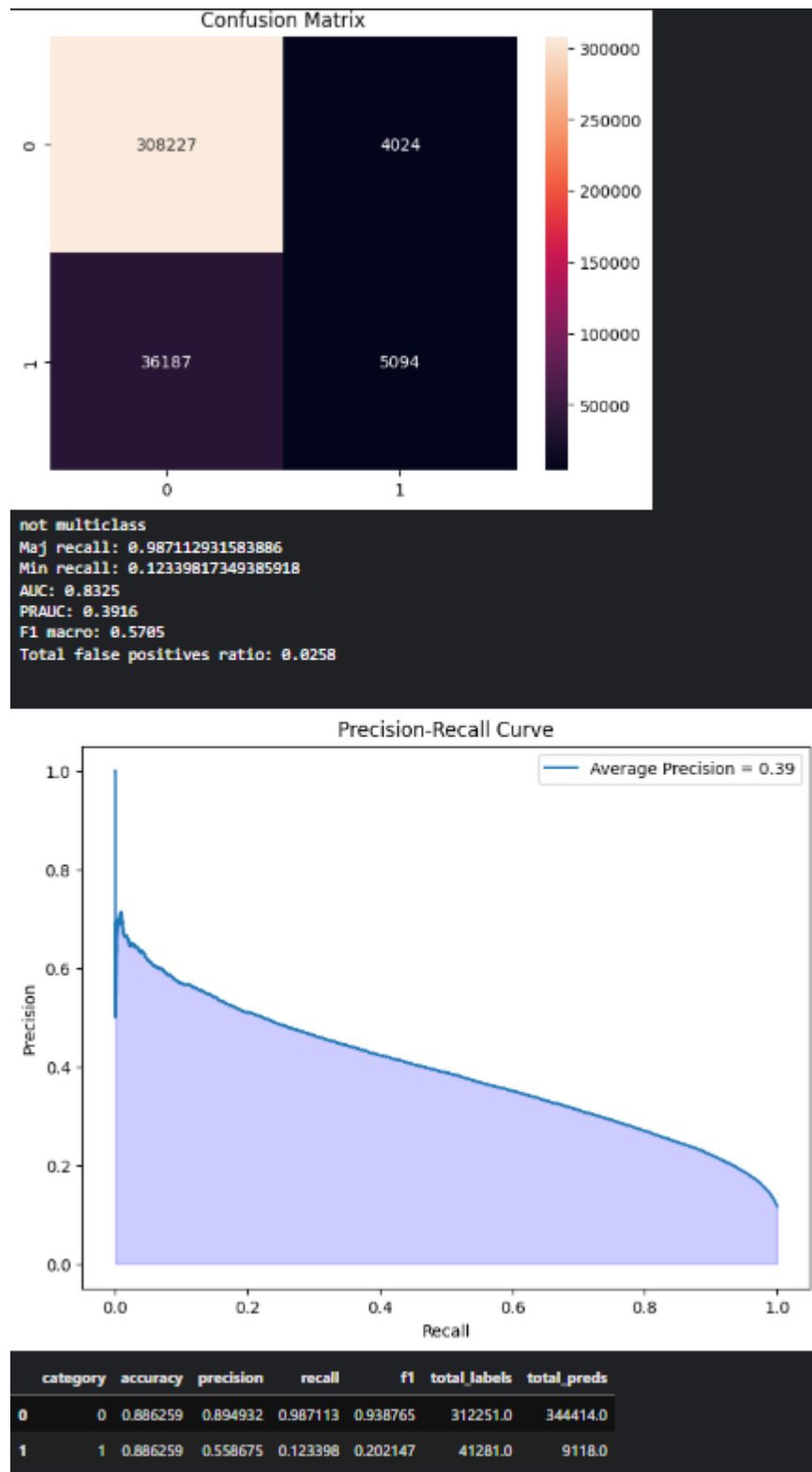


Рис. 2.20 - Результати тестування Lightgbm

З даних результатів можна зробити кілька висновків:

1. У даних домінують лінійні залежності
2. Мультиколінеарність у даних практично відсутня, так як логістична регресія дуже чутлива до неї.

3. Скоріш за все, за допомогою оптимізації гіперпараметрів, ціною ризику витоку даних, можливо підвищити середній precision ще на кілька пунктів.
4. У ознаках практично відсутні зайві ознаки, так як ансамблеві методи більше чутливі до цього, ніж логістична регресія.

Висновки до розділу 2

Завдяки застосування сучасних методів аналізу, обробки та узагальнення даних вдалось підтвердити можливість побудови моделі передбачення хвороб серця на основі виключно поведінкових даних. Окрім того, вдалось підтвердити впливовість загальноприйнятих основних факторів ССЗ.

У результаті вдалось досягнути середньої точності передбачення серцевої хвороби в 40% по всьому простору трешхолдів моделі. Також, існує можливість досягнути близько 70% точності за умови зменшення рівня покриття до 10-20%.

Порівняльний аналіз моделей показав, що залежності в даних є переважно лінійними і легко інтерпритованими.

РОЗДІЛ 3. СТВОРЕННЯ, ТЕСТУВАННЯ ТА ДЕПЛОЙ ВЕБ-ЗАСТОСУНКУ

3.1. ОБҐРУНТУВАННЯ ВИБОРУ ТЕХНОЛОГІЙ ТА АРХІТЕКТУРА ДОДАТКУ

Для розробки застосунку було наслідовано мову програмування, на якій було виконано тренування моделі, а саме Python. Це необхідно для узгодженої роботи модулів самої моделі та веб-застосунку в єдиній програмній оболонці, без розділення архітектури на кілька окремих мікросервісів. Кінцевий продукт представляє опитувальник, який за результатами відповідей користувача на питання буде обраховувати групу ризику користувача.

Тобто, додаток полишений важких обрахунків та необхідності функціонувати з максимально можливою швидкістю, мінімізуючи затримку. Це дозволяє не фокусуватись на продуктивності, але більше зосередитись на надійності та зрозумілості.

Серед головних вимог до сервісу:

1. Зрозумілість та простота інтерфейсу для користувача. Інтерфейс повинен бути написаний повністю українською мовою, а питання поставлено чітко та зрозуміло.
2. Зовнішня привабливість та відповідність тематиці. Інтерфейс повинен використовувати строгі, але прості кольори, які будуть відповідати тематиці, асоційованій зі здоров'ям.
3. Надійність та стабільність. Додаток повинен не допустити можливість некоректного вводу користувачем та продовжувати функціонувати без збоїв.

4. Легкість інтерпритації кінцевого результату. Користувач без знань у галузі машинного навчання повинен без додаткових уточнень та двозначності зрозуміти результати оцінки ризику.

Для того, щоб забезпечити простоту та припустиму швидкодію було вирішено використати асинхронний веб-фреймворк FastApi, відомий своєю мінімалістичністю, легкістю та лаконічністю. Його асинхронна природа дозволяє йому без додаткових налаштувань опрацьовувати кілька запитів одночасно при роботі в один процес. Він вже має гарну документацію та інтеграцію з популярним ASGI сервером uvicorn.

Серед головних викликів, які постали під час розробки були:

1. Забезпечення ідентичності препроцесингу даних між даними на продакшені та при тренуванні. Адже якщо модель отримає дані, яких не бачила під час тренування – в найкращому випадку модель недооцінить чи переоцінить ризик. В найгіршому – відбудеться помилка, а застосунок припинить роботу.
2. Узгодженість між українським інтерфейсом та англійськими тренувальними даними. Необхідно було зібрати набір питань, за допомогою яких збирались оригінальні початкові дані та перекладаючи на українську забезпечити збереження максимум контексту, який мали респонденти, що формували початкові дані.
3. Забезпечити інтерпритованість результату. Кінцева модель видає «скор» ризику у межах від 0 до 1. Але насправді це не є вірогідністю настання серцевої хвороби. В реальності це «впевненість» моделі у тому, що серцева хвороба настане. А вірогідність можливо дізнатись лише переглянувши рівень precision та recall виданого скору на тестових даних.

Для перенесення моделі та кроків препроцесингу використовувалась бібліотека серіалізації dill з рекурсивною параметризацією. Особливість цього інструменту від подібних – здатність зберігати разом з функціями усі їх зовнішні замикання та залежності в автоматичному режимі. Це дозволяє

переносити дані зі збереженням максимум контекста початкового оточення, уникаючи помилок та небажаної поведінки.

В межах аплі було реалізовано 2 роути:

1. Кореневий, який представляє базовий домен застосунку та є його основою. На ньому розміщено опитування та увесь інтерфейс, з яким буде напряду взаємодіяти користувач
2. Обрахунку ризику. Виконує обрахунок ризик на основі поведінкових даних після заповнення користувачем опитування та натиснення відповідної кнопки. Тут фактично відбувається весь бекенд та внутрішні розрахунки моделі

Для легкої підтримуваності та масштабованості код було розбито на перелік вузько-спеціалізованих функцій, які поділились на 3 спеціалізації:

1. Створення форми опитування, яку бачить користувач
2. Декодування вхідного запису, який є готовою формою опитування
3. Повний цикл препроцесингу форми опитування з доступну для моделі форму
4. Обрахунку ризику закодованого запису шляхом виклику відповідного методу моделі
5. Категоризація скору по 5 можливим зонам ризику: дуже низький, помірний, середній, високий та дуже високий та його повернення користувачу.

3.2. Пояснення логіки функціонування додатку

Для створення форми було використано гнучкий підхід з динамічним формуванням на основі заделегідь визначеного .json словника усіх питань українською версією. Який, для узгодженості містить також мапу внутрішніх представлень англійською мовою ознак, які представляють дані питання. Зі словника легко видалити питання або додати нове, без необхідності переписувати будь-який код у застосунку.

Самі поля форми обмежують можливий ввід користувача, задаючи зазделегідь максимальні та мінімальні можливі значення «чутливих» полів. Наприклад, людина не може спати більше 24 годин на добу або важити менше нуля кілограм.

Декодування вхідного запису відбувається за допомогою зазделегідь визначеного словника, який містить мапу відповідності українських відповідей до їх англійських версій, які точно наслідують оригінальним даним. При надходженні готової форми сполучаються представлення ознак англійською мовою словника питань та англійській маппінг відповідей словника декодування.

Повний цикл препроцесингу відбувається через виклик того самого препроцесора, який було використано під час тренування моделі.

Категоризація скору ризику відбувається за набором зазделегідь визначених трешхолдів. Трешхолди визначались дещо спорадично, але головною мірою була збережена закономірність – чим вище показник precision та нижче показник recall, тим вище буде зона.

Повний код роутів у Додаток В.

Головна ідея залишається наступною – чим точніше стає модель на визначеному скорі, тим вищим є зона ризику. Найвищий трешхолд, наприклад, відповідає всього 1% покриття та 80% точності (рис. 3.1).

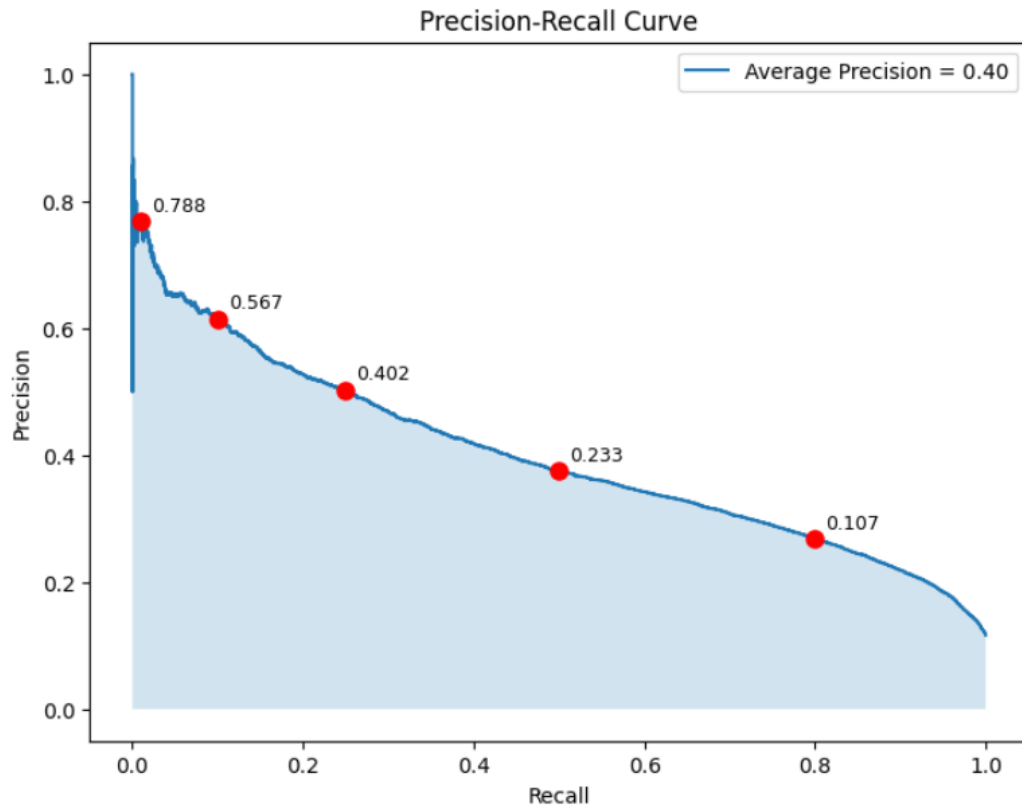


Рис. 3.1 - Графік залежності трешхолів від збільшення показника precision та зменшення показник recall

Таким чином (рис. 3.2) було вирішено головні поставлені задачі та дотримано вимоги до функціонування застосунка.

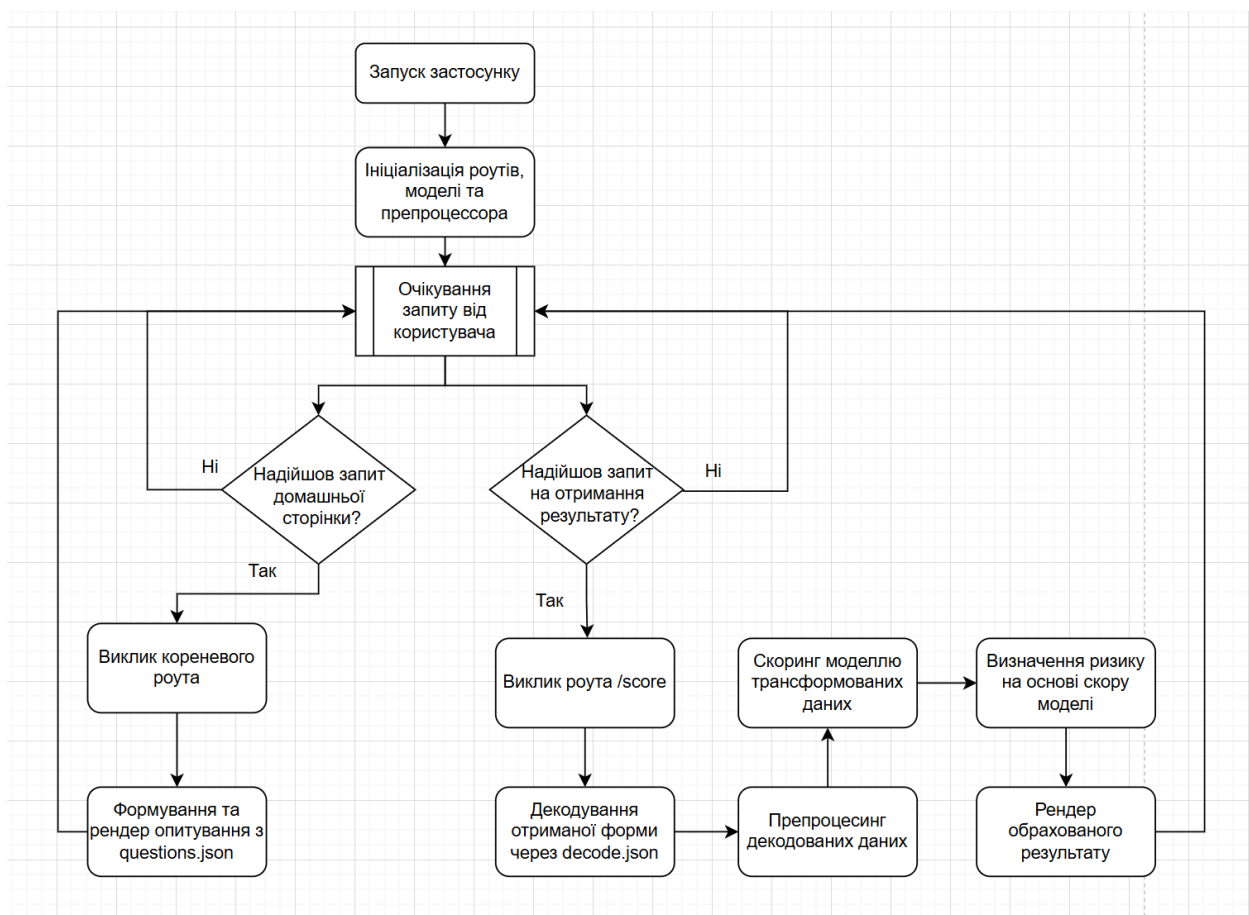


Рис. 3.2 - Повний цикл обробки запиту користувача

3.3. ТЕСТУВАННЯ ТА РОЗГОРТАННЯ ДОДАТКУ

Для автоматизованого тестування було використано selenium. Він надає простий та гнучкий інтерфейс для тестування веб-додатків, має відкритий апі та гарну документацію. Хоча його швидкодія може бути його недоліком, сама задача не має великого простору можливих значень для дослідження, тому це не є проблемою.

У ході тестування скрипт випадково визначав які поля буде заповнено, значення самих полів теж визначались випадковим чином. Скрипт було повторено 1 тисячу раз, з логуванням помилок. За фактом завершення повного циклу тестування помилок у роботі веб-додатку не було виявлено.

Апі коректно обробив усі 1 тисячу запитів з різними параметрами. Повний код скрипту тестування у Додаток 4.

Для розгортання застосунку було обрано Render за свою простоту та наявність безкоштовних планів для хоббі проектів. Також сервіс надає власні безкоштовні доменні ім'я. Розгортання проводить відповідно до заданої docker конфігурації у репозиторії проекту, який повинен розташовуватись на github.

Отже, пройшовши повний цикл реєстрації, створивши файл конфігурації докер та завантаживши проект на github репозиторій залишилось лише вказати його адресу всередині render, після чого застосунок було автоматично публічно розгорнуто.

Застосунок (рис. 3.3) повинен функціонувати за посиланням <https://hap-webapp.onrender.com/>

hap-webapp.onrender.com

Оцінка ризику кардіо-васкулярних захворювань

Яка ваша стать?

— Оберіть варіант —

До якої вікової категорії ви належите?

— Оберіть варіант —

Який ваш зріст (у метрах)?

Яка ваша вага (у кілограмах)?

Скільки годин на добу ви спите в середньому?

Скільки днів протягом останнього місяця ви переживали психологічні труднощі?

Про проєкт

Цей проєкт використовує методи машинного навчання для оцінки ризику серцевого нападу на основі поведінкових даних. Він націлений на швидку оцінку вашого способу життя та ризик-групи. Вам не потрібні аналізи чи медична освіта, лише кілька хвилин на заповнення опитування. Всього існує 5 рівнів ризику - низький, помірний, середній, високий та дуже високий.

Оцінка, яку Ви отримаєте, не може бути гарантією відсутності ризику, так як під час дослідження даних я знайшов, що кардіо-васкулярні хвороби трапляються у всіх групах ризику. Тим не менш, чим вище рівень ризику - тим частіше трапляються серцеві напади, інсульти та інші хвороби серця.

Про модель

У якості моделі використовується логістична регресія newton-cholesky варіації, разом з поглибленим інжинірингом фіч. Модель заснована на даних з опитування CDC, що є частиною Behavioral Risk Factor Surveillance System (BRFSS). Це щорічне телефонне опитування, яке проводиться для збору даних про стан здоров'я громадян США. База даних BRFSS є найбільшою безперервно проведеною системою опитувань у світі. Щорічно опитуванням охоплюється понад 400 000 дорослих.

В якості цільової змінної використовувалась наявність серцевого нападу, інсульту чи болей в грудях

Рис. 3.3 - Інтерфейс розгорнутого застосунка

Висновки до розділу 3.

Було виконано розробка, тестування та публічне розгортання застосунка, який інтегрує раніше побудовану модель оцінки ризику хвороб серця у зручний та зрозумілий для українського користувача веб-інтерфейс.

У ході виконання поставленого завдання було використано в якості технологічного стеку FastApi, Uvicorn, Selenium, Docker, Render.

Було вирішено виклики забезпечення узгодженості даних на тренуванні та продакшені, забезпечення ідентичності контексту питань з контекстом оригінальних респондентів та інтерпритації кінцевих результатів обрахунку ризику, розділивши увесь простір трешхолдів на 5 зон.

ВИСНОВКИ

В межах аналізу трендів, проблем та можливостей було з'ясовано, що хвороби серця є головною причиною смертності світі, займаючи третину від усіх смертей. Тренд не показує тенденції до спадання, навпаки, скоріше проблема набирає обертів з часом, зумовлена зміною образу життя людей.

Наслідки ж захворювань серця є багатограними та існують на самих різних рівнях масштабів. Вони мають значний економічний ефект на країни всього світу, збільшуючи видатки на охорону здоров'я та зменшуючи продуктивність праці. Окрім цього, що важливіше, вони мають значний психологічний та фізичний вплив на людину. У випадку виживання після факту серцевої хвороби людина решту життя знаходиться у підвищеній зоні ризику, що веде до значно частіших депресивних та тривожних станів. І все одно велика частина людей після першого інсульту чи серцевого нападу помирають протягом п'яти років від рецидиву або пов'язаних факторів.

Хоча хвороби серця добре піддаються профілактиці, обізнаність людей про головні ризик-фактори залишається низькою навіть у країнах розвиненої системи охорони здоров'я.

У ході дослідження доступних публічних даних було з'ясовано, що вибір є дуже обмеженим у зв'язку з поведінковою специфікою вимог та конфіденційним характером медичних даних.

У ході дослідження природи даних вдалось підтвердити сильний зв'язок загальноприйнятих ризик-факторів з підвищеним ризиком серцевих хвороб. Також вдалось встановити, що хворіють на ССЗ навіть ті люди, які позитивно сприймають свій стан здоров'я та ведуть здоровий спосіб життя.

Під час препроцесингу даних було проведено дослідження доменного контексту для встановлення потенційних витоків. Було видалено потенційно витікаючі ознаки, що могли б знизити практику предективну здібність моделі.

У процес побудови моделей машинного навчання з'ясувалось що дані містять здебільшого прості, лінійні залежності, що спрощує інтерпритацію результатів роботи моделі.

Розроблюючи додаток було вирішено проблему перенесення моделі та препроцесингу в продакш середовище та забезпечення узгодженості даних на продакшен та тренуванні.

Врешті-решт застосунок було успішно протестовано та публічно розгорнуто. Тепер він доступний будь-кому з доступ в інтернет.

СПИСОК ДЖЕРЕЛ

1. WHO (World Health Organization), Cardiovascular diseases (CVDs), [Електроний ресурс] - [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)#:~:text=,income%20countries](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=,income%20countries)
2. Our World in Data, Cardiovascular Diseases, [Електроний ресурс] - <https://ourworldindata.org/cardiovascular-diseases#:~:text=In%202000%2C%20around%2014%20million,close%20to%2018%20million%20died>
3. WHO (World Health Organization), leading causes of death and disability worldwide: 2000-2019, [Електроний ресурс] - <https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019#:~:text=Heart%20disease%20has%20remained%20the,%5B1>
4. CDC (Center for disease control and prevention), Morbidity and Mortality Weekly Report, [Електроний ресурс] - <https://www.cdc.gov/mmwr/volumes/71/wr/mm7123a4.htm#:~:text=In%202020%2C%206.3,aged%2018%E2%80%9344%20or%20%E2%89%A565%20years>
5. CDC (Center for disease control and prevention), Heart Disease Facts, [Електроний ресурс] - <https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html#:~:text=In%20the%20United%20States%3A>
6. World Heart Federation, Cardiovascular disease (CVD) is the leading cause of death in women worldwide, [Електроний ресурс] - <https://world-heart-federation.org/what-we-do/women-cvd/#:~:text=Cardiovascular%20disease%20,of%20death%20in%20women%20worldwide>
7. Sonia S Anand¹, Shofiquel Islam, Annika Rosengren, Maria Grazia Franzosi, Krisela Steyn, Afzal Hussein Yusufali, Matyas Keltai, Rafael Diaz, Sumathy Rangarajan, Salim Yusuf; Risk factors for myocardial infarction in women and men: insights from the INTERHEART study

8. American College of Cardiology, Ultra-Processed Foods Are Breaking Your Heart, [Электронный ресурс] - <https://www.acc.org/About-ACC/Press-Releases/2021/03/22/18/46/Ultra-Processed-Foods-Are-Breaking-Your-Heart#:~:text=%28CVD%29,found%20to%20further%20increase%20risk>
9. WHO (World Health Organization), Global Status Report on Physical Activity 2022, [Электронный ресурс] - <https://www.who.int/teams/health-promotion/physical-activity/global-status-report-on-physical-activity-2022#:~:text=Regular%20physical%20activity%20promotes%20both,and%20society%20as%20a%20whole>
10. HCCI (Health Care Cost Institute), Spending on Preventive Services Represents a Small Fraction of Total Health Care Spending, but Costs to Individuals Could Be High without ACA Protection, [Электронный ресурс] - <https://healthcostinstitute.org/hcci-originals-dropdown/all-hcci-reports/spending-on-preventive-services-represents-a-small-fraction-of-total-health-care-spending-but-costs-to-individuals-could-be-high-without-aca-protection>
11. Donella Meadows, Thinking in Systems (2008)
12. Thomas K. Hillers, Gordon H. Guyatt, Neil Oldridge, Jean Crowe, Andrew Willan, Lauren Griffith, David Feeny; Quality of life after myocardial infarction
13. Ajar Kochar, Anita Y. Chen, Puza P. Sharma, Neha J. Pagidipati, Gregg C. Fonarow, Patricia A. Cowper, Matthew T. Roe, Eric D. Peterson and Tracy Y. Wang; Long-Term Mortality of Older Patients With Acute Myocardial Infarction Treated in US Clinical Practice
14. American Heart Federation, Life After a Heart Attack, [Электронный ресурс] - <https://www.heart.org/en/health-topics/heart-attack/life-after-a-heart-attack#:~:text=Life%20After%20a%20Heart%20Attack,of%20death%20within%20five%20years>
15. Redford B. Williams; Depression After Heart Attack: Why Should I Be Concerned About Depression After a Heart Attack?

16. World Heart Federation, WHY CIRCULATORY DISEASES MATTER...
[Инфографика] -
<https://world-heart-federation.org/wp-content/uploads/2021/04/Infographic-Why-Circulatory-Health-Matters.pdf#:~:text=3,percent%20of%20all%20CVD%20expenditure18>
17. Acemoglu D. & Robinson J. A. (2012); Why nations fail the origins of power prosperity and poverty
18. Richard Kahn, Rose Marie Robertson, Robert Smith, David Eddy; The Impact of Prevention on Reducing the Burden of Cardiovascular Disease
19. Institute of Medicine (US) Roundtable on Evidence-Based Medicine; Yong PL, Saunders RS, Olsen LA; The Healthcare Imperative: Lowering Costs and Improving Outcomes: Workshop Series Summary.
20. Amirah Alshakarah, Deema Muriyah; Awareness and Knowledge of Cardiovascular Diseases and Its Risk Factors Among Women of Reproductive Age: A Scoping Review
21. World Heart Federation; Cardiovascular disease (CVD) is the leading cause of death in women worldwide, [Электроний ресурс] -
<https://world-heart-federation.org/what-we-do/women-cvd/#:~:text=Cardiovascular%20disease%20is%20responsible%20for,more%20than%20it%20does%20women>
22. Abdo Ahmed AA, Mohammed Al-Shami A, Jamshed S, Fata Nahas AR, Mohamed Ibrahim MI. Public Awareness of and Action towards Heart Attack Symptoms: An Exploratory Study. Int J Environ Res Public Health. 2020 Dec 2;17(23):8982. doi: 10.3390/ijerph17238982. PMID: 33276636; PMCID: PMC7729488.
23. CDC, Awareness of Heart Attack Symptoms and Response Among Adults — United States, 2008, 2014, and 2017, [Электроний ресурс] -
<https://www.cdc.gov/mmwr/volumes/68/wr/mm6805a2.htm#:~:text=chance%20for%20survival%20%284%29,adults%20who%20knew%20to%20call>

24. Pronk NP, Anderson LH, Crain AL, Martinson BC, O'Connor PJ, Sherwood NE, Whitebird RR. Meeting recommendations for multiple healthy lifestyle factors. Prevalence, clustering, and predictors among adolescent, adult, and senior health plan members. *Am J Prev Med.* 2004 Aug;27(2 Suppl):25-33. doi: 10.1016/j.amepre.2004.04.022. PMID: 15275671.
25. Wolpert, David H. and William G. Macready. "No Free Lunch Theorems for Search." (1995).
26. National Kidney Foundation; The Heart and Kidney Connection, [Электроний ресурс] - <https://www.kidney.org/kidney-topics/heart-and-kidney-connection>
27. Harvard Medical School; Understanding COPD from a cardiovascular perspective, [Электроний ресурс] - <https://www.health.harvard.edu/heart-health/understanding-copd-from-a-cardiovascular-perspective>
28. Martijn W. Heymans, Jos W.R. Twisk, Handling missing data in clinical research,
29. Li, Cheng. (2013). Little's Test of Missing Completely at Random. *Stata Journal.* 13. 795-809. 10.1177/1536867X1301300407.
30. Pantanowitz, Adam & Marwala, Tshilidzi. (2009). Missing Data Imputation Through the Use of the Random Forest Algorithm
31. Lecun, Yann & Bottou, Leon & Orr, Genevieve & Müller, Klaus-Robert. (2000). Efficient BackProp.
32. National Institute of Mental Health, Major Depression, [Электроний ресурс] - <https://www.nimh.nih.gov/health/statistics/major-depression>
33. Jennifer N. Carpenter, Anthony W. Lynch, Survivorship bias and attrition effects in measures of performance persistence
34. C. Seger, 'An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing', Dissertation, 2018.

35. Daniele Micci-Barreca. 2001. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems.
36. Sanford Weisberg, Yeo-Johnson Power Transformations
37. Scikit-learn, MissingIndicator, [Електроний ресурс] - <https://scikit-learn.org/stable/modules/generated/sklearn.impute.MissingIndicator.html>
38. Andrew P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms
39. Yanminsun, & Wong, Andrew & Kamel, Mohamed S.. (2011). Classification of imbalanced data: a review
40. Daniel Berrar, Cross-validation
41. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu; LightGBM: A Highly Efficient Gradient Boosting Decision Tree
42. І.В Машкіна, ВО Абрамов, ТІ Носенко, ЄВ Іваніченко. Навчально-методичний посібник до виконання і захисту кваліфікаційної роботи бакалавра спеціальностей 122 Комп'ютерні науки для здобувачів вищої освіти денної форми навчання, Київський столичний університет імені Бориса Грінченка. Київ, 2024.- 43 с.
43. Теоретичні та практичні аспекти використання математичних методів та інформаційних технологій в освіті й науці: моногр. / за заг. ред. О. Литвин. — К.: Київ. ун-т ім. Б. Грінченка, 2021. — 332 с.
44. Aralova N. I., Mashkin V. I., Mashkina I. V. Athletes heart hypertrophy as result of long-term adaptation to loadings. Studying at mathematical model. Scientific achievements of modern society. Abstracts of the 5th International scientific and practical conference. Cognum Publishing House. Liverpool, United Kingdom. 2020, P. 286– 292. URL: <http://sci-conf.com.ua>

ДОДАТКИ

ДОДАТОК А

Повний код кастомного імпутера

```
class CustomImputer(BaseEstimator, TransformerMixin):
```

```
    def __init__(self):
```

```
        self.statistics_ = {}
```

```
        self.numeric_columns_ = None
```

```
        self.non_numeric_columns_ = None
```

```
    def fit(self, X, y=None):
```

```
        self.statistics_ = {}
```

```
self.numeric_columns_ = X.select_dtypes(include=["float32", "float64",
"float16"]).columns
```

```
self.non_numeric_columns_ = X.select_dtypes(exclude=["float32",
"float64", "float16"]).columns
```

```
for col in self.numeric_columns_:
    self.statistics_[col] = X[col].mean()
```

```
for col in self.non_numeric_columns_:
    mode_value = X[col].mode()
    self.statistics_[col] = mode_value.iloc[0] if not mode_value.empty else
```

None

```
return self
```

```
def transform(self, X):
```

```
numeric_data =
X[self.numeric_columns_].to_numpy(copy=True).astype(np.float32)
```

```
non_numeric_data = X[self.non_numeric_columns_]
```

```
if np.isnan(numeric_data).any():
```

```
    for i, col in enumerate(self.numeric_columns_):
```

```
        value = self.statistics_.get(col, None)
```

```
        mask = np.isnan(numeric_data[:, i])
```

```
        numeric_data[mask, i] = value
```

```
numeric_df = pd.DataFrame(numeric_data,
columns=self.numeric_columns_, index=X.index)
```

```
non_numeric_data = non_numeric_data.fillna(self.statistics_)
```

```

        result = pd.concat([numeric_df, non_numeric_data], axis=1)
        return result[self.numeric_columns_.tolist() +
self.non_numeric_columns_.tolist()]

```

```

def set_output(self, transform=None):
    return self

```

```

def __call__(self, X, y):
    return self.transform(X, y)

```

ДОДАТОК Б

Повний код препроцесора

```

def format_age(row, **kwargs):
    if type(row) == float:
        return row

    replacements = {
        'Age': '',
        'to': '-',
        'or older': '+',
        '': ''
    }

```

```
pattern = re.compile('|'.join(re.escape(key) for key in replacements.keys()))
```

```
formatted = pattern.sub(lambda match: replacements[match.group(0)], row)  
return formatted
```

```
class CustomOrdinalEncoder(BaseEstimator, TransformerMixin):
```

```
    def __init__(self, mappings):
```

```
        self.mappings = mappings
```

```
    def fit(self, X, y=None):
```

```
        self.mappings = {key:value for key, value in self.mappings.items() if key in  
X.columns.to_list()}
```

```
        self.inverse_mappings = {col: {v: k for k, v in mapping.items()} for col,  
mapping in self.mappings.items()}
```

```
        return self
```

```
    def transform(self, X):
```

```
        X_transformed = X.copy()
```

```
        for col, mapping in self.mappings.items():
```

```
            X_transformed[col] = X_transformed[col].map(mapping)
```

```
        return X_transformed
```

```
    def inverse_transform(self, X):
```

```
        X_inverse = X.copy()
```

```
        for col, inverse_mapping in self.inverse_mappings.items():
```

```
            X_inverse[col] = X_inverse[col].map(inverse_mapping)
```

```
        return X_inverse
```

```
    def get_feature_names_out(self, input_features=None):
```

```
        return input_features if input_features is not None else  
list(self.mappings.keys())
```

```
class Preprocessor:
```

```
    def __init__(self, is_bin_features=False):  
        self.is_bin_features=is_bin_features  
        self.bimodal_features = ["MentalHealthDays", "PhysicalHealthDays"]  
        self.bins = [0, 3, 7, 14, 29, float("inf")]  
        self.labels = ["0-3", "4-7", "8-14", "15-29", "30"]  
        self.ordinal_mappings = {  
            "GeneralHealth": {  
                'Poor': 0,  
                'Fair': 1,  
                'Good': 2,  
                'Very good': 3,  
                'Excellent': 4,  
            },  
            "AgeCategory":{  
                '80+': 82.5,  
                '75-79': 77,  
                '70-74': 72,  
                '65-69': 67,  
                '60-64': 62,  
                '55-59': 57,  
                '50-54': 52,  
                '45-49': 46.5,  
                '40-44': 42,  
                '35-39': 37,  
                '30-34': 32,
```

```

        '25-29': 27,
        '18-24': 21,
    },
    "RemovedTeeth": {
        'None of them': 0,
        '1 to 5': 1,
        '6 or more, but not all': 2,
        'All': 3
    },
    'LastCheckupTime': {
        'Within past year (anytime less than 12 months ago)': 0,
        'Within past 2 years (1 year but less than 2 years ago)': 1,
        'Within past 5 years (2 years but less than 5 years ago)': 2,
        '5 or more years ago': 3
    }
}

```

```

def bin_features(self, X):
    for feature in self.bimodal_features:
        X[feature] = pd.cut(X[feature], bins=self.bins, labels=self.labels,
right=False)

    def create_transformers(self, num_features, binary_features,
multiclass_features, ordinal_features, is_poly=False):
        transformers = []

        num_steps = [("imputer", CustomImputer())]
        if is_poly:
            num_steps.append(("poly", PolynomialFeatures(degree=2)))

```

```

num_steps.append(("log", FunctionTransformer(lambda x: np.log1p(x))))
num_steps.append(("scaler", MinMaxScaler()))
num_pipe = Pipeline(num_steps)

transformers.append(("num", num_pipe, num_features))
binary_pipe = Pipeline([
    ("imputer", CustomImputer()),
    ("encoder", OneHotEncoder(sparse_output=False))
])
transformers.append(("binary", binary_pipe, binary_features))
multiclass_pipe = Pipeline([
    ("imputer", CustomImputer()),
    ("encoder", OneHotEncoder(sparse_output=False))
])
if ordinal_features:
    multiclass_features = [feature for feature in multiclass_features if feature
not in ordinal_features]

transformers.append(("multiclass", multiclass_pipe, multiclass_features))
if ordinal_features:
    ordinal_pipe = Pipeline([
                                                                    # ("imputer",
IterativeImputer(estimator=RandomForestClassifier(random_state=42),
initial_strategy='most_frequent', random_state=42)),
    ("imputer", CustomImputer()),
    ("encoder", CustomOrdinalEncoder(self.ordinal_mappings))
])
transformers.append(("ordinal", ordinal_pipe, ordinal_features))

```

```
self.preprocessor = ColumnTransformer(transformers, remainder='drop',
verbose=True)
```

```
# Використання pandas на виході
```

```
self.preprocessor.set_output(transform='pandas')
```

```
def fit(self, X, y):
```

```
    if self.is_bin_features:
```

```
        X = self.bin_features(X)
```

```
        print("binned")
```

```
    X["AgeCategory"] = X["AgeCategory"].apply(format_age, axis=1)
```

```
self.preprocessor.fit(X, y)
```

```
def transform(self, X, y, add_missing_indicators=False):
```

```
    if self.is_bin_features:
```

```
        X = self.bin_features(X)
```

```
    if add_missing_indicators:
```

```
        missing_indicator_tf = MissingIndicator(sparse=False).fit(X)
```

```
        missing_indicator_tf.set_output(transform="pandas")
```

```
        missing_indicators_df = missing_indicator_tf.transform(X)
```

```
X["AgeCategory"] = X["AgeCategory"].apply(format_age, axis=1)
```

```
X = self.preprocessor.transform(X)
```

```
    X.columns = X.columns.str.replace(r'^\w', '_', regex=True) #remove
special symbols like spaces
```

```
    y = y.map({"Yes": 1, "No": 0}).fillna(y)
```

```
    if add_missing_indicators:
```

```
        X = pd.concat([X, missing_indicators_df], axis=1)
```

```
    return X, y
```

ДОДАТОК В

Повний код app.py роутів

```
from fastapi import FastAPI, Request
from fastapi.responses import HTMLResponse, JSONResponse
from fastapi.staticfiles import StaticFiles
from fastapi.templating import Jinja2Templates
import uvicorn
import json
from utils.preprocessing import get_preprocessor
from models.model import get_model
from utils.process_request import validate_request, decode_request
from utils.utils import get_risk_level
```

```

app = FastAPI()
app.mount("/static", StaticFiles(directory="static"), name="static")
templates = Jinja2Templates(directory="templates")

with open("data/questions.json", "r", encoding="utf-8") as f:
    questions = json.load(f)

preprocessor = get_preprocessor()
model = get_model()

@app.get("/", response_class=HTMLResponse)
async def get_risk(request: Request):
    return templates.TemplateResponse("risk.html", {"request": request,
"questions": questions, "result": None, "errors": {}, "form_data": {}})

@app.post("/api/risk")
async def api_risk(request: Request):
    form = await request.form()
    valid_res = validate_request(form)
    if len(valid_res['errors']) > 0:
        return JSONResponse({"errors": valid_res["errors"]})
    request_df = decode_request(form)
    request_enc = preprocessor.transform(request_df)
    risk = model.predict_proba(request_enc)[0][1]
    level, color = get_risk_level(risk)
    return JSONResponse({ "level": level, "color": color})

if __name__ == "__main__":

```

```
uvicorn.run(app, host="127.0.0.1", port=8001)
```

ДОДАТОК Г

Повний код скрипта тестування

```
import time
import random
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC

def random_number_for_field(field_id):
    if field_id == "HeightInMeters":
        return round(random.uniform(1.3, 2.2), 2)
    elif field_id == "WeightInKilograms":
```

```

        return random.randint(50, 120)
    elif field_id == "SleepHours":
        return random.randint(4, 10)
    elif field_id == "MentalHealthDays":
        return random.randint(0, 31)
    else:
        return random.randint(1, 10) # 3

def main():
    driver = webdriver.Chrome()

    try:
        url = "http://127.0.0.1:8200"
        driver.get(url)

        time.sleep(2)

        form_inputs = driver.find_elements(By.TAG_NAME, "input")
        form_selects = driver.find_elements(By.TAG_NAME, "select")

        for inp in form_inputs:
            input_type = inp.get_attribute("type")
            field_id = inp.get_attribute("id")

            if input_type == "number":
                val = random_number_for_field(field_id)
                inp.clear()
                inp.send_keys(str(val))

```

```

elif input_type == "checkbox":
    is_checked = inp.is_selected()
    want_to_check = random.choice([True, False])
    if is_checked != want_to_check:
        inp.click()

for sel in form_selects:
    options = sel.find_elements(By.TAG_NAME, "option")
    if len(options) > 1:
        chosen_option = random.choice(options[1:])
        chosen_option.click()

        submit_button = driver.find_element(By.XPATH,
        "//button[@type='submit']")
        submit_button.click()

try:
    wait = WebDriverWait(driver, 10)
    result_block = wait.until(
        EC.presence_of_element_located((By.CLASS_NAME, "result-block"))
    )
    print("Знайшли блок з результатом")
except:
    print("Не дочекалися результату")

time.sleep(5)

finally:
    driver.quit()

```

```
if __name__ == "__main__":  
    main()
```