Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів заповнення пропущених значень в медичних вибірках даних.
(тема)

Виконав: студент 2 курсу, групи СШІм-18-2
Рябой Д.В.
(прізвище, ініціали)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо- наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва спеціалізації)

Керівник проф. Кулішова Н.Є.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2020 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва)

Тип програми освітньо -наукова
(освітньо-проіфесійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Рябому Денису Вікторовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів заповнення пропущених значень в медичних вибірках даних

затверджена наказом по університету від 30 березня 2020 р. № 480 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 19 травня 2020 р.

3. Вихідні дані до роботи

3.1 Досліджувані методи: ЕМ-оцінювання, регресійне моделювання, метод на основі часткових відстаней

3.2 Кількість медичних вибірок даних з репозиторію – 3, клінічні дані – 1

3.3 Тип даних безперервні, бінарні, рангові

3.4 Загальна кількість пропущених значень 645

3.5 Середовище моделювання: Python 3.7 на базі Jupiter Notebook

4. Перелік питань, що потрібно опрацювати в роботі

4.1. Аналіз предметної галузі та постановка завдання дослідження

4.2. Методи заповнення пропущених значень в даних

4.3 Апробація методів заповнення пропущених значень на медичних даних

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	проф. Кулішова Н.Є.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Отримання завдання на атестаційну роботу	30.03.2020	виконано
2.	Аналіз предметної галузі і постановка		
	завдання дослідження	01.04.2020 – 07.04.2020	виконано
3.	Огляд методів заповнення пропусків	08.04.2020 – 15.04.2020	виконано
4.	Аналіз існуючих проблем даної предметної галузі	16.04.2020 – 22.04.2020	виконано
5.	Розробка та апробація методу заповнення пропусків		
	на основі частковий відстаней	23.04.2020 – 04.05.2020	виконано
6.	Оформлення пояснювальної записки	05.05.2020 – 12.05.2020	виконано
7.	Оформлення графічних матеріалів	13.05.2020	виконано
8.	Попередній захист	16.05.2020	виконано
9.	Захист перед ЕК	19.05.2020	виконано

Дата видачі завдання 30 березня 2020 р.

Студент _____
(підпис)

Керівник роботи _____ ___проф. Кулішова Н.Є._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Записка пояснювальна: 71 с., 20 рис., 10 табл., 2 дод., 43 джерела.

ВІДСТАНЬ ЧАСТКОВА, ВИБІРКА МЕДИЧНА, ЗАПОВНЕННЯ ПРОПУСКУ, ЗНАЧЕННЯ ПРОПУЩЕНЕ, СТУПІНЬ НАЛЕЖНОСТІ

Об'єкт дослідження – процес заповнення пропущених значень в медичних вибірках даних.

Предмет дослідження – дані пульмонологічних досліджень.

Мета роботи – аналіз методів заповнення пропущених значень та вибір тих, що є доцільним застосовувати в медичних вибірках.

Методи дослідження – EM-оцінювання, регресійний аналіз, метод, що базується на часткових відстанях.

У роботі розглянуло аналіз області прийняття рішень в медицині, розглянуто методи заповнення пропущених значень, виділено переваги та недоліки кожного з них. Обрано методи, які можна застосовувати для заповнення пропущених значень в медичних вибірках даних, проведено апробацію таких методів на даних з репозиторію та клінічних медичних даних.

Результатом роботи є вибір та апробація методів заповнення пропущених значень для медичних вибірок даних, які можуть знайти застосування для різноманітних завдань інтелектуального аналізу медичних даних.

РЕФЕРАТ

Пояснительная записка: 71 с., 20 рис., 10 табл., 2 прил., 43 источника.

ВЫБОРКА МЕДИЦИНСКАЯ, ЗАПОЛНЕНИЕ ПРОПУСКОВ, ЗНАЧЕНИЕ ОТСУТСТВУЮЩЕЕ, РАССТОЯНИЕ ЧАСТИЧНОЕ, СТЕПЕНЬ ПРИНАДЛЕЖНОСТИ

Объект исследования – процесс заполнения пропущенных значений в медицинских выборках данных.

Предмет исследования – данные пульмонологических исследований.

Цель работы – анализ методов заполнения пропущенных значений и выбор таких, которые целесообразно применять в медицинских выборках.

Методы исследования – ЭМ-оценивание, регрессионный анализ, метод, основанный на частичных расстояниях.

В работе рассмотрен анализ области принятия решений в медицине, рассмотрены методы заполнения пропущенных значений, выделены преимущества и недостатки каждого из них. Выбраны методы, которые можно применять для заполнения пропущенных значений в медицинских выборках данных, проведена апробация таких методов на данных из репозитория и клинических медицинских данных.

Результатом работы является выбор и апробация методов заполнения пропущенных значений для медицинских выборок данных, которые могут найти применение для различных задач интеллектуального анализа медицинских данных.

ABSTRACT

Explanatory note: 71 p., 20 fig., 10 tabl., 2 ann.,43 sources.

DEGREE OF MEMBERSHIP, FILL IN GAPS, MEDICAL DATASET, MISSING VALUE, PARTIAL DISTANCE

The object of study – the process of filling in the missing values in medical datasets.

The subject of research – the data of pulmonary research.

The purpose of the work is to analyze the methods of filling in the missing values and selecting those that are appropriate to use in medical datasets.

Research methods – EM-estimation, regression analysis, method based on partial distances.

The paper considers the analysis of the field of decision-making in medicine, considers the methods of filling in the missing values, highlights the advantages and disadvantages of each of them. Methods that can be used to fill in the missing values in medical data samples have been selected, and such methods have been tested on repository data and clinical medical data.

The result is the selection and testing of methods for filling in the missing values for medical data samples, which can be used for various tasks of medical data mining.

ЗМІСТ

ВСТУП

Основним завданням медичної діагностики є визначення можливих діагнозів хворого на основі знань предметної області та даних його обстеження. До них відносяться значення ознак (в моменти їхнього спостереження), значення анатомо-фізіологічних особливостей (постійні в часі) і значення подій, що відбулися (в моменти, коли вони відбувалися).

І в медицині, і в техніці на початковому етапі розвитку все завдання вирішував один чоловік. Його зір і слух були основними вимірювальними інструментами, а досвід визначав якість діагнозу. У міру розвитку засобів вимірювань функції вимірювання передавалися приладам і обслуговуючому персоналу, які, як правило, не мали спеціальної діагностичної підготовки. Постановку діагнозу з урахуванням результатів виконаних вимірів продовжував здійснювати фахівець з фундаментальної діагностичної підготовкою, який має, при необхідності, додаткову спеціалізацію за певним видом органів або систем організму. В даний час в медицині чітко проявляється тенденція об'єднання завдань вимірювання і постановки діагнозу вже на базі автоматизованої системи комплексної діагностики, яка не тільки виконує вимірювання, а й ставить діагноз.

Завдання постановки вірного діагнозу в медицині ускладнюється наявністю великої кількості пропущених значень в медичних вибірках даних. Причиною цьому є те, що:

– збір інформації про пацієнтів, що мають різний початковий діагноз відбувається за різними протоколами та стандартами;

– оскільки лабораторні бази багатьох лікарень України мають обмежену кількість обладнання, тому деякі дослідження пацієнтові або не були проведені зовсім, або відкладаються на певний час;

– пацієнти, що потрапляють в ургентному порядку (швидка допомога) або у стані втрати свідомості можуть не мати частини даних, що стосується анамнезу та суб’єктивного опису стану.

І тому частина пацієнтів можуть мати певну кількість пропущених значень. Тому для аналізу таких даних необхідно або використовувати методи, що можуть обробляти дані з пропусками (а таких небагато) або проводити заповнення пропущених значень на етапі попереднього оброблення даних. Серед існуючих методів заповнення пропущених значень використовуються прості процедури (заповнення середнім, даними попереднього пацієнта у вибірці даних, тощо) та складні підходи (використання для заповнення інформації про взаємозв’язки у даних). Таким чином, метою роботи є дослідження різних методів заповнення пропущених значень в медичних вибірках даних.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАВДАННЯ ДОСЛІДЖЕННЯ

1.1 Завдання медичної діагностики

Виходячи з розглянутої задачі, можна виділити наступні труднощі, з якими стикаються лікарі на етапі проведення медичної діагностики:

– неоднозначність діагнозу, тобто наявність супутніх захворювань, які сильно викривлюють картину симптомів;

– часто, невиправдано велика кількість показників, які вимірюються з кожної людини, що також призводить до викривлення даних;

– масові обстеження, тобто відсутність індивідуального підходу до кожного пацієнта, а, як відомо, значення чутливості і специфічності для кожної людини необхідно підбирати персонально;

– наявність пропущених показників у даних пацієнта з причин невиконання деяких вимірювань;

– лікування захворювання вже на етапі його прояви і розвитку, а не на етапі профілактики його появи;

– кількість існуючих захворювань більш ніж на порядок більше, ніж кількість методів їх діагностики, що відповідно підвищує ризик неправильної постановки діагнозу;

– відсутність статистики по рідкісних форм захворювання;

– суб'єктивізм лікаря, оскільки саме лікар ставить остаточний діагноз і, як показує практика, більше половини всіх лікарських помилок відбувається за рахунок цього фактору.

На сьогодні лікарська помилка трактується як неправильні дії або бездіяльність лікаря при виконанні ним своїх професійних обов'язків, які не є наслідком його недобросовісності і не містять складу злочину або ознак проступку (сформульовано академіком Давидовським) [1].

Лікарські помилки можуть бути декількох видів:

1. діагностичні, тобто пов'язані з постановкою діагнозу;
2. лікувально-тактичні, сюди входять помилки у виборі методів дослідження і в оцінці їх результатів;
3. лікувально-технічні, це неповне обстеження хворого і помилки діагностичних або лікувальних маніпуляцій;
4. організаційні, сюди включаються неправильна організація робочого місця і лікувального процесу;
5. помилки ведення медичної документації;
6. помилки поведінки медичного персоналу.

Причини лікарських помилок можна розділити на дві великі групи:

– об'єктивні (мінливість медичних концепцій, проблеми з технічним забезпеченням, організаційні труднощі);

– суб'єктивні (на думку багатьох дослідників, на частку суб'єктивного фактору припадає 60-70% причин діагностичних помилок). Це низький професіоналізм лікаря, недосконаліть знань лікаря, недотримання принципів деонтології.

За спостереженням патологоанатомів, причин великої кількості лікарських помилок кілька. На першому місці, незважаючи на всі досягнення сучасної медицини, як і раніше знаходяться труднощі діагностики. Правда, в 50% випадків неправильний діагноз лікарі ставлять з об'єктивних причин. До таких причин належать нетиповий перебіг хвороби і смертність, коли пацієнт помирає раніше, ніж його встигають обстежити. А ось решта 50% випадків – це як раз ті самі суб'єктивні причини, коли можна говорити безпосередньо про провину лікаря. До них відносяться: недооблік клінічних даних, коли лікар недооцінив тяжкість стану хворого (7-16%), недостатність обстеження (20-22%) і неправильне трактування результатів лабораторних досліджень.

Невтішні дані статистики говорять про масштабність проблеми. З лікарськими помилками за даними Канадського інституту медичної

інформації стикаються в середньому 187000 чоловік, причому від 9250 до 23750 з них гинуть (таким чином, частка летальних помилок 4,9-12,7%).

Департамент охорони здоров'я Великобританії в доповіді за 2000 рік зазначив, що несприятливі наслідки медичних втручань склали 850 тис. випадків і стали причиною 10% всіх госпіталізацій. Дослідження якості медичної допомоги в Австралії в 1995 році показало, що несприятливі побічні ефекти мали місце у 16,6% госпіталізованих хворих. Подібне становище спостерігається і в інших розвинених країнах. Так, у Франції в 1997 році близько 10% всіх госпіталізованих становили хворі з побічними реакціями.

Згідно з даними Інституту медицини Національної академії наук США (1999), в структурі причин смерті лікарські помилки посідають п'яте місце, випередивши такі поширені захворювання, як цукровий діабет, пневмонію, хвороба Альцгеймера і ниркову недостатність. Від помилок, що допускаються медичним персоналом, в Сполучених Штатах щорічно гине до 195 тисяч чоловік. Таку цифру оприлюднили дослідники з приватної корпорації HealthGrade. Згідно з їхніми даними, в 2000-2002 рр., щороку на 37 млн. госпіталізацій реєструвалося в середньому 1,14 мільйона лікарських помилок, з тієї чи іншої причини створювали загрозу для здоров'я і життя пацієнтів. З них 15-20% приводили до загибелі хворих.

На превеликий жаль, офіційної статистики з цієї проблеми в нашій країні немає, і найближчим часом навряд чи буде. В першу чергу, це пов'язано з тим, що структура, яка перевіряє факти лікарських помилок є частиною медичної системи України. Таким чином, в з'ясуванні реальної ситуації ніхто не зацікавлений. Хоча, проводячи прості аналогії з розвиненими країнами і з огляду на стан вітчизняної медицини, можна уявити, що цифри будуть жахливими.

А ось за неофіційними даними кожен третій діагноз ставиться вітчизняними лікарями невірно. Ця цифра не враховує тих хворих, яким медичний діагноз був поставлений не в повному обсязі. Для порівняння, в

США відсоток лікарських помилок становить 3-4%, у Великобританії – 5%, у Франції – 3% [0-0]. Через неправильний або несвоєчасно поставлений діагноз у нас вмирають 12% хворих на пневмонію. Серед розвинених країн країни СНД на першому місці за кількістю інсультів, оскільки погано поставлений лікарський контроль перебігу артеріальної гіпертонії. Дуже низький рівень надання медичної допомоги печінковим хворим. Через нестачу знань лікарі первинної ланки виявляють лише близько 30% хворих, які потребують високотехнологічної допомоги.

Принципово важливим для практичного рішення про оцінку стану здоров'я людини є питання про межу між нормою і патологією, тобто про виникнення донозологічних змін, які отримали назву «примежовий» стан [5]. Діагностика примежових станів в даний час ведеться практикуючими лікарями для виявлення можливих ускладнень стану пацієнтів [6-9], оскільки такі донозологічні зміни при недостатньому контролі можуть перерости в серйозні захворювання. Цим пояснюється підвищений інтерес сучасної медицини до профілів патології і груп ризику. Відповідно для діагностики примежових станів слід досліджувати методики прийняття рішення експертом-лікарем і при підборі методу обробки вибірок медико-біологічних даних враховувати особливості такого типу даних.

1.2 Порівняльна оцінка методик формування діагнозу експертом-лікарем

Діагностика грає в медицині найважливішу роль, і постановка діагнозу вимагає від лікаря великого майстерності, знань і інтуїції. Точність діагнозу і швидкість, з якою його можна поставити, залежать від дуже багатьох чинників: від стану хворого, від наявних даних про симптоми і ознаки захворювання і результатах лабораторних аналізів, від загального обсягу медичної інформації про спостереження таких симптомів при самих різних захворюваннях і, нарешті , від кваліфікації самого лікаря. Своєчасно

поставлений точний діагноз часто полегшує вибір методу лікування і значно підвищує ймовірність одужання хворого.

Виходячи з усіх цих міркувань, цілком природно спробувати визначити умови, при яких діагноз може бути поставлений максимально швидко і точно. Протягом багатьох століть лікарі зі змінним успіхом робили спроби вирішити це завдання. Однак в останні роки завдяки застосуванню сучасних методів лікування і діагностики, заснованих на новітніх досягненнях науки і техніки, можливості отримання успішних результатів значно зросли. Тому важливо знайти точні методи опису, дослідження, оцінки і контролю процесу постановки діагнозу. Оскільки зазвичай лікарі мають справу з великим числом взаємозалежних чинників, то необхідно використовувати комп'ютер, щоб шукані результати можна було отримати за досить короткий час. Такий підхід звільняє лікаря від необхідності займатися такими проблемами, які можна сформулювати в чисельній і логічній формі. Наявні в даний час дані свідчать про те, що обчислювальні машини грають важливу роль при постановці діагнозу [5, 13].

Слід вказати перспективні напрямки використання обчислювальної техніки в діагностиці. Це, перш за все:

– постановка диференціального діагнозу в відповідних умовах;

– оцінка точності діагнозів, які ставлять лікарі, з метою підвищення загального рівня діагностики;

– збір, узагальнення і, найголовніше, попередня обробка клінічних даних для кваліфікованого використання їх лікарями при постановці діагнозу.

Розробка методів діагностики за допомогою сучасних інформаційних технологій в нашій країні поки що не отримала досить великого поширення, і основною причиною цього є недостатнє оснащення лікувальних установ. Незважаючи на це, дослідниками вже отримані дуже обнадійливі результати, і подальші дослідження в цій області слід вважати вельми перспективними [14-18].

Зрозуміло, концентрація уваги на постановці диференціального діагнозу є у багатьох відношеннях надмірно спрощеним та обмеженим підходом до проблеми в цілому. Зазвичай передбачається, що всі альтернативні діагнози, з яких потрібно вибрати один, чітко і однозначно визначені. Однак на практиці виникає ряд проблем, пов'язаних з тим, що, по-перше, думки фахівців про найкращі способи класифікації хвороб часто розходяться, і нові дані можуть зажадати перегляду існуючих схем. По-друге, крім наявності одного, нерідко виявляються і супутні йому захворювання, які досить сильно «збурюють» картину симптомів і системи, що працюють за цим принципом, встановлять помилковий діагноз.

З усього вищесказаного можна зробити висновок, що формування діагнозу відбувається за кілька етапів, які залежать від тих технічних засобів, які використовує лікар для постановки діагнозу.

У тій ситуації, коли лікар не використовує ніяких технічних засобів, формування діагнозу відбувається за простою схемою, яка в більшості випадків використовується в лікарських кабінетах нашої країни (рис.1.1).

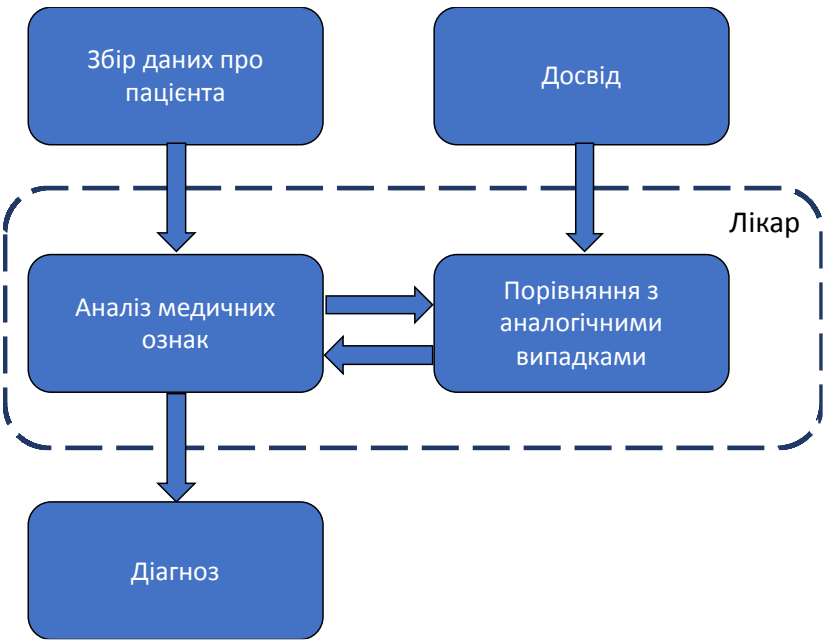

Рисунок 1.1 – Схема формування діагнозу без використання обчислювальних методів

Набір даних про пацієнтів складають або дані з історії хвороби, або та інформація про пацієнта, яка необхідна лікарю для постановки діагнозу. Недоліком такої схеми є в першу чергу суб'єктивізм лікаря, оскільки при постановці діагнозу він може покладатися тільки на свої знання і свій досвід, накопичений роками лікарської практики. Лікар не отримує ніяких додаткових даних про можливі діагнози. Також він не отримує інформації про схожості даного випадку на ті, що раніше зустрічалися, оскільки далеко не завжди лікар може виділити з отриманої ним кількості показників значимість кожного з них або незвичайне співвідношення показників. Слід також зазначити, що в разі помилки в даних дуже велика ймовірність того, що лікар не зверне на це увагу і подальша діагностика буде проходити під впливом цієї помилки.

Таким чином, очевидна необхідність впроваджувати в процес постановки діагнозу для допомоги лікаря. Вже відомий підхід, коли для поліпшення якості діагностики та для допомоги лікаря використовувалися системи прийняття рішення (рис. 1.2) [14-18].

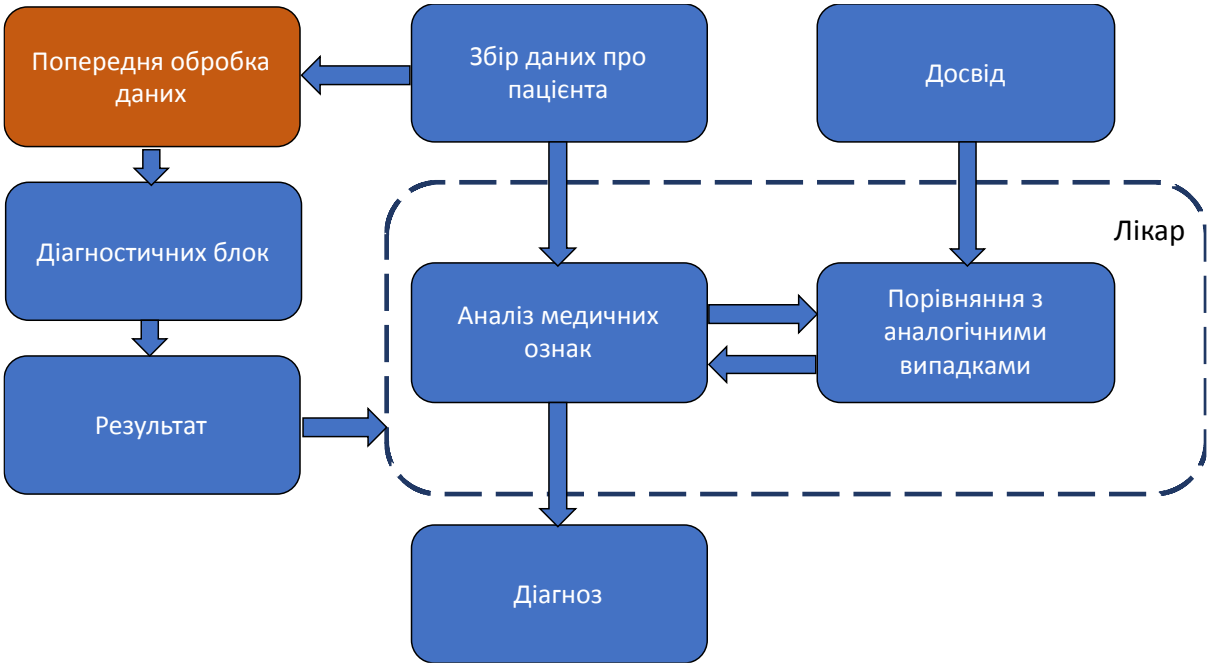

Рисунок 1.2 – Схема формування діагнозу з використанням діагностувальних систем

Такий підхід є найбільш доцільним для використання, при цьому у якості діагностичного блоку можна використовувати велику кількість систем прийняття рішення, що побудовані на нейронних мережах, нечітких системах, нейро-фаззі системах (поєднання нейронних мереж та нечітких систем) і т.ін. Наразі існувало багато спроб використання у якості діагностичного блоку експертних систем, але у такого підходу існує кілька недоліків.

По-перше, як і в першому випадку, немає ніякого захисту від аномальних спостережень (викидів в даних). По-друге, існуючі потужні експертні системи створені під певні галузі медицини (тобто кардіологія, онкозахворювання і т.ін.) і не існує такої експертної системи, яка б визначила діагноз при будь-якому типі захворювання з будь-якої області. В даний час використання експертних систем обґрунтовано тільки в кардіології, а в інших областях медицини шукають нові підходи для вирішення практичних завдань. Таким чином, запропонована схема може бути використана лише в обмеженій кількості випадків.

Слід зазначити, що існує багато вдалих спроб впровадження в процес постановки діагнозу будь-яких комп'ютерних діагностувальних систем, але більшість із них не здатна проводити аналіз даних у ситуаціях наявності пропущених значень. Тому слід докладніше розглянути особливості медичних даних, щоб виявити причини появи таких пропусків.

1.3 Особливості даних медичних досліджень

Медичні дані – одна з форм представлення інформації, що характеризує стан хворого. Такий тип даних відрізняється від інших складністю накопичення з точки зору їх повноти по апріорно заданій вихідній вибірці, а також неоднозначністю аналізу.

В експериментальних наборах даних ознаки можуть бути як однотипними, що можна зустріти вкрай рідко, так і різнотипними, що в реальних вибірках зустрічається практично завжди.

З точки зору їх походження, медичні дані можна розділити на такі групи [13]:

– кількісні або числові ознаки (причому абсолютно різних порядків). До них відносяться заміряні в певних шкалах медичні показники, такі як артеріальний тиск, зріст, вага і т.ін. Ці ознаки можуть вимірюватися в абсолютній шкалою, шкалою інтервалів і шкалою відносин

– якісні ознаки – це ознаки, що характеризують стан хворого в історії хвороби, записані або зі слів пацієнта, або зі слів лікаря. До них відносяться опис зовнішнього вигляду пацієнта, його скарги і т.ін. В цьому випадку вкрай важливим для подальшого аналізу фактом є можливість обмеження кількості цих ознак і можливість їх угруповання або ранжування для оптимального опису стану пацієнта. Такі ознаки вимірюються в шкалі найменувань.

Якісні ознаки в свою чергу можна поділити на:

– рангові або бальні ознаки, що вимірюються в шкалі порядку. До них може відноситись, наприклад, тяжкість стан хворого в залежності від ступеня тяжкості (наприклад, 1 – неважкий стан, 2 – стан середньої тяжкості, 3 – дуже важкий стан). Кількість градацій залежить від розв'язуваної задачі і вимог лікаря-фахівця до деталізації цього показника;

– класифікаційні ознаки, що вимірюються в шкалі найменувань. До них може ставитися статева приналежність, група крові пацієнта;

– графічні ознаки – це графічні залежності, отримані шляхом зняття таких показників, як електрокардіограма, електроенцефалограма і т.ін.

З точки зору достовірності медичні дані можна розділити на достовірні (до них відносять всі ознаки, крім якісних, з попередньої класифікації) і ймовірні (якісні ознаки).

Досліджувані в медицині об'єкти є складними системами, які функціонують при постійному впливі на них множини вхідних факторів (рис.1.3).

Частина з цих факторів X1, X2, ..., Xm є контрольованими. Їх можна виміряти кількісно, оцінити в балах або номінально. Кількість вхідних параметрів визначається в залежності від мети і завдань дослідження. Інша частина відноситься до неконтрольованих і випадкових факторів, які неможливо виміряти, але вони впливають на людину. Результатом цього впливу є випадковість стану і функціонування всієї системи.

Стан системи характеризується великою кількістю вихідних параметрів Y1, Y2, ..., Yn. Ці параметри можуть бути кількісними, вимірюватися в балах або бути випадковими величинами. Їх кількість також визначається в залежності від мети і завдань дослідження.

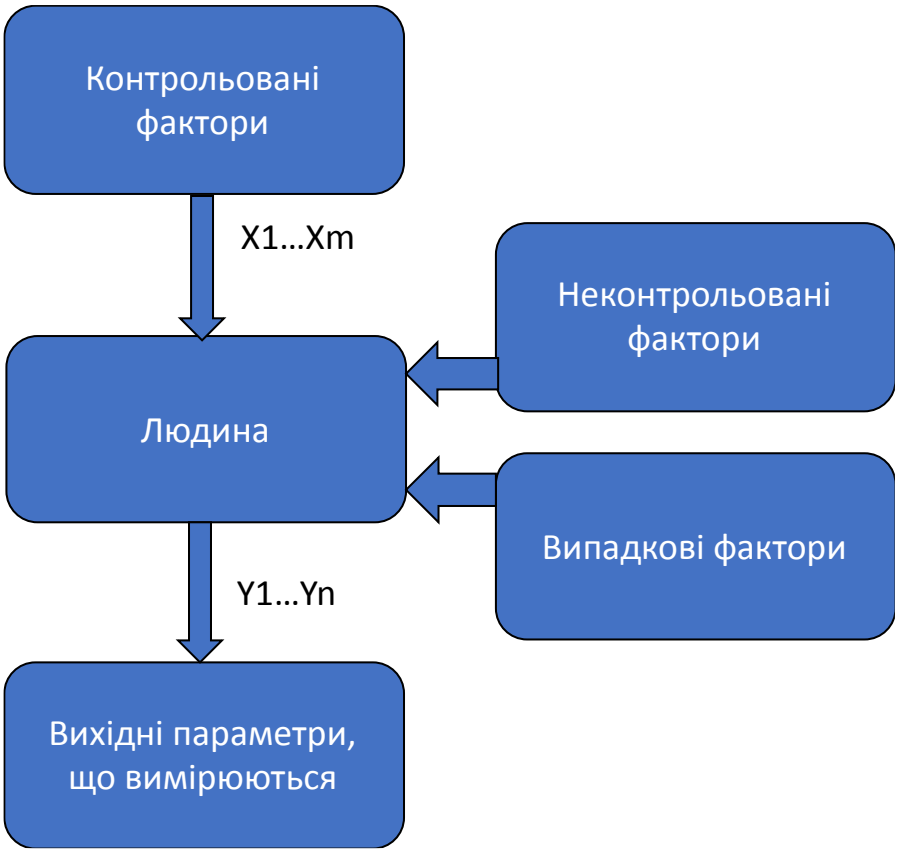

Рисунок 1.3 – Представлення біологічного об'єкта як складної системи

Розглянемо набір медичних показників, як складний об'єкт [10]:

– одна з характерних рис складних об'єктів – це їх багатовимірність. Це означає, що для опису таких об'єктів використовується велика кількість змінних різної фізичної природи при цьому частина цих змінних може мати пропуски;

– при вивченні медичного об'єкта часто виникають проблеми, пов'язані з формуванням його моделі. Причина цього – брак знань про його внутрішні функціональні взаємозв'язки;

– при описі складних об'єктів вимірюються різнотипні їхні властивості. Це означає, що поряд з кількісними оцінками ряду властивостей або показників зустрічаються і якісні оцінки.

Для передбачення деякого виділеної дослідником властивості об'єкта в тому випадку, коли відсутня інформація про структуру та внутрішні взаємозв'язки цього об'єкта використовують метод міркування за аналогією. Для цього вимірюються деякі властивості об'єкта в результаті виходить емпірична таблиця (таблиця «Об'єкт-Властивість») [19], в якій рядки відповідають множині об'єктів, а стовпці – множині властивостей (медичних ознак) цих об'єктів. Як відзначають автори [19], під емпіричною таблицею розуміється таблиця, елементи якої є результатами вимірювань ряду ознак підмножини об'єктів $A = \{a_1, a_2, ..., a_N\}$, обраних з безлічі Г, яка задається в залежності від мети дослідження. Будь-якій підмножині $a \in \Gamma$ можна зіставити вектор $x = (x_1, x_2, ..., x_j, ..., x_n)$ і значення x_0 в просторі ознак $X_1, ..., X_j, ..., X_n; X_0$, де X_0 – цільова ознака. Для кожної ознаки X_j визначена область її значень D_j $(j = 1, ..., n;\ 0)$ і зазначений тип шкали, для якої вона виміряна.

1.4 Постановка завдання дослідження

На підставі розглянутих вище положень, можна виділити основні відмінності медичних даних від інших типів даних:

– медичні дані доводиться обробляти в умовах високої апріорної невизначеності, коли практично нічого не відомо про вигляд функцій розподілу ймовірностей в просторі ознак. Будь-яке припущення про вид функції розподілу (нормальний закон розподілу, незалежності ознак, тощо) ставить питання про адекватність такого припущення [19]. З усього цього випливає, що методи заповнення пропущених значень повинні бути універсальними, орієнтованими на досить слабкі обмеження на вид функції розподілу;

– експериментальні медичні вибірки мають досить малу кількість розглянутих об'єктів, що обмежує вибір методу дослідження та створює необхідність пошуку нових методів заповнення пропущених значень;

– виникають великі труднощі із завданням вихідної системи ознак. Як показує досвід вирішення прикладних завдань [16-19], у кінцеву систему ознак часто входить велика кількість дублюючих і шумливих ознак із пропущеними значеннями. Звідси випливає проблема вибору найбільш інформативної підсистеми ознак, оскільки як показують дослідження [20-23] зменшення числа ознак вихідної системи часто покращує якість рішення. При цьому слід враховувати ситуації, коли заповненні пропущені значення можуть впливати на інформативність медичних ознак, чого за можливістю слід уникати;

– у медичних даних є великий відсоток артефактів. Вони можуть бути наслідком випадкових коливань, обумовлених, наприклад, зовнішнім впливом на пацієнта, його психофізичним станом в момент виміру параметра, або бути артефактом внаслідок похибки дослідження або фіксування результатів дослідження. Досить часто такі артефакти є

тотожними до пропущеного значення у вибірці даних, тобто виникає необхідність їхнього заповнення.

Таким чином, враховуючі відмінності медичних даних слід провести аналіз методів заповнення пропущених значень та обрати той, який забезпечить найкращий варіант відсотків вірно класифікованих пацієнтів в медичних вибірках даних.

2 МЕТОДИ ЗАПОВНЕННЯ ПРОПУЩЕНИХ ЗНАЧЕНЬ В ДАНИХ

Наявність пропусків у даних, так само як і аналіз тільки повних спостережень (після виключення спостережень з пропусками), може призвести до отримання зміщених результатів, і як наслідок до спотворення висновків, які можуть бути зроблені за результатами дослідження і прийняття неправильних діагностичних рішень.

Дану проблему зазвичай вирішують по-різному. Деякі просто виключають з розгляду спостереження з пропущеними даними, що з точки зору персоналізації медицині є неприпустимим, оскільки усі пацієнти мають однакове право на отримання адекватного рівня діагностичних послуг. Іншим підходом є виключення медичних ознак, що містять пропущені значення. Це має більший сенс, оскільки якщо така медична ознака міститься лише у половини усіх пацієнтів, то її можна видаляти. Але інколи лікарі вважають за доцільне залишити ознаку незважаючи на наявність пропущених значень. Тоді слід підходити до вирішення проблеми пропущених даних більш раціонально – на етапі первинної обробки даних заповнити пропуски в уже наявних даних, для того щоб відновити вихідну залежність.

Основною метою було систематизувати основні підходи і методи заповнення пропусків в даних і продемонструвати на практиці використання найбільш універсальних методів заповнення пропусків в даних.

2.1 Існуючі методи заповнення пропусків в даних

Існує безліч способів заповнення пропусків вже після етапу збору даних [24-37]: заповнення середнім значенням, пропорційне розміщення спостережень з пропущеними даними за заздалегідь відомими градаціями шкали, розрахунок можливого значення за допомогою регресійної моделі і

таке інше [24]. Заповнення пропусків дозволяє не тільки отримати додаткову інформацію (передбачені значення), але і зберегти вже наявну, часто дуже важливу і отриману ціною значних зусиль інформацію, за рахунок збереження спостережень, які спочатку містили пропуски.

Крім очевидних переваг, заповнення пропусків як спосіб вирішення проблеми недостатньої інформації має кілька недоліків, які не можна не враховувати:

1. використання для передбачення пропусків наявних повних даних спотворює структуру результуючих даних (після заповнення), яка зміщується в бік структури тільки повних спостережень;

2. існують дослідження, в яких розглянуто ситуації, коли штучна підстановка пропусків вносить в масив певну частку штучних даних, які в свою чергу призводять до зміщення значущості одержуваних на їх основі результатів [25].

Можливо що, моделі, побудовані по заповненим даним, будуть менш точними в порівнянні з ідеальною моделлю, побудованій тільки на повних спостереженнях. Втрати в їх точності залежатимуть від якості передбачення (заповнення) відсутніх значень. Втративши в точності, можна виграти в репрезентативності результатів. Але при цьому заповнення пропущених значень може призвести до збільшення кількості пацієнтів, яким будуть проведені діагностичні заходи.

При виборі конкретного методу відновлення пропущених значень слід враховувати, що, оскільки алгоритми заповнення пропусків не універсальні, можливості застосування того чи іншого способу заповнення пропусків залежить від методу аналізу даних, який планується використовувати в подальшому.

Найбільш популярні з існуючих методів заповнення пропущених значень, представлені в найбільш повній і докладній класифікації Р. Літла [26] і відображені на рисунку 2.1.

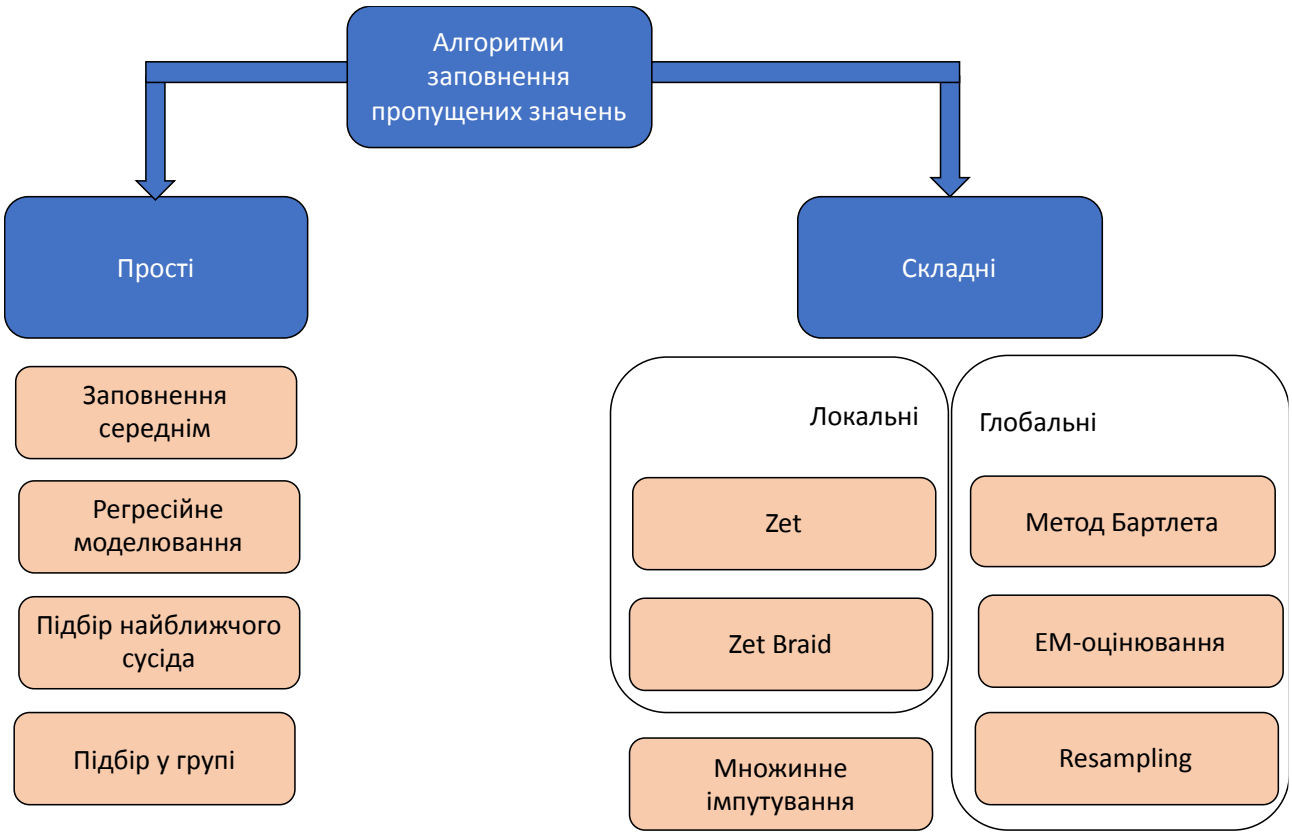

Рисунок 2.1 – Класифікація алгоритмів заповнення пропущених значень

Охарактеризуємо кожну групу методів:

Прості алгоритми – неітеративні алгоритми, засновані на простих арифметичних операціях, відстанях між об'єктами, регресійному моделюванні, тощо. До них відноситься: заповнення пропусків середнім арифметичним (або нульовими значеннями), регресійне моделювання пропусків, метод HotDeck (підбор найближчого сусіда) і підбір в групі.

Складні алгоритми – ітеративні алгоритми, які передбачають оптимізацію деякого функціоналу, що відображає точність розрахунку величини, що підставляється на місце пропущеного значення. Їх можна розділити на глобальні та локальні.

Глобальні алгоритми – це такі алгоритми, коли в оцінюванні (прогнозі) кожного пропущеного значення беруть участь всі об'єкти даної сукупності: метод Бартлета, EM - оцінювання та Resampling.

Локальні алгоритми – такі алгоритми, коли в оцінюванні (прогнозі) кожного пропущеного значення беруть участь повні спостереження, що знаходяться в деякому околі об'єкта з пропущеними значеннями. До даної групи належать алгоритми Zet і Zet Braid.

2.1.1 Заповнення середнім і підбір всередині груп

Метод заповнення середнім значенням передбачає, що всі пропущені значення замінюються середнім значенням цього показника, розрахованим за наявними даними.

Підбір всередині груп передбачає, що вся сукупність об'єктів розбивається на групи за певною ознакою, всередині кожної групи для заповнення пропусків використовуються тільки одне з наявних в ній значень [26].

2.1.2 Підбір найближчого сусіда (Hot Deck)

Hot Deck використовується в одномоментних дослідженнях, і являє собою підстановку замість пропуску значення по даній змінній у найбільш близького об'єкта з повною інформацією. Причому підбір може здійснюватися як з усієї сукупності повних спостережень, так з її деякої підгрупи – кластера, до якого належить цільовий об'єкт.

Для заповнення пропуску по даній характеристиці у цільового об'єкта, використовується значення даної характеристики у об'єкта, найближчого до цільового (відстань до якого від цільового об'єкта менше, менше ніж до всіх інших об'єктів).

Тип функції відстані для визначення спостереження, найближчого до цільового (з пропуском), вибирається виходячи з типу даних, що використовуються, та уявлень дослідника про характер зв'язку між змінними і завдань кожного конкретного дослідження [27].

2.1.3 Метод Бартлета

Метод Бартлета складається з двох етапів: підстановці замість пропусків початкових значень на першому етапі і проведенні на другому етапі коваріаційного аналізу цільової змінної і дихотомічного індикатора повноти спостереження за цільовою змінною. Індикатор повноти спостереження завжди дорівнює нулю, за винятком одного єдиного випадку, коли *i*-е значення є цільовою змінною і є пропущеним, тільки в цьому випадку він приймає значення, що дорівнює одиниці [26].

2.1.4 Алгоритм ZET

Суть алгоритму Zet полягає в підборі для кожного пропуску значення не з усієї сукупності повних спостережень, а з певної її частини, званої компонентної матрицею. Вона складається з композитних рядків і стовпців. Компонентність деякого рядка або об'єкта є оберненою величиною декартової відстані до цільового рядка (неповного спостереження з пропуском) в просторі, осі якого задані змінними – розглянутими характеристиками об'єктів.

За даними компонентної матриці потім будується функціональна залежність прогнозованого значення від відповідного значення в компетентній матриці, на основі якої, потім, прогнозується значення пропуску [28].

2.1.5 Алгоритм ZETBRAID

Основна відмінність і одночасно перевага алгоритму ZetBraid (плетіння) від алгоритму Zet полягає в тому, що в ньому закладено апарат для об'єктивного визначення розмірності компонентної матриці.

У процесі роботи алгоритму відбувається послідовний почерговий відбір компетентних рядків і компетентних стовпців. При кожному новому відборі рядка або стовпця формується нова компетентна матриця. За заданим критерієм визначається її ефективність при прогнозуванні пропусків [29].

2.1.6 Алгоритм RESAMPLING

У ітеративному алгоритмі Resampling рядки, які містять пропущені дані замінюють випадково підібраними рядками з матриці повних спостережень. Потім будується регресійне рівняння для передбачення відсутнього значення.

Процедура побудови регресійного моделювання повторюється кілька разів. Після певної кількості повторень значення отриманих регресійних коефіцієнтів усереднюють і отримують остаточне рішення, що дає максимальну точність прогнозу пропущеного значення [26].

2.1.7 Множинне імпутування.

Метод множинного імпутування був розроблений Дональдом Рубіним в 1970-х роках 20 століття. З точки зору Рубіна приписування кожному пропуску декількох потенційних значень покликане відобразити ступінь невизначеності, з якою здійснюється заповнення пропущеного значення – безпосередньо імпутування. Зараз цей метод є одним з найбільш перспективних і реалізований в спеціалізованому програмному забезпеченні, на жаль, в більшій частині комерційному [30].

Техніка множинного імпутування передбачає підстановку відразу декількох значень на місце кожного з пропусків. Істотний розкид цих значень свідчить про невизначеність моделі і не дозволяє зробити однозначний висновок про їхній тип і причини виникнення.

Дані з кожним набором заповнених пропусків зберігаються в окремий масив, кожен з яких потім аналізується як такий, що складається тільки з повних спостережень [31-34].

2.1.8 EM-оцінювання

Метод максимізації очікувань (EM – expectation maximization), в деяких джерелах так само званий EM – оцінюванням, дозволяє не тільки відновлювати пропущені значення з використанням двохетапного ітеративного алгоритму, а й оцінювати середні значення, коваріаційні і кореляційні матриці для кількісних змінних. EM-алгоритм, в найзагальнішому сенсі являє собою ітераційну процедуру, призначену для вирішення завдань оптимізації деякого функціоналу, через аналітичний пошук екстремуму цільової функції.

Цей алгоритм реалізується в 2 етапи. Перші букви назв, яких утворюють загальну абревіатуру алгоритму:

- етап E;

На першому етапі E (expectation) на основі наявних абсолютно повних або частково (за цільовою змінною) повних спостережень розраховуються умовні очікувані значення цільової змінної для кожного неповного спостереження. потім після отримання масиву повних спостережень, оцінюються основні статистичні параметри: заходи середньої тенденції і розкиду, показники взаємної кореляції і коваріації змінних.

У разі роботи з неповними даними на E-етапі визначається функція умовного математичного очікування логарифма повної функції правдоподібності при відомому значенні цільової змінної X $Q\left(\Theta;\Theta^{(m)}\right)$:

$$Q\left(\Theta;\Theta^{(m)}\right) = E_{\Theta(m)}\left(\log f_{\Theta}(X,Y)\middle|X\right) \qquad (2.1)$$

Коли мають справу з повним спостереженням, у якого характеристика X приймає значення x, вираз (2.1) для обчислення значень функції $Q\left(\Theta;\Theta^{(m)}\right)$ набуває вигляду:

$$Q\left(\Theta;\Theta^{(m)}\right)=\int_{R^m}\left(\log f_{\Theta}(x,y)\right) f_{\Theta^{(m)}}\left(y\middle|x\right)\mu_Y dy \qquad (2.2)$$

Після визначення виду цієї функції починається другий етап роботи алгоритму – М-етап.

- етап М.

На другому етапі М (maximization), задача алгоритму максимізувати ступінь взаємної відповідності очікуваних даних та тих, що реально підставляються, а також відповідності структури імпутованих даних структурі даних повних спостережень.

У класичному варіанті алгоритму, формально завдання по максимізації очікування можна виразити таким чином:

$$\Theta^{(m+1)}=\arg\max Q\left(\Theta;\Theta^{(m)}\right) \qquad (2.3)$$

Тут Θ позначає розраховане очікуване умовне значення, відсутньої характеристики для деякого спостереження [35].

2.1.9 Регресійне моделювання пропусків

У більшості випадків, імпутування за допомогою регресійних моделей здійснюється в два етапи:

1. На першому етапі за сукупністю повних спостережень відбудовується регресійна модель, і оцінюються коефіцієнти в рівнянні, де

в якості залежної змінної виступає цільова змінна, пропущені значення по якій необхідно відновити;

2. Потім за отриманим на попередньому етапі рівняння, в яке підставляються відомі значення незалежних змінних предикторів, для кожного цільового об'єкта розраховується відсутнє значення по залежній цільовій змінній. У разі інтервальних і абсолютних змінних розраховується конкретне значення, а для порядкових і номінальних змінних з певною ймовірністю передбачається категорія, до якої повинен бути віднесений об'єкт.

Вибір регресійної моделі для розрахунку пропущених значень змінної, визначається рівнем вимірювання цільової залежної змінної (значення якої необхідно відновити) і незалежних змінних, за якими будуть пророкувати відсутні значення [24].

2.2 Метод заповнення пропущених значень на основі часткових відстаней між даними

Крім теоретичного опису різних методів імпутування важливо зрозуміти, як вони працюють на практиці. Існує істотна залежність між вибором методу заповнення пропущених значень і тим датасетом, в якому необхідно заповнити пропущені значення. Для частини методів заповнення є необхідним той факт, щоб пропущені значення були лише в деякій частині даних і не перекривали увесь датасет (регресійне моделювання пропусків). Інші потребують наявність нормального (гаусівського) закону розподілу даних (ЕМ-оцінювання). Оскільки в медичних вибірках даних кількість пропущених значень є непрогнозованою, а нормальний закон розподілу майже не виконується, слід обирати такі підходи до заповнення пропущених значень, які не мають додаткових умов їхнього використання.

Основним завданням експерименту, крім оцінки втрат інформації, внаслідок неповноти даних, є порівняння якості заповнення пропущених значень різними методами.

Нехай є вибірка спостережень, що складається з даних пацієнтів $x(k)$, кожен з яких описаний кількістю ознак, що дорівнює n. Загальний обсяг вибірки складає N. Тобто ми маємо таку таблицю даних $X = \{x(1), x(2), ..., x(N)\} \in R^n$. Нехай в даній вибірці деякі пацієнти мають пропущені значення [38-40].

На першому етапі необхідно розрахувати часткові відстані між ознаками пацієнта, якому необхідно заповнити пропущені значення $x(k)$ та усіма іншими пацієнтами $x(p)$, які також можуть містити пропуски в даних:

$$dist_p(k, p) = \frac{1}{n - n_k - n_p + n_{kp}} \sum_{i=1}^{n} \left| x_i(k) - x_i(p) \right|, \qquad (2.4)$$

де n_k – кількість пропущених значень у k-го вектора спостережень,

n_p – кількість пропущених значень у p-го вектора спостереження,

n_{kp} - кількість загальних пропущених значень в одній той самій ознаці в обох об'єктів.

Таким чином, величини $\left| x_i(k) - x_i(p) \right|$ розраховуються тільки для заповнених позицій вхідного вектору.

На наступному етапі усі розраховані значення відстаней упорядковуються таким чином, що:

$$dist_p(k, p_1) < dist_p(k, p_2) < ... < dist_p(k, m) < dist_p(k, p_{N-1}) \qquad (2.5)$$

та визначається рівень належності $\mu_m(k)$ k-го пацієнта до m-го спостереження, який обчислюється за допомогою співвідношення:

$$\mu_m(k) = \frac{dist_p^{-1}(k,m)}{\sum_{m=1}^{N-1} dist^{-1}(k,m)}. \quad (2.6)$$

Неважко помітити, що рівні належності (2.6) задовольняють умовам одиничного розподілу, тобто є нормованими.

На останньому кроці роботи алгоритму у k-му векторі ознак пацієнта, де в ознаці $x_i(k)$ знаходиться пропущене значення, заповнення його реалізується, за допомогою формули:

$$\hat{x}_i(k) = \sum_m \mu_m(k)\, x_i(m) \quad \forall \quad \vec{m} \quad (2.7)$$

Заповнення пропусків проводиться послідовно доти, доки в ознаці i будуть заповнені усі пропущені значення. Таким чином, пропонований метод може бути використаний для заповнення пропущених значень в пакетному та послідовному (онлайн) режимах.

2.3 Висновки за розділом

У розділі розглянуто методи заповнення пропущених значень, які найбільш часто використовуються в багатьох областях. Деякі з них неможливо використовувати в медицині, оскільки вони є занадто примітивними або орієнтованими на закон розподілу даних чи на наявність загально відомої кількості об’єктів у вибірці даних. Тому автор пропонує використання методу заповнення пропущених значень на основі часткових відстаней між даними, який орієнтується на ступінь схожості між

пацієнтами за рахунок використання рівня належності під час заповнення пропущеного значення [38-40]. Такий метод враховує наскільки усі інші ознаки пацієнта є схожими і відновлює пропущене значення базуючись саме на ступені схожості між пацієнтами. Це є логічним з точки зору медицини, оскільки зміни у показниках пацієнтів зі схожими захворюваннями корелюють.

3 АПРОБАЦІЯ МЕТОДІВ ЗАПОВНЕННЯ ПРОПУЩЕНИХ ЗНАЧЕНЬ НА МЕДИЧНИХ ДАНИХ

3.1 Перевірка медичних даних на нормальний закон розподілу

Для використання підходів заповнення пропущених значень середнім значенням по ознаці серед пацієнтів з однаковим діагнозом або методу EM-оцінювання слід бути впевненим, що вибірка даних має нормальний або близький до нього закон розподілу. Тоді такі методи є доцільними для використання. Але, як відомо з літературних джерел, розподіл даних в медичних вибірках зазвичай не є нормальним (гаусівським). Для підтвердження цього факту слід провести дослідження закону розподілу найбільш відомих вибірок медичних даних з репозиторію UCI [41-43], які є вбудованими у бібліотеку sklearn.datasets у Python.

3.1.1 Ділянка квантіль-квантіль

Популярним графіком для перевірки розподілу вибірки даних є графік квантіль-квантіль, графік Q-Q або графік QQ. Цей графік генерує власну вибірку ідеалізованого розподілу, з яким ми порівнюємо, в даному випадку розподіл Гауса. Ідеалізовані зразки діляться на групи (наприклад, 5), які мають назву квантіль. Кожна точка даних у вибірці пов'язана з аналогічним елементом з ідеалізованого розподілу з тим же кумулятивним розподілом. Результуючі точки побудовані у вигляді точкової діаграми з ідеалізованим значенням на осі x і вибіркою даних на осі y.

Ідеальний збіг для розподілу буде показано лінією точок під кутом 45 градусів від нижнього лівого кута графіка до правого верхнього кута. Відхилення по точкам від лінії показують відхилення від очікуваного розподілу.

У Python для побудови графіку QQ використовується функція qqplot() statsmodels, функція бере вибірку даних і передбачає, що ми порівнюємо її з гаусовим розподілом.

Для роботи цієї функції необхідно завантажити медичні дані, на яких буде проведено апробацію, а саме медичні вибірки даних diabetes dataset, breast_cancer dataset, linnerud dataset (Приклад 3.1) [411-43].

```
import numpy as np
import pandas as pd

from sklearn.datasets import load_diabetes, load_breast_cancer,
load_linnerud
line = load_linnerud()
breast = load_breast_cancer()
diabetes = load_diabetes()
```

Приклад 3.1 – Частина програмного коду з файлу Diplom.ipynb для завантаження даних датасетів

На прикладі diabetes dataset можна розглянути які дані містяться у датасеті. diabetes.data містить дані усіх ознак (рис. 3.1), diabetes.target містить інформацію про цільову ознаку (рис. 3.2), diabetes.DESCR містить опис ознак (рис. 3.3), diabetes.feature_names містить перелік ознак.

```
array([[ 0.03807591,  0.05068012,  0.06169621, ..., -0.00259226,
         0.01990842, -0.01764613],
       [-0.00188202, -0.04464164, -0.05147406, ..., -0.03949338,
        -0.06832974, -0.09220405],
       [ 0.08529891,  0.05068012,  0.04445121, ..., -0.00259226,
         0.00286377, -0.02593034],
       ...,
       [ 0.04170844,  0.05068012, -0.01590626, ..., -0.01107952,
        -0.04687948,  0.01549073],
       [-0.04547248, -0.04464164,  0.03906215, ...,  0.02655962,
         0.04452837, -0.02593034],
       [-0.04547248, -0.04464164, -0.0730303 , ..., -0.03949338,
        -0.00421986,  0.00306441]])
```

Рисунок 3.1 – Вміст diabetes.data

```
array([151.,  75., 141., 206., 135.,  97., 138.,  63., 110., 310., 101.,
        69., 179., 185., 118., 171., 166., 144.,  97., 168.,  68.,  49.,
        68., 245., 184., 202., 137.,  85., 131., 283., 129.,  59., 341.,
        87.,  65., 102., 265., 276., 252.,  90., 100.,  55.,  61.,  92.,
       259.,  53., 190., 142.,  75., 142., 155., 225.,  59., 104., 182.,
       128.,  52.,  37., 170., 170.,  61., 144.,  52., 128.,  71., 163.,
       150.,  97., 160., 178.,  48., 270., 202., 111.,  85.,  42., 170.,
       200., 252., 113., 143.,  51.,  52., 210.,  65., 141.,  55., 134.,
        42., 111.,  98., 164.,  48.,  96.,  90., 162., 150., 279.,  92.,
        83., 128., 102., 302., 198.,  95.,  53., 134., 144., 232.,  81.,
       104.,  59., 246., 297., 258., 229., 275., 281., 179., 200., 200.,
       173., 180.,  84., 121., 161.,  99., 109., 115., 268., 274., 158.,
       107.,  83., 103., 272.,  85., 280., 336., 281., 118., 317., 235.,
        60., 174., 259., 178., 128.,  96., 126., 288.,  88., 292.,  71.,
       197., 186.,  25.,  84.,  96., 195.,  53., 217., 172., 131., 214.,
        59.,  70., 220., 268., 152.,  47.,  74., 295., 101., 151., 127.,
       237., 225.,  81., 151., 107.,  64., 138., 185., 265., 101., 137.,
       143., 141.,  79., 292., 178.,  91., 116.,  86., 122.,  72., 129.,
       142.,  90., 158.,  39., 196., 222., 277.,  99., 196., 202., 155.,
        77., 191.,  70.,  73.,  49.,  65., 263., 248., 296., 214., 185.,
        78.,  93., 252., 150.,  77., 208.,  77., 108., 160.,  53., 220.,
       154., 259.,  90., 246., 124.,  67.,  72., 257., 262., 275., 177.,
        71.,  47., 187., 125.,  78.,  51., 258., 215., 303., 243.,  91.,
       150., 310., 153., 346.,  63.,  89.,  50.,  39., 103., 308., 116.,
       145.,  74.,  45., 115., 264.,  87., 202., 127., 182., 241.,  66.,
        94., 283.,  64., 102., 200., 265.,  94., 230., 181., 156., 233.,
        60., 219.,  80.,  68., 332., 248.,  84., 200.,  55.,  85.,  89.,
        31., 129.,  83.])
```

Рисунок 3.2 – Вміст diabetes.target

```
'.. _diabetes_dataset:\n\nDiabetes dataset\n----------------\n\nTen baseline variables, age, sex, body ma
ss index, average blood\npressure, and six blood serum measurements were obtained for each of n =\n442 di
abetes patients, as well as the response of interest, a\nquantitative measure of disease progression one
year after baseline.\n\n**Data Set Characteristics:**\n\n  :Number of Instances: 442\n\n  :Number of Attr
ibutes: First 10 columns are numeric predictive values\n\n  :Target: Column 11 is a quantitative measure
of disease progression one year after baseline\n\n  :Attribute Information:\n      - Age\n      - Sex\n
- Body mass index\n      - Average blood pressure\n      - S1\n      - S2\n      - S3\n      - S4\n
- S5\n      - S6\n\nNote: Each of these 10 feature variables have been mean centered and scaled by the st
andard deviation times `n_samples` (i.e. the sum of squares of each column totals 1).\n\nSource URL:\nhtt
ps://www4.stat.ncsu.edu/~boos/var.select/diabetes.html\n\nFor more information see:\nBradley Efron, Trevo
r Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression," Annals of Statistics (wit
h discussion), 407-499.\n(https://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf)'
```

Рисунок 3.3 – Вміст diabetes.DESCR

Наступним кроком є безпосередньо побудування графіку QQ з використанням функції qqplot() з statsmodels (Приклад 3.2).

Результатом роботи є зображення розподілу точок даних в порівнянні з лінією гаусівського розподілу, що зображено червоною лінією на рисунку 3.4 для diabetes.data, рисунку 3.5 для breast_cancer.data та рисунку 3.6 для linnerud.data.

```
from statsmodels.graphics.gofplots import qqplot
from matplotlib import pyplot
data = np.array(diabetes.data)
qqplot(data, line='s')
pyplot.show()
```

Приклад 3.2 – Частина програмного коду з файлу Diplom.ipynb для побудування графіку QQ

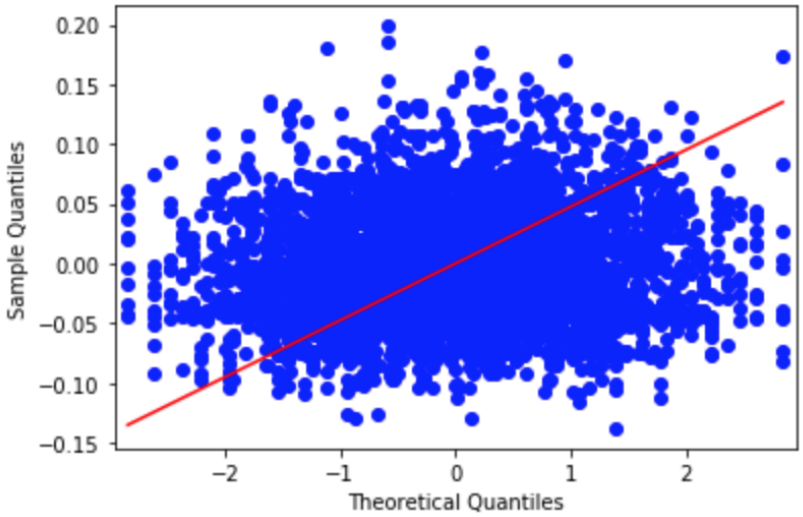

Рисунок 3.4 – Графік QQ для diabetes dataset

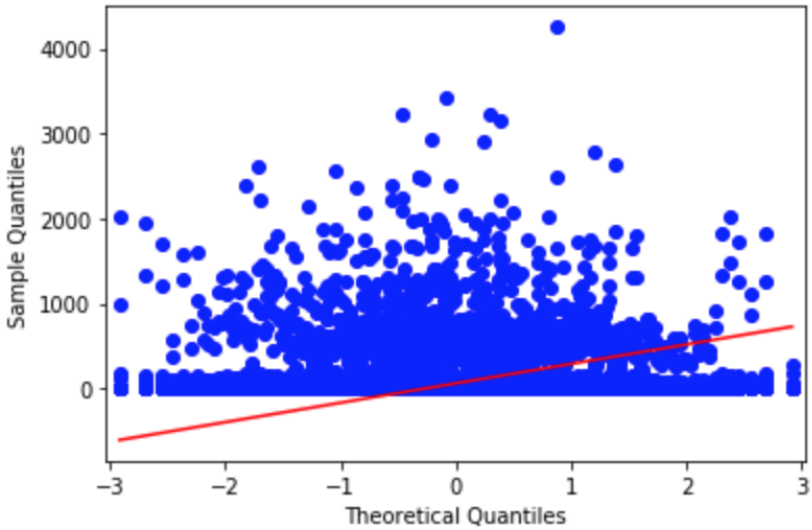

Рисунок 3.5 – Графік QQ для breast_cancer dataset

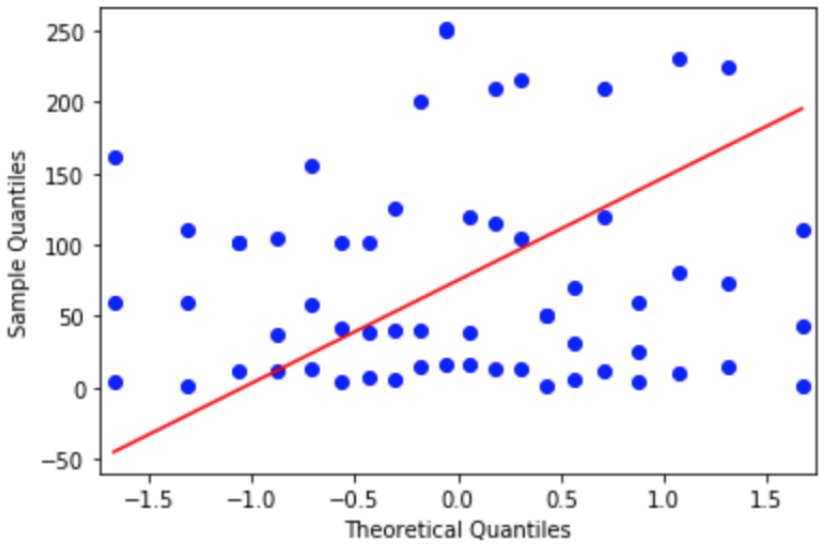

Рисунок 3.6 – Графік QQ для linnerud dataset

З рисунків добре видно, що жоден з медичних датасетів не є близьким до нормального закону розподілу. Але для того, щоб підтвердити це чисельним чином, необхідно розрахувати спеціальні статистики, які розглядаються у наступному пункті.

3.1.2 Тест Шапіро-Уілка

Тест Шапіро-Уілка проводить оцінку вибірки даних і дає кількісну оцінку ймовірності того, що дані були отримані з гаусівського розподілу, названого на ім'я Самуеля Шапіро та Мартіна Уілка.

На практиці тест Шапіро-Уілка вважається надійним тестом нормальності, хоча є деякі припущення, що цей тест може бути придатний для невеликих вибірок даних, наприклад, тисячі спостережень або менше.

У Python для розрахунку функції Шапіро-Уілка по заданому набору даних використовується shapiro() з SciPy. Функція повертає W-статистику, розраховану тестом, і значення p. Перш за все необхідно вірним чином інтерпретувати результати:

- W-статистика – кількість, розрахована за допомогою тесту, яку можна інтерпретувати в контексті тесту, порівнюючи її з критичними значеннями з розподілу статистики тесту;

- p-значення використовується для інтерпретації тесту, в даному випадку, чи була вибірка отримана з гаусівского розподілу. Технічно це називається нульовою гіпотезою, або H0. Обрано пороговий рівень, званий альфа, зазвичай 5% (або 0,05), який використовується для інтерпретації значення p.

При реалізації цих тестів за допомогою SciPy можна інтерпретувати значення *p* наступним чином:

- $p < \alpha$: відхилити нульову гіпотезу H0, закон розподілу не є нормальним;

- $p \geq \alpha$: неможливо відхилити нульову гіпотезу H0, закон розподілу є нормальним.

Це не означає, що нульова гіпотеза вірна, а лише значить, що це цілком ймовірно, з огляду на наявні докази. Значення *p* не є ймовірністю того, що дані відповідають гаусівському розподілу; це можна розглядати як значення, яке допомагає нам інтерпретувати статистичний тест.

Тобто, загалом, ми шукаємо результати з великим значенням *p*, щоб підтвердити, що вибірки даних, ймовірно, були взяті з гаусівського розподілу.

Застосуємо тест Шапіро-Уілка для кожної з вибірок даних (Приклад 3.3).

В результаті роботи програми отримано такі дані, що жоден з датасетів не відноситься до нормального закону розподілу (рисунок 3.7 для diabetes.data, рисунок 3.8 для breast_cancer.data та рисунок 3.9 для linnerud.data).

```
from scipy.stats import shapiro
stat, p = shapiro(data)
print('Statistics=%.3f, p=%.3f' % (stat, p))
alpha = 0.05
if p > alpha:
      print('Sample looks Gaussian (fail to reject H0)')
else:
      print('Sample does not look Gaussian (reject H0)')
```

Приклад 3.3 – Частина програмного коду з файлу Diplom.ipynb для розрахунку тест Шапіро-Уілка

```
Statistics=0.986, p=0.000
Sample does not look Gaussian (reject H0)
```

Рисунок 3.7 – Результат розрахунку тесту Шапіро-Уілка для diabetes.data

```
Statistics=0.295, p=0.000
Sample does not look Gaussian (reject H0)
```

Рисунок 3.8 – Результат розрахунку тесту Шапіро-Уілка для breast_cancer.data

```
Statistics=0.854, p=0.000
Sample does not look Gaussian (reject H0)
```

Рисунок 3.9 – Результат розрахунку тесту Шапіро-Уілка для linnerud.data

3.2 Дані пульмонологічних захворювань

Для перевірки роботи різних методів заповнення пропущених значень на реальних клінічних даних було отримано датасет, що містить інформацію про пацієнтів із пульмонологічними захворюваннями. Датасет містить

інформацію про 132 пацієнти, кожен з яких описаний 104 ознаками, серед яких 24 ознаки описують стать, вік та скарги пацієнта, 14 ознак складають анамнез, 26 ознак для об'єктивного опису стану пацієнта, 10 ознак описують клінічний аналіз крові, 8 ознак – біохімічний аналіз крові, 10 ознак для клінічного аналізу сечі, рентгенографія грудної клітки описано 6-ма ознаками, 8-ма ознаками описано електрокардіографічні дослідження (ЕКГ), 2-ма ознаками спірометричні дослідження. Кожен з пацієнтів відноситься до одного з трьох діагнозів: хронічне обструктивне захворювання легень (ХОЗЛ) – 46 пацієнтів, бронхіальна астма – 53 хворих, пневмонія – 33 пацієнти. Повну таблицю медичних даних наведено в додатку А.

3.2.1 Формування DataFrame

Усі дані пульмонологічних досліджень представлено у вигляді файлу формату Dataset.xlsx (Додаток А). Для завантаження даних до Python Data Frame необхідно прописати Header таким чином, як показано у Прикладі 3.4.

```
import numpy
import pandas
pandas.set_option('display.max_rows', 500)
pandas.set_option('display.max_columns', 500)
pandas.set_option('display.width', 500)
file = pandas.ExcelFile('Dataset.xlsx')
df = pandas.read_excel(
    file,
    sheet_name = 0,
    header = 1,
    names = None,
    index_col = None,
    usecols = None,
    squeeze = False,
    dtype = None,
    engine = None,
    converters = None,
```

```
    true_values = None,
    false_values = None,
    skiprows = None,
    nrows = None,
    na_values = None,
    keep_default_na = True,
    verbose = False,
    parse_dates = False,
    date_parser = None,
    thousands = None,
    comment = None,
    skip_footer = 0,
    skipfooter = 0,
    convert_float = True,
    mangle_dupe_cols = True)
df.replace("ХОЗЛ", 1, inplace=True)
df.replace("БА", 2, inplace=True)
df.replace("Пневмонія", 3, inplace=True)
df.drop(df.columns[[0]], axis=1, inplace=True)
```

Приклад 3.4 – Частина програмного коду з файлу Diplom.ipynb для завантаження даних до Python Data Frame

3.2.2 Перевірка на нормальний закон розподілу

Наступним кроком необхідно перевірити дані на нормальний закон розподілу для того, щоб обмежити можливі методи заповнення пропущених значень. Для цього необхідно використані програмний код з Прикладів 3.2 та 3.3. В результаті отримаємо загальний вигляд графіку квантіль-квантіль (рисунок 3.10) та значення W-статистики та p для тесту Шапіро-Уілка (рисунок 3.11). З обох результатів добре видно, що дані зовсім не відповідають нормальному закону розподілу, тому для заповнення пропущених значень в таких даних не можна використовувати методи EM-оцінювання та заповнення середнім значенням.

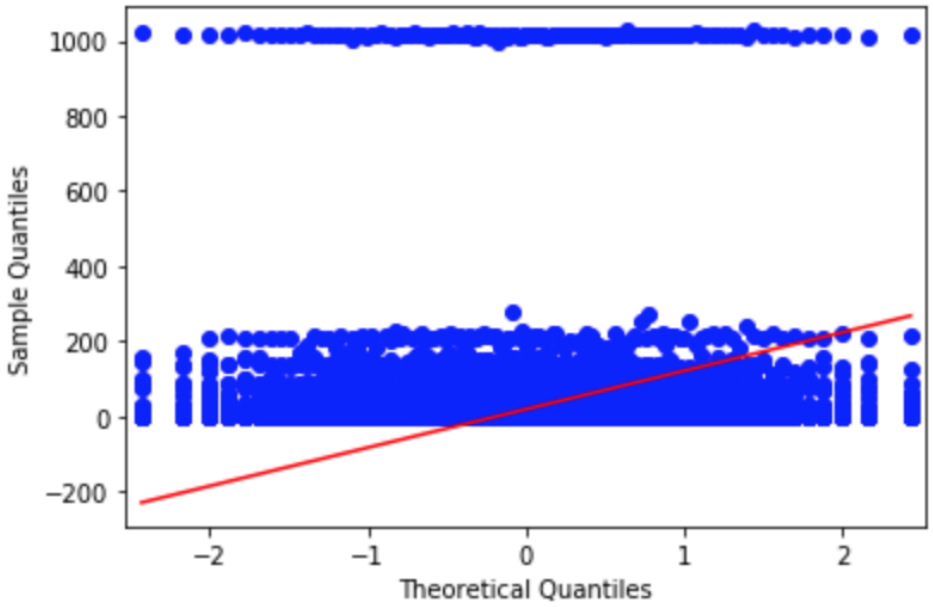

Рисунок 3.10 – Графік QQ для клінічних даних пацієнтів пульмонологічними захворюваннями

```
Statistics=0.167, p=0.000
Sample does not look Gaussian (reject H0)
```

Рисунок 3.11 – Результат тесту Шапіро-Уілка для клінічних даних пацієнтів пульмонологічними захворюваннями

3.2.3 Аналіз кількості пропущених значень та їх заповнення

Першим пунктом при аналізі цієї вибірки даних є оцінювання кількості пропущених значень. Для цього використовується код, наведений в Прикладі 3.5.

```
gap = df.isnull().sum()
gap.to_excel("gap.xlsx")
```

Приклад 3.5 – Частина програмного коду з файлу Diplom.ipynb для підрахування кількості пропущених значень

У додатку Б наведено перелік усіх ознак та кількість пропусків в них (вміст файлу gap.xlsx), а в таблиці 3.1 наведено перелік лише тих ознак, які містять пропущені значення та кількість таких пропусків.

Таблиця 3.1 – Перелік тих ознак, які містять пропущені значення та кількість таких пропусків

Ознака	Кількість пропущених значень
ЧДД в минуту	10
АД сист	10
АД диаст	10
Пульс	7
эритроциты	3
Hb	2
цветовой показатель	8
лейкоциты	2
эозинофилы	3
палочкоядерные	2
сегментоядерные	2
лимфоциты	2
Моноциты	3
СОЭ	4
билирубин 1	55
билирубин 2	55
билирубин 3	55
АСТ	97
АЛТ	97
глюкоза	34
сиаловые кислоты	74
серомукоиды	74
удельный вес	30
белок	2
лейк п/зр	2
эритр неизм п/зр	2

З таблиці добре видно, що загальна кількість пропущених значень складає 645, що є 4,698% від загальної кількості ознак (13728 ознак).

Наступним шагом є заповнення пропущених значень з використанням регресійного методу та методу, що базується на часткових відстанях. Але попередньо необхідно оцінити доцільність заповнення пропущеного значення при співвідношенні кількості пропущених значень до загальної кількості ознак. Таким чином, добре видно, що в деяких ознаках пропущена більша половина ознак і заповнення пропущених значень в них є можливим, але не має сенсу. Тому перед заповненням пропущених значень слід видалити ту частину ознак, пропущені значення в яких перевищують 10% від загальної кількості ознак. Для цього використано програмний код, наведений в прикладі 3.6.

```
reducedAllData = allData.copy()
columnCount = len(allData.index)
indexesToDelete = []
for x, row in allData.iterrows():
    nanCount = 0
    for y, value in row.iteritems():
        if numpy.isnan(value):
            nanCount += 1
    if ((nanCount/columnCount) * 100) > 10:
        indexesToDelete.append(x)
reducedAllData.drop(reducedAllData.index[indexesToDelete], inplace=True)
availableData = reducedAllData.dropna()
resultData = reducedAllData.copy()
```

Приклад 3.6 – Частина програмного коду з файлу Diplom.ipynb для видалення частини ознак

Таким чином було видалено 9 ознак, кількість пропусків в яких перевищувала 10%. Усі інші пропуски в ознаках було заповнено за допомогою двох методів: перший – метод регресійного моделювання вже реалізований в середовищі Python; другий – метод, що базується на

часткових відстанях, програмний код якого наведено в прикладі 3.7., а його застосування в прикладі 3.8.

```
def getMissingValue(missingValueX, missingValueY, rowCount):
    distances = []
    for x, availableRow in availableData.iterrows():
        res = 0
        for y, value in allData.iloc[missingValueX].iteritems():
            if pandas.isnull(value) != True:
                res += abs(availableRow[y] - value)
        distances.append(res/rowCount)
    inverseDistancesSum = 0
    for distance in distances:
        inverseDistancesSum += 1/distance
    affiliationLevels = []
    for distance in distances:
        affiliationLevels.append((1/distance)/inverseDistancesSum)
    missingValue = 0
    iterator = 0
    for x, value in availableData[missingValueY].iteritems():
        missingValue += value * affiliationLevels[iterator]
        iterator += 1
    return missingValue
```

Приклад 3.7 – Частина програмного коду з файлу Diplom.ipynb, що реалізує метод заповнення пропущених значень, що базується на часткових відстанях

```
for x, row in allData.iterrows():
    for y, value in row.iteritems():
        if pandas.isnull(value):
            resultData.loc[x, y] = getMissingValue(x, y, len(availableData.
index) + 1)
```

Приклад 3.8 – Частина програмного коду з файлу Diplom.ipynb для заповнення пропущених значень з використанням функції getMissingValue

3.2.3 Аналіз отриманих результатів

Таким чином ми маємо датасет, що було заповнено двома різними методами і наступним завданням є оцінка отриманих результатів для того, щоб обрати найкращій метод заповнення.

На першому етапі проведемо аналіз кожної ознаки з використанням методу df.describe(). В Таблицях 3.2 та 3.3 наведено основні статистичні характеристики даних з пропущеними значеннями.

Таблиця 3.2 – Основні статистичні характеристики даних з незаповненими пропусками

	ЧДД в минуту	АД сист	АД диаст	Пульс	Эритр.	Hb	цветовой показатель	Лейк.	Эозин.
count	122	122	122	125	129	130	124	130	129
mean	23,53	133,85	82,28	88,73	4,40	137,27	0,93	8,33	1,47
std	3,38	15,18	11,52	14,04	0,48	15,03	0,05	5,03	1,02
min	15	60	8	56	3,27	92	0,72	1	1
25%	22	120	80	80	4,2	127,25	0,9	6,2	1
50%	24	130	80	90	4,37	137	0,91	7,3	1
75%	26	140	90	95	4,51	148	0,94	9,2	2
max	33	180	110	150	8,1	176	1,1	57,6	7

Таблиця 3.3 – Основні статистичні характеристики даних з незаповненими пропусками

	Палочко-ядерные	Сегменто-ядерные	Лимфо-циты	Моно-циты	СОЭ	белок	лейк п/зр	эритр неизм п/зр
count	130	130	130	129	128	130	130	130
mean	5,80	62,95	24,36	4,55	16,30	0,01	4,37	0,38
std	10,08	12,92	9,69	2,47	13,18	0,02	4,38	1,30
min	1	3	3	2	2	0	0	0
25%	2	59	18	3	7	0	2	0
50%	3	64,5	25	4	11,5	0	3	0
75%	5,75	69	31	6	21	0	5	0
max	69	87	69	20	74	0,09	40	8

Важливою складовою процесу заповнення є незмінність внутрішньої структури даних. Для даних, що було відновлено за допомогою регресійного моделювання основні статистичні характеристики наведено в таблицях 3.4 та 3.5.

Таблиця 3.4 – Основні статистичні характеристики даних з пропусками, заповненими на основі регресійного моделювання

	ЧДД в минуту	АД сист	АД диаст	Пульс	Эритр.	Hb	цветовой показатель	Лейк.	Эозин.
count	132	132	132	132	132	132	132	132	132
mean	23,78	134,12	81,30	88,55	4,40	137,27	0,93	8,31	1,46
std	3,42	15,63	11,25	13,87	0,48	14,91	0,05	4,99	1,01
min	15	60	8	56	3,27	92	0,72	1	1
25%	21,50	120	80	80	4,2	127,75	0,9	6,2	1
50%	23	130	80	88	4,39	137	0,91	7,35	1
75%	25,25	140	90	94,25	4,50	148	0,94	9,2	1,59
max	33	180	110	150	8,1	176	1,1	57,6	7

Таблиця 3.5 – Основні статистичні характеристики даних з пропусками, заповненими на основі регресійного моделювання

	Палочко-ядерные	Сегменто-ядерные	Лимфо-циты	Моно-циты	СОЭ	белок	лейк п/зр	эритр неизм п/зр
count	132	132	132	132	132	132	132	132
mean	5,83	62,99	24,36	4,55	16,33	0,01	4,38	0,38
std	10,00	12,82	9,62	2,44	12,98	0,02	4,34	1,29
min	1	3	3	2	2	0	0	0
25%	2	59	18	3	7,75	0	2	0
50%	3,5	65	24,60	4	12,5	0	3	0
75%	6	69	31	6	21	0	5	0
max	69	87	69	20	74	0,09	40	8

Для даних, що було відновлено за допомогою регресійного моделювання основні статистичні характеристики наведено в таблицях 3.6 та 3.7.

Таблиця 3.6 – Основні статистичні характеристики даних з пропусками, заповненими на основі методу часткових відстаней

	ЧДД в минуту	АД сист	АД диаст	Пульс	Эритр.	Hb	цветовой показатель	Лейк.	Эозин.
count	132	132	132	132	132	132	132	132	132
mean	23,38	133,59	82,46	88,32	4,40	137,27	0,92	8,32	1,46
std	3,29	14,62	11,09	13,76	0,48	14,91	0,05	5,00	1,00
min	15	60	8	56	3,27	92	0,72	1	1
25%	21,50	120	80	80	4,2	127,75	0,9	6,2	1
50%	23	130	80	88	4,39	137	0,91	7,35	1
75%	25,25	140	90	94,25	4,50	148	0,94	9,2	1,59
max	33	180	110	150	8,1	176	1,1	57,6	7

Таблиця 3.7 – Основні статистичні характеристики даних з пропусками, заповненими на основі методу часткових відстаней

	Палочко-ядерные	Сегменто-ядерные	Лимфо-циты	Моно-циты	СОЭ	белок	лейк п/зр	эритр неизм п/зр
count	132	132	132	132	132	132	132	132
mean	5,83	62,99	24,36	4,55	16,33	0,01	4,38	0,38
std	10,00	12,82	9,62	2,44	12,98	0,02	4,34	1,29
min	1	3	3	2	2	0	0	0
25%	2	59	18	3	7,75	0	2	0
50%	3,5	65	24,60	4	12,5	0	3	0
75%	6	69	31	6	21	0	5	0
max	69	87	69	20	74	0,09	40	8

При порівнянні таблиць добре видно, що є вкрай малі зміни таких параметрів, як середнє значення (mean), стандартне відхилення (std), в деяких ознаках також змінилися параметри розподілу між максимальним та мінімальним значенням.

Також було проведено порівняння тих самих статистичних ознак для тих ознак, що було видалено з чого отримано обґрунтування необхідності видалення показників з кількістю пропусків більшою за 10%. Основні статистичні характеристики таких показників наведено в таблиці 3.8,

результати відновлення ознак, що містять більше за 10% пропусків наведено в таблиці 3.9,

Таблиця 3.8 – Основні статистичні характеристики ознак, що містять більше за 10% пропусків

	Билирубин 1	Билирубин 2	Билирубин 3	АСТ	АЛТ	Глюкоза	Сиал. кисл.	Серомукоиды	Удельный вес
count	77	77	77	35	35	98	58	58	102
mean	13,11	3,82	10,24	37,01	29,01	5,25	184,4	0,18	1016,19
std	3,49	1,07	8,01	14,32	15,63	1,63	36,57	0,060	4,38
min	9,5	2	5,2	0,33	0,33	3,5	100	0,12	1000
25%	11	3,1	7,9	30,5	18,5	4,3	160	0,1425	1014
50%	12,2	3,8	8,7	34	27	4,9	190	0,18	1016
75%	13,9	4	10,1	43	35,5	5,575	200	0,2	1018
max	31	9	75	69	88	16,5	280	0,46	1028

Таблиця 3.9 – Основні статистичні характеристики заповнених ознак, що містять більше за 10% пропусків

	Билирубин 1	Билирубин 2	Билирубин 3	АСТ	АЛТ	Глюкоза	Сиал. кисл.	Серомукоиды	Удельный вес
count	132	132	132	132	132	132	132	132	132
mean	13,08	3,77	9,862	33,99	23,42	5,33	200,39	0,19	1016,2
std	2,66	0,82	6,11	7,54	8,65	1,41	28,10	0,04	3,85
min	9,5	2	5,2	0,33	0,33	3,5	100	0,12	1000
25%	11,85	3,5	8,4	32,41	20,99	4,67	190	0,18	1015
50%	12,87	3,71	9,20	32,99	21,53	5,39	210,23	0,20	1016,1
75%	13,21	4	9,59	33,66	21,96	5,64	214,01	0,21	1018
max	31	9	75	69	88	16,5	280	0,46	1028

Порівняння цих двох таблиць дозволяє стверджувати, що при зростанні кількості пропущених значень істотно змінюються показники

середнього значення, стандартного відхилення та розподілу між максимум та мінімумом, що представлено в таблиці 3.10.

Таблиця 3.10 – Основні статистичні характеристики заповнених ознак, що містять більше за 10% пропусків

	Били-рубин 1	Били-рубин 2	Били-рубин 3	АСТ	АЛТ	Глю-коза	Сиал. кисл.	Серо-мукоиды	Удель-ный вес
count	55	55	55	97	97	34	74	74	30
mean	-0,029	-0,049	-0,378	-3,011	-5,576	0,086	15,90	0,012	0,027
std	-0,8286	-0,253	-1,891	-6,775	-6,977	-0,217	-8,467	-0,018	-0,532
min	0	0	0	0	0	0	0	0	0
25%	0,85	0,4	0,5	1,919	2,497	0,375	30	0,045	1
50%	0,678	-0,085	0,503	-1,003	-5,465	0,49	20,23	0,024	0,197
75%	-0,687	0	-0,506	-9,339	-13,53	0,069	14,01	0,014	0
max	55	55	55	97	97	34	74	74	30

3.2.4 Перевірка якості заповнення шляхом класифікації даних

Для перевірки якості заповнення даних пропонується провести класифікацію даних за допомогою методу логістичної регресії. Попередньо необхідно провести нормування даних за допомогою методу MinMaxScaler з бібліотеки sklearn.preprocessing (прикладу 3.9).

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(0, 1))
scaler_x = scaler.fit(dfProc.values)
x_norm = scaler_x.transform(dfProc.values)
```

Приклад 3.9 – Частина програмного коду з файлу Diplom.ipynb для нормування даних

Після цього проводиться безпосередньо навчання моделі логістичної регресії за допомогою методу LogisticRegression з sklearn.linear_model. Для оцінки якості роботи методу слід провести оцінку роботи моделі за допомогою метрик з бібліотеки sklearn.metrics, результат роботи методу на даних, що заповнені методом, що базується на часткових відстанях представлено на рисунку 3.12. Візуалізація клінічних даних пульмонологічних досліджень на базі метода головних компонент приведена на рисунку 3.13 (червоний колір – пацієнти з ХОЗЛ, зелений колір – пацієнти з бронхіальною астмою, синій колір – пацієнти з пневмонією).

```
import numpy as np
from sklearn.linear_model import LogisticRegression
logistic_model = LogisticRegression()
logistic_model.fit(x_norm, y)
y_predicted = logistic_model.predict(x_norm)
error = np.mean(y != y_predicted)
print('Error =', error)
from sklearn.metrics import mean_squared_error, r2_score,
mean_absolute_error
mae = mean_absolute_error(y, y_predicted)
rmse = mean_squared_error(y, y_predicted)
r2 = r2_score(y, y_predicted)
print('Mean Absolute Error:', mae)
print('Root mean squared error: ', rmse)
print('R2 score: ', r2)
```

Приклад 3.10 – Частина програмного коду з файлу Diplom.ipynb для класифікації за допомогою логістичної регресії та оцінка роботи моделі

```
Error = 0.007575757575757576
Mean Absolute Error: 0.007575757575757576
Root mean squared error:  0.007575757575757576
R2 score:  0.9871332488546642
```

Рисунок 3.12 – Значення метрик MAE, RMSE та R^2

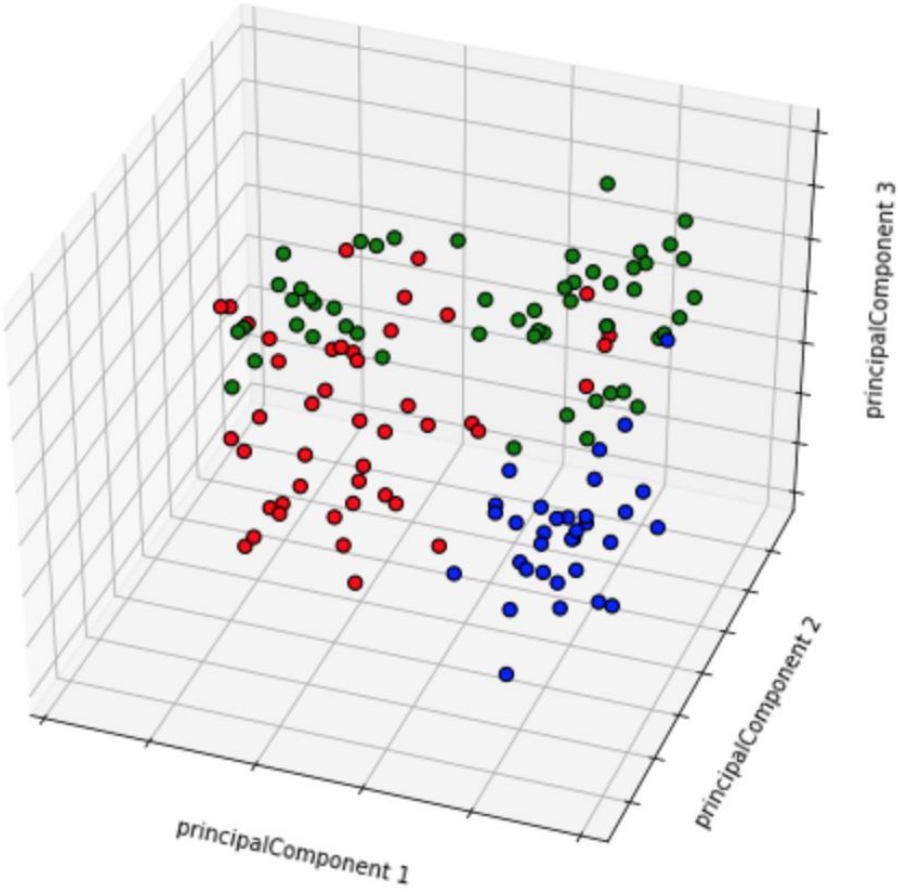

Рисунок 3.13 – Візуалізація даних з використанням методу PCA

Для порівняння використаного методу з іншими слід провести класифікацію датасету, що заповнений іншими методами. Для методу заповнення середніми значеннями результати наведено на рисунку 3.14, для методу регресійного моделювання – на рисунку 3.15.

```
Error = 0.09645467235
Mean Absolute Error: 0.09645467235
Root mean squared error: 0.09645467457
R2 score: 0.8534262748596
```

Рисунок 3.14 – Значення метрик MAE, RMSE та R^2 для датасета, заповненого середніми значеннями

```
Error = 0.0094765985
Mean Absolute Error: 0.0094765985
Root mean squared error: 0.0094766575
R2 score: 0.9534262748596
```

Рисунок 3.15 – Значення метрик MAE, RMSE та R^2 для датасета, заповненого методом регресійного моделювання

Для того, щоб зрозуміти де саме модель логістичної регресії невірно класифікувала пацієнтів необхідно побудувати матрицю спряження (Confusion Matrix) за допомогою програмного коду з Прикладу 3.11. Результат роботи методу наведено на рисунку 3.16. Добре видно, що класифікатор невірно класифікував 1 пацієнта з бронхіальною астмою встановивши йому діагноз ХОЗЛ.

```
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
print(confusion_matrix(y, y_predicted))
import seaborn as sns
plt.figure(figsize=(5, 5))
g = sns.heatmap(confusion_matrix(y, y_predicted), annot=True, fmt='.3f',
cmap='BuPu')
```

Приклад 3.11 –Частина програмного коду з файлу Diplom.ipynb для побудування Confusion Matrix

Матриця спряження побудована лише для того датасету, який було відновлено за допомогою методу часткових відстаней, оскільки з метрик добре видно, що інші методи заповнення дали гірші результати при подальшій класифікації.

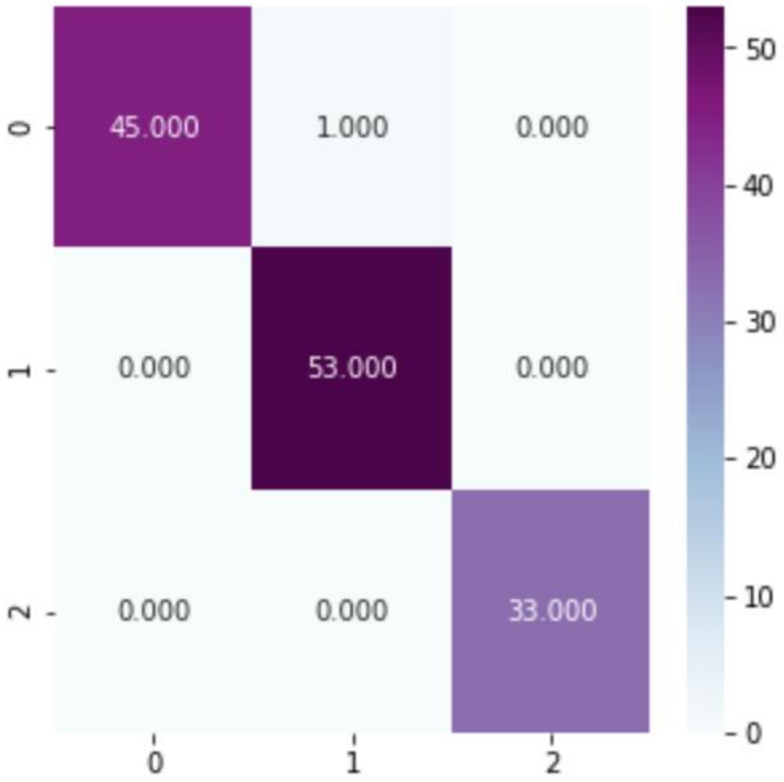

Рисунок 3.16 – Confusion Matrix

Таким чином можна зробити висновок, що метод, який базується на часткових відстанях продемонстрував найкращій результат.

ВИСНОВКИ

Під час виконання атестаційної роботи проведено огляд літератури на тему дослідження, а саме підходів до заповнення пропущених значень в медичних вибірках даних. По-перше, було проведено аналіз предметної галузі, при якому розглянуто основні завдання медичної діагностики та причини виникнення пропусків у медичних даних. Проведено порівняльну оцінку методик формування діагнозу лікарем та виділено відмінні риси даних медичних досліджень від інших видів даних. Було представлено аналіз людини, як складної системи, що піддається впливу контрольованих та неконтрольованих зовнішніх факторів.

Таким чином було проведено постановку завдання дослідження, що полягає в проведенні аналізу методів заповнення пропущених значень та виборі того, який забезпечить найкращий варіант відсотків вірно класифікованих пацієнтів в медичних вибірках даних.

Для аналізу була взята вибірка клінічних даних в області пульмонології, що містила 645 пропущених значень при загальній кількості пацієнтів 132 людини та кількості ознак 104. Для заповнення пропущених значень було використано метод регресійного моделювання та метод, що базується на часткових відстанях. Результат заповнення пропусків було перевірено шляхом аналізу статистичних характеристик даних до та після заповнення пропусків та шляхом класифікації даних методом логістичної регресії з використанням метрик MAE, RMSE та R^2 та побудуванням Confusion Matrix. Найкращі значення метрик продемонстрував метод, що базується на часткових відстанях тому можна зробити висновок, що цей метод є найбільш придатним для використання в медичних датасетах.

ПЕРЕЛІК ПОСИЛАНЬ

1. Акопов В.И. Медицинское право в вопросах и ответах. М.: Приор, 2001. 208 с.

2. Imbs J.L., Pouyanne P., Haramburu F. et al. Iatrogenic medication: estimation of its prevalence in French public hospitals. *Regional Centers of Pharmacovigilance Therapie*. 1999. 54 (1). P. 21-27.

3. Johnson J.A., Bootman J.L. Drug-related morbidity and mortality. A cost-of-illness model. *J. Arch. Intern. Med.*, 1995 Oct. 9. 155 (18). P.1949-1956.

4. Бейли Н. Математика в биологии и медицине. Москва: Мир, 1970. 326 с.

5. Кобринский Б.А. Автоматизированные диагностические и информационно-аналитические системы в педиатрии. *Русский медицинский журнал*. 1999, т. 7. №4. С. 5-8.

6. Илларионова А.Р., Фридман Н.В. Диспансерное наблюдение больных глаукомой в условиях поликлиники. *Русский медицинский журнал*. 2001, т. 2. №3. С. 118-121.

7. Ю.И. Фещенко, О.М. Рекалова, С.Д. Кузовкова Морфологические предпосылки для фунгальной инвазии в легкие и ее влияние на течение неспецифических заболеваний легких у больных. *Український пульмонологічний журнал.* 2007. № 1. С. 17-21.

8. Е. Фриджлинг-Шредер Пограничные состояния у детей. *Журнал практической психологии и психоанализа.* 2003. №2. С. 22-25.

9. Глова С.Е., Кательницкая Л.И., Хаишева Л.А. и др. Скрининг сердечно-сосудистой патологии и ассоциированных поведенческих факторов риска у жителей г. Ростова-на-Дону. *Российский кардиологический журнал.* 2006. №3(59). С. 31-36.

10. Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981. 160 с.

11. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М.: Наука, 1974. 415 с.

12. Раудис Ш.Ю. Ограниченность выборки в задачах классификации. *Статистические проблемы управления.* 1976. вып. 18. 180 с.

13. Юнкеров В.И., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований. СПб.: ВМедА, 2002. 266с.

14. Урбах В.Ю. Статистический анализ в биологических и медицинских исследованиях. М.: Медицина, 1975. 295с.

15. Зайцев Г.Н. Математический анализ биологических данных. М.: Наука, 1991. 183с.

16. Pliss I., Perova I. Diagnostic Neuro-Fuzzy System and Its Learning in Medical Data Mining Tasks in Conditions of Uncertainty about Numbers of Attributes and Diagnoses. *Automatic Control and Computer Sciences*. 2017. 51(6). pp.391-398. doi: 10.3103/ S0146411617060062

17. Бражнікова Є.М., Перова І.Г. Інформаційна технологія аналізу потоків медичних даних за умов невизначеності. *Прикладная радиоэлектроника*. 2019. № 1-2. с. 46-51

18. Бодянский Е.В., Перова И.Г. Нейро-фаззи система для задач обработки медицинских данных в ситуациях множества диагнозов. *Бионика интеллекта*. Харьков: ХНУРЭ, 2015. Вип. 2 (85). с. 86-89

19. Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981. 160 с.

20. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М.: Наука, 1974. 415 с.

21. Раудис Ш.Ю. Ограниченность выборки в задачах классификации. *Статистические проблемы управления*. 1976. вып. 18. 180 с.

22. Perova I., Bodyanskiy Ye. Adaptive Human Machine Interaction Approach for Feature Selection-Extraction Task in Medical Data Mining. *International Journal of Computing*. 17(2). 2018. 113-119 p.

23. Bodyanskiy Ye., Perova I., Zhernova P. Online fuzzy clustering of high-dimensional data based on ensembles in data stream mining tasks. *Сучасний стан наукових досліджень та технологій в промисловості*. 2019. №1(7). с. 16-24.

24. Крыштановский А.О. Анализ социологических данных с помощью пакета SPSS. М.: ГУ, ВШЭ. 2006. 263 с.

25. Rubin, D.B. Multiple Imputation for Nonresponse in Surveys. Ney York: Willey, 1987. P. 64-66.

26. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных. *Computer Modeling & New Technologies*. Vol. 6. 2004. с. 55 – 56.

27. Kalton, G., Kasprzyk, D. The treatment of missing survey data. *Survey Methodology*, №12, 1986. P. 1-16.

28. Снитюк В.Е., Эволюционный метод восстановления пропусков в данных, 2008. Електронний ресурс. Режим доступу: http://iissvit.narod.ru/index_a.htm (дата звернення 25.04.2020).

29. Алгоритм ZetBraid. *Информационные интеллектуальные системы*. Вып.40, 2008. Електронний ресурс. Режим доступу: http://iissvit.narod.ru/rass/vip40.htm (дата звернення 20.04.2020).

30. Horton N. J., Lipsitz S.R. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. *The American Statistician*, Vol. 55, № 3. (Aug., 2001), P. 244-254 Електронний ресурс. Режим доступу: http://links.jstor.org/sici?sici=0003-1305%28200108%2955%3A3 %3C244%3AMIIPCO%3E2.0.CO%3B2-J

31. Rubin, D.B. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, № 91, 1996. P. 473-489.

32. Lipsitz S.R., Lue Ping Zhao, Molenberghs G.A. Semiparametric Method of Multiple Imputation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 60, № 1. 1998, P. 127-144. Електронний ресурс. Режим доступу: http://links.jstor.org/sici?sici=1369-7412%281998%

2960%3A1%3C12 7%3AASMOMI%3E2.0.CO%3B2-5 (дата звернення 30.04.2020)

33. Schafer, J.L. Multiple Imputation: A Primer, *Statistical Methods in Medical Research*, Vol. 8, 1999. P. 3-15.

34. Schulte Nordholt E. Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistical Review / Revue Internationale de Statistique*, Vol. 66, № 2. 1998, P. 157-180.

35. Королев В.Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор. М.: 2007. 102 с.

36. Bodyanskiy Ye. V., Kulishova N. Ye. Memory based neuro-fuzzy system for printing inks color reproductive properties description. *Wissenschaftliche Berichte von Hochschule Zittau*. Goerlitz. Heft 100, Nr. 2360 – 2395, 2008. – P. 61 – 69.

37. Bodyanskiy, Y.V., Kulishova, N.Y. Memory-based neuro-fuzzy system for interpolation of reflection coefficients of printing inks. *Cybern Syst Anal*. 2008. 44. pp. 625–632. https://doi.org/10.1007/s10559-008-9034-8

38. Бодянский Е.В., Перова И.Г. Нечеткая классификация данных медико- биологических исследований в условиях дефицита информации. *Системи обробки інформації.* 2015. Вип. 11(136). С.161-163.

39. Бодянский Е.В., Перова И.Г. Восстановление пропусков в таблицах данных на основе метода нечеткой пространственной экстраполяции с использованием манхэттенской метрики. *Интеллектуальные системы принятия решения и проблемы вычислительного интеллекта ISDMCI'2015.* (Железный порт, Украина 18-22 травня 2015). Железный порт. 2015. с. 306-308

40. Mulesa P., Perova I. Fuzzy Spacial Extrapolation Method Using Manhattan Metrics for Tasks of Medical Data Mining. *Computer Science and Information Technologies CSIT'2015*. (Lviv, Ukraine, 14-17 September 2015). Lviv. 2015. P. 104- 106.

41. Dermatology dataset, May 2008. [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning- databases/dermatology/ dermatology.data.

42. Breast Cancer in Wisconsin dataset, 03 Dec 1996. [Електронний ресурс]. Режим доступу: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer- wisconsin/breast-cancer-wisconsin.data (дата звернення 03.05.2020)

43. Sklearn.datasets.load_linnerud May 2015. [Електронний ресурс]. Режим доступу: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_linnerud.html#sklearn.datasets.load_linnerud (дата звернення 03.05.2020)