ЗМІСТ

ВСТУП

На сьогоднішній день в Україні стоїть проблема безпліддя у жінок та недоношеність плоду. В Україні недоношеними народжуються 7% дітей. Це є проблемою, тому що передчасні пологи можуть закінчитися перинатальною чи неонатальною смертю. Також у недоношених дітей проявляються складні захворювання:

- Дитячий церебральний параліч
- Хронічні респіраторні захворювання
- Сліпота
- інші

Проблема безпліддя та недоношеності стоїть також і в розвинутих країнах.

В Україні на сьогоднішній день існують медичні заклади (один із них знаходиться в м. Києві – НМАПО ім. П. Л. Шупика), які допомагають подружнім парам в лікуванні безпліддя та недоношеності. Вони проводять аналізи на генетичному рівні з метою отримання повної картини, на основі яких призначають лікування та роблять передбачення про подальший розвиток подій.

Аналізи беруть в подружніх парах (чоловіка та дружини) з метою отримання даних по безпліддю пари. Коли жінка завагітніла і є ризики недоношеності плоду проводять дослідження ще й матері жінки. Також збирають генеалогічну інформації і по батьківській лінії.

Щоб дати оцінку можливості безпліддя чи недоношеності плоду подружньої пари беруть до уваги велику кількість параметрів. Приблизне їх число 70 параметрів.

Лікарі призначають лікування і надають рекомендації на основі своїх знань та досвіду роботи. Проте бажають мати докази своїх рішень на основі статистичних даних обробки інформації хворих.

Тому проблема, яка розглядається в дослідженні − безпліддя у подружньої пари.

Об'єкт дослідження: безпліддя подружньої пари та недоношеність плоду.

Предмет дослідження: медична статистика та прогноз в медицині.

Мета дослідження: покращення коректності лікарського висновку, на основі розробленої моделі отриманої в результаті аналізу медичних даних пацієнтів.

ПЕРЕЛІК СКОРОЧЕНЬ

ML– MachineLearning (машинне навчання)

AUC – Area Under Curve (площа під кривою)

ROC – Receiver Operating Characteristic (робочахарактеристикаприймача)

1. ВИХІДНІ ДАНІ

1.1. Аналіз наявних рішень

Аналізуючи проблему безпліддя пари та недоношеності був проведений огляд програмних засобів, які використовують лікарі для отримання статистичних даних та моделей на основі яких можна проводити прогнозування. Зазвичай це програмні пакети в яких реалізовані алгоритми DataMining. До них відносяться RapidMiner, MDR, SNPstat, Orange та інші. Проблема безпліддя вимагає досліджувати її трьома методами. Перший – це аналіз генетичної інформації, а саме генотипів та їх комбінації. Другий – аналіз генеалогічних даних жінки та чоловіка. Третій – аналіз факторів, що впливають на здоров'я. Отже для повноцінного дослідження не вийде обмежитись одним програмним продуктом. Наприклад, MDRбув розроблений для оцінки сили впливу певної комбінації генів і його не можна застосувати для побудови моделі, яка включатиме в себе вік та інші числові атрибути. Тому потрібно використовувати ще й інші програмні засоби.

За кордоном використовують вище описані програмні продукти для побудови моделей та прогнозу на їх основі. Українські лікарі також намагаються використовувати дані продукти. Але вони є занадто складними для лікаря, які не мають спеціальної освіти. Додатки з дружнім програмним інтерфейсом, що дозволяють аналізувати безпліддя у подружньої пари,отримувати статистичнуінформацію з вибірки та на її основі прогнозувати, в Україні немає.

Дане дослідження є важливим для української медицини, оскільки не проводилися подібні. Ніхто не звертав особливої уваги на пари чоловік-дружина. Тобто, не було створеної моделі, яку можна було б використовувати в прогнозуванні. Саме це є метою дослідження. Отримання найбільш адекватної моделі, яку можна використати в прогнозуванні. А також розробка програмного забезпечення, що матиме дружній інтерфейс для лікаря без спеціальної освіти і за допомогою якого можна буде створювати нові моделі на основі обраних вибірок даних, оцінювати їх і використовувати в прогнозуванні.

1.2. Дані дослідження.

Дослідження проводиться з даними отриманими в НМАПО ім. П. Л. Шупика. Лікарі генетики визначили фактори, на основі яких вони проводять аналіз подружніх пар і дають лікарські висновки. Надані фактори складають чимале число. Загальна їхня кількість дорівнює 45.

Ознайомившись з факторами відразу стає зрозумілим, що їх можна поділити на три логічні частини, де кожну з частин можна досліджувати окремо.

Головною частиною дослідження є генотипи чоловіка та жінки. Саме отримання статичних даних по цій вибірці є найважливішими, тому що доведено, що при певній комбінації генів у подружньої пари не може бути дітей.

Однак, тільки даних по генотипах не достатньо, тому що є ситуації в яких при однакових комбінаціях генів для різних пар маємо різні результати. Одні можуть завагітніти інші ні.

Саме тому потрібно досліджувати інші фактори. Так маємо дві інші групи, одна з яких це – генеалогічна інформація предків. Оскільки безпліддя вважається мультифакторним захворюванням, то в даній групі містяться дані не генотипів, а інформація про тяжкі захворювання. Наприклад, інсульти в родині.

До третьої групи відносяться дані, що описують стан здоров'я пацієнта та умови в яких він проживає.

Для дослідження було надано інформацію по 230 подружніх парах. 100 з яких не мають проблем з вагітністю і 130 з яких мають.

Нижче приведені фактори з кожної групи даних.

1.2.1. Генотипи подружніх пар

Вже доведені багатьма дослідженнями гени, які впливають на безпліддя. Вони є однаковими як для чоловіків так і для жінок. Нижче наведені гени, які завжди перевіряють українські лікарі в НМАПО ім. П. Л. Шупика, та їх всі можливі варіанти.

Назва гена	**Значення 1**	**Значення 2**	**Значення 3**
PAI	5G/5G	5G/4G	4G/4G
MTHFR_C677T	CC	CT	TT
MTHFR_A1298C	AA	AC	CC
FV	GG	GA	AA
FII	GG	GA	AA
MTRR_A66C	GG	GA	AA
FGB_C148T	CC	CT	TT
FGB_455GA	GG	GA	AA
ITGA2a2	CC	CT	TT
ITGB3b1	CC	CT	TT

Таблиця 1.2.1.1. Генотипи подружніх пар

Отримані дані були приведені в зручний формат для аналізу.

В результаті отримали таблицю, що містить 11 атрибутів.

Перед кожним геном було доставлені символи або F_, абоM_ (F – Female, M – male) для того, щоб розмежувати атрибути чоловіка та жінки. Також було додано останню колонку, яка вказує на наявність проблеми з вагітністю. 1 якщо є проблеми з вагітністю, 0 − немає.

Нижче приведена одна таблиця розділена на дві з підготовленими даними для аналізу.Показані перші 6 записів з 230.

F_PAI	F_MTHFR_C677T	F_MTHFR_A1298C	F_FV	F_FII	F_MTRR_A66C	F_FGB_C148T	F_F
4G/4G	CT	AC	GG	GG	AG	CT	GA
5G/5G	CC	AC	GG	GG	AG	CC	GG
5G/4G	CC	AC	GG	GG	AA	CT	GA
5G/4G	CC	AA	GG	GG	GG	CT	GA
5G/4G	CT	AA	GG	GG	GG	CT	GA
5G/4G	CT	AA	GG	GG	AG	CT	GA
M_PAI	M_MTHFR_C677T	M_MTHFR_A1298C	M_FV	M_FII	M_MTRR_A66C	M_FGB_C148T	M_
5G/4G	CC	AC	GG	GG	AG	CC	GG
4G/4G	CC	AC	GG	GG	GG	CT	GA
5G/4G	CT	AA	GG	GG	AA	CC	GG

Таблиця 1.2.1.2. Дані генотипів подружньої пари

1.2.2. Генеалогічні дані поколінь

До подружніх пар, що мають випадки захворювань у старших поколіннях, додається інформація генеалогічних даних хвороб чоловіка та/або дружини. В таблиці наведені атрибути з описом генеалогічних даних.

Назва	**Тип даних**	**Опис**
Ступінь спорідненості	фактор	Номер покоління
Спадковість	біноміальні	
Материнська	біноміальні	
батьківська	біноміальні	
загальне число захворювань у родичів	число	
інфаркти	число	
число інфарктів в родині	число	
інсульти	число	
число інсультів в родині	число	
тромбоемболії	число	
число тромбоемболій в родині	число	
варикоз	число	
число варикозу в родині	число	
безпліддя	число	
частота безпліддя	число	
невиношування вагітності	число	
частота невиношування	число	
естрогензалежні захворювання	число	
число естрогензалежних	число	
онкологічні захворювання	число	
число онкологічних захворювань	число	
стать	біноміальні	

Таблиця 1.2.2.1. Генеалогічні фактори

Генеалогічні дані представлені 22 факторами.

1.3. Оцінка даних

Дослідження проводяться з даними по 230 парах. Це означає, що маємо вибірку для дослідження з 230 записами. Кожен запис містить результат, що вказує на наявність безпліддя чи на його відсутність у пари. Результат має всього лише два стани, а отже його можна представити в біноміальному

вигляді, тобто значенням true (правда – у пари безпліддя) абоfalse(не правда – пара може завагітніти). В подальшому аналізі замість true, falseбудемо використовувати числа 1(true) та 0 (false).

230 записів це достатня кількість на основі яких можна будувати моделі і використовувати в прогнозуванні. Однак загальна кількість факторів для кожного запису досягає 45. Це означає, що не можна побудувати найкращої моделі якщо використовувати всі фактори. В такому випадку потрібно застосувати методи, що проаналізують зв'язки між факторами і оберуть ті, які мають найбільший вплив на результат. Фактори, що не впливають або майже не впливають можна відкидати.

В отриманих даних відсутні деякі значення, а конкретно в генеалогічних. Числовідсутніх даних є достатньо високим і тому це не може не вплинути на оцінку. В такому випадку потрібно застосовувати методи, що певним чином заповнять відсутні дані, використовувати алгоритми, що враховують відсутні дані або видалення колонок з великою кількість відсутніх даних.

Оскільки вибірку даних можна розділити на дві логічні частини їх будемо досліджувати окремо. Після отримання моделей можна визначити наскільки добре кожна з них прогнозує результат. Потім потрібно змінювати отримані моделі до тих пір поки не буде знайдено найкращу модель.

1.4. Висновок до розділу

Для дослідження безпліддя у подружньої пари установа НМАПО ім. П. Л. Шупика надала230 записів по їх генотипах та обтяженої спадковості родоводу. Цієї кількості достатньо, щоб за допомогою методів машинного навчання отримати модель прогнозування.

Такий набір даних дозволить провести дослідження генотипів пари з урахуванням генеалогічної спадкової інформації, що не робилося до цього в Україні.

2. МЕТОДИ ОБРОБКИ ДАНИХ

2.1.Машинне навчання

Машинне навчання (англ. Machine Learning) – великий підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних навчатися. Розрізняють два типи навчання. Навчання по прецедентах, або індуктивне навчання, засноване на виявленні закономірностей в емпіричних даних. Дедуктивне навчання передбачає формалізацію знань експертів і їх перенесення в комп'ютер у вигляді бази знань. Дедуктивне навчання прийнято відносити до області експертних систем, тому терміни машинне навчання і навчання по прецедентах можна вважати синонімами.

Машинне навчання знаходиться на стику математичної статистики, методів оптимізації та дискретної математики, але має також і власну специфіку, пов'язану з проблемами обчислювальної ефективності і перенавчання. Багато методів індуктивного навчання розроблялися як альтернатива класичнимстатистичним підходам. Багато методи тісно пов'язані з витяганням інформації, інтелектуальним аналізом даних.

Машинне навчання – це технологія використання наявних статистичних даних минулих подій для вирішення якихось проблем, наприклад, прогнозування майбутніх тенденцій. Сьогодні подібні інструменти використовуються все частіше: пошуковики, онлайнові рекомендації в інтернет-магазинах, вказівки напрямку в GPS-навігаторах та персональні асистенти на кшталт Cortana. Усі вони працюють за допомогою машинного навчання, однак їхня функціональність – це лише частина того, що можна досягнути.

Подібна технологія здатна поліпшити обслуговування споживачів, передбачати епідемії хвороб, а також прогнозувати та запобігати злочинності. Щоб сучасні програмні інструменти могли це робити, необхідно, щоб машинне навчання було доступним – спочатку для корпорацій, а згодом і кожному.

Сьогодні тренуванням комп’ютерів займаються самі корпоративні споживачі для власних мереж. Це потребує роботи вчених, яких бракує. Цьому

також заважають ліцензії на програмне забезпечення, які можуть коштувати великих грошей, а популярні мови програмування для вирішення статистичних завдань можуть бути складними у вивченні. Навіть якщо бізнес подолає ці проблеми, він все одно не зможе впровадити нові моделі машинного навчання у свої кінцеві інструменти, оскільки це часто потребує декількох місяців інженерної роботи. Масштабування, управління та моніторинг подібних систем вимагає таких ресурсів та спеціалістів, які мають лише декілька корпорацій.

Нижче наведені методи, які будуть використовуватися в дослідженні.

2.1.1. Логістична регресія

Логістична регресія – це вид нелінійної множинної регресії, яка аналізуєфункціональну залежність між декількома незалежними змінними(регресорами) і залежною змінною [1, 2]. Безпосередньо у медичних задачах застосовується бінарна логістична регресія, оскільки значення вихідноїзмінної (наявність хвороби) може приймати два значення: хворий (1 або true), або здоровий (0 або False). У загальному вигляді регресійна модель записується так:

$$y = F(x1,x2,...,xn). \quad (2.1.1.1)$$

У випадку застосування бінарної логістичної регресії ймовірність *p*наявність хворобипов’язана зі змінними (x1, ..., xn), як показано у рівнянні (2.1.1.2):

$$\ln\left(\frac{P}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \qquad (2.1.1.2)$$

Оцінка β_0 називається перетином, а інші оцінки – β_1. – ще називаютькутовими коефіцієнтами. Дана модель коректно представляє ймовірність міжнулем і одиницею, незалежно від вхідних значень. Використовуючи векторидля визначення параметрів моделі і вхідних змінних, ймовірність хворобиможна представити у такій формі(2.1.1.3):

$$p = \frac{\exp(\beta X)}{1 + \exp(\beta X)} \qquad (2.1.1.3)$$

2.1.2. Дерева рішень

Метод дерева рішень – метод ситуаційного аналізу, сутність якого полягає у процедурі прийняття управлінських рішень з погляду оцінки рівня ризику з певного питання, яке виникає в результаті реалізації будь-яких проектів[2, 3].

Метод дерева рішень найбільш популярний в менеджменті для визначення та вибору оптимального напряму дій із наявних варіантів. Метод дерева рішень – це схематичне подання проблеми прийняття рішень. Дерево рішень подають графічно у вигляді деревовидної структури. Порівнявши рівень витрат і рівень доходу, аналітик (фінансовий менеджер) визначає рівень чистого виграшу і відображає на вузлах дерева через його гілки. Кожна гілка визначає раціональність цього рішення, враховуючи ймовірність настання негативної події.

Таким чином, метод дерева рішень дає змогу керівнику врахувати різні напрями дій, узгодити з ними фінансові результати, скорегувати їх зі згідно приписаної їм імовірності, зробити порівняння альтернатив. Невід'ємна частина цього методу – концепція очікуваного значення.

Основні етапи обґрунтування та прийняття рішення щодо здійснення проекту за допомогою побудови дерева рішень:

Етап 1. Формулювання кінцевої мети проекту.

Етап 2. Визначення сукупності можливих дій для розгляду та аналізу проекту.

Етап 3. Оцінка можливих варіантів дій та їх ймовірностей.

Етап 4. Оцінка математичного очікування можливого доходу.

Етап 5. Прийняття рішення (рис. 2.1.2.1).

Метод побудови дерева рішень доцільно застосовувати на початковій стадії розробки проекту, коли прогнозований стан структурують, виділяючи ключові моменти, в яких слід приймати рішення з певною ймовірністю.

Методику побудови дерева рішень розглянемо на конкретному прикладі.

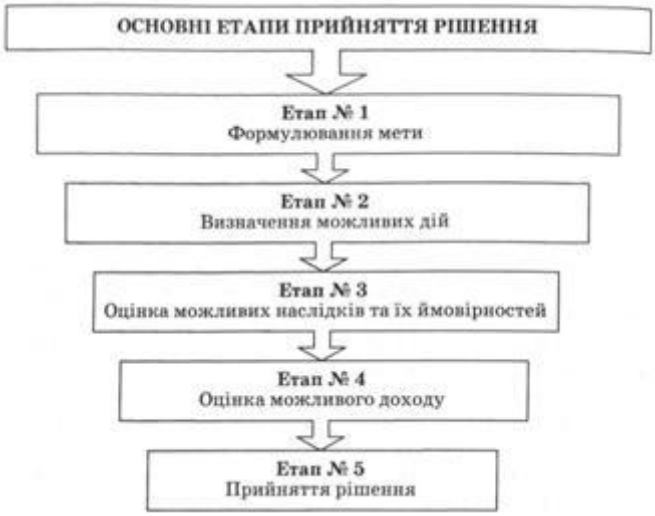

Рис. 2.1.2.1. Основні етапи прийняття рішення за допомогою методу побудови дерева рішень

Випадковий ліс (random forest) – алгоритм машинного навчання, запропонований Лео Брейманом і Адель Катлер, що полягає у використанні ансамблю дереврішень [1]. Алгоритм поєднує в собі дві основні ідеї: метод баггінга (bagging – від bootstrap aggregation, тобто дерева навчаються по випадковим підвибіркам) і метод випадкових підпросторів. Алгоритм застосовується для задач класифікації, регресії і кластеризації.

Випадковим лісом називається класифікатор, який складається з набору дерев $\{h(x,\theta_k)k=1,\ldots\}$, де $\theta_k -$ незалежні однаково розподілені випадкові вектори і кожне дерево вносить один голос при визначенні класу x. Відповідно до цього визначення, випадковим лісом є класифікатор, який складається з ансамблю дерев рішень, кожне з якого будується з використанням баггінга. В цьому випадку θ_k представляє собою незалежні однаково розподілені *l*-мірні випадкові вектори, координати яких є незалежними дискретними випадковими величинами, які приймають значення з підмножини $\{1,2,3,\ldots,l\}$ з рівними

ймовірностями. Реалізація випадкового вектора θ_k визначає номера прецедентів навчальної вибірки, які утворюють бутстреп (bootstrap – тобто вибірка з поверненням) вибірку, яка використовується при побудові k-го дерева рішень.

Алгоритм індукції випадкового лісу може бети представлений у такому вигляді [2]:

1. Для $i = 1, 2, \dots, B$ (де B – кількість дерев в ансамблі) виконати:

   Сформувати бутстреп вибірку S розміру l по вихідній навчальній вибірці $D = \{x_i, y_i\}_i^l$;

   По бутстреп вибірці S індукувати не усічене дерево рішень T_i з мінімальною кількістю спостережень в термінальних вершинах, що дорівнює n_{min}, рекурсивно слідуючи наступному алгоритму:

   a) з вихідного набору n ознак випадковообрати p ознак;

   b) з p ознак обрати ознаку, яка забезпечує найкраще розщеплення;

   c) розщепити вибірку, яка відповідає обробленій вершині, на дві підвибірки;

2. В результаті виконання кроку 1 отримуємо ансамбль дерев рішень $\{T_i\}_{i=1}^B$;

3. Передбачення нових спостережень здійснювати наступним чином:

   a) для регресії:

$$\hat{f}_r^B{}_f(x) = \frac{1}{B}\sum_{i=1}^{B} T_i(x);$$

(2.1.2.1)

b) для класифікації: нехай $\widehat{w}_i(x) \in \{w_1, w_2, \ldots, w_c\}$ – клас, передбачений деревом рішень T_i, тобто $T_i(x) = \widehat{w}_i(x)$; тоді $\widehat{w}_{rf}^{B}(x)$ – клас, який найчастіше зустрічається в множині $\{\widehat{w}_b(x)\}_{i=1}^{B}$.

Одна з переваг випадкових лісів полягає в тому, що для оцінки ймовірності помилкової класифікації немає необхідності використовувати крос-перевірку (cross-validation) або тестову вибірку. Оцінка ймовірності помилкової класифікації випадкового лісу здійснюється методом "Out-Of-Bag" (OOB), який полягає у наступному. Так як, кожна бутстреп вибірка не містить приблизно 37% спостережень вихідної навчальної вибірки (оскільки вибірка з поверненням, то деякі спостереження в неї не потрапляють, а деякі – декілька раз). Класифікуємо деякий вектор $x \in D$. Для класифікації використовуються тільки ті дерева випадкового лісу, які будувалися по бутстреп вибіркам, що не містять x, і як звичайно використовується метод голосування. Частота помилково класифікованих векторів навчальної вибірки при такому способі класифікації і представляє собою оцінку ймовірності помилкової класифікації випадкового лісу методом OOB. Практика застосування оцінки OOB показала, що у випадку, якщо кількість дерев велика, то ця оцінка має високу точність.

Випадкові ліси мають цілу низку переваг, в порівнянні з деревами рішень, що і зумовило їх широке застосування, а саме:

1. Випадкові ліса забезпечують істотне підвищення точності, так як дерева в ансамблі слабокорельовані внаслідок подвійної ін'єкції випадковості в індуктивний алгоритм – за допомогою баггінга і використання методу випадкових підпросторів при розщепленні кожної вершини;

2. Методично і алгоритмічно складне завдання обрізання повного дерева рішень знімається , оскільки дерева у випадковому лісі не обрізаються (це також призводить до високої обчислювальної ефективності);

3. Відсутня проблема перенавчання (overfitting) (навіть при кількості ознак, що перевищує кількість спостережень навчальної вибірки і великій кількості дерев). Тим самим знімається складна проблема відбору ознак, необхідна для інших ансамблевих класифікаторів;

4. Простота застосування: єдиними параметрами алгоритму є кількість дерев в ансамблі і кількість ознак, випадково відібраних для розщеплення в кожній вершині дерева;

5. Легкість організації паралельних обчислень.

Недоліками випадкових лісів у порівнянні з деревами рішень є відсутність візуального представлення процесу прийняття рішень і складність інтерпретації їх рішень.

2.1.3. Нейронна мережа

Штучні нейронні мережі (ШНМ) – математичні моделі, а також їх програмні або апаратні реалізації, побудовані за принципом організації й функціонування біологічних нейронних мереж – мереж нервових кліток живого організму. Це поняття виникло при вивченні процесів, що протікають у мозку, і при спробі змоделювати ці процеси. Першою такою спробою були нейронні мережі Маккалока й Піттса. Згодом, після розробки алгоритмів навчання, одержувані моделі стали використовувати в практичних цілях: у завданнях прогнозування, для розпізнавання образів, у завданнях керування й ін.

ШНМ являють собою систему з'єднаних і взаємодіючих між собою простих процесорів (штучних нейронів). Такі процесори звичайно досить прості, особливо в порівнянні із процесорами, використовуваними в персональних комп'ютерах. Кожний процесор подібної мережі має справу тільки із сигналами, які він періодично одержує, і сигналами, які він періодично посилає іншим процесорам. Проте, з'єднавши їх в досить велику мережу з керованою взаємодією, такі локально прості процесори разом здатні виконувати досить складні завдання. З погляду машинного навчання, нейронна мережа являє собою окремий випадок методів розпізнавання образів, методів

кластеризації й т.п. З математичної точки зору, навчання нейронних мереж – це багатопараметричне завдання нелінійної оптимізації. З погляду кібернетики, нейронна мережа використовується в завданнях адаптивного керування і як алгоритми для робототехніки. З погляду розвитку обчислювальної техніки й програмування, нейронна мережа – спосіб розв'язку проблеми ефективного паралелізму. А з погляду штучного інтелекту, ИНС є основним напрямком у структурному підході по вивченню можливості побудови (моделювання) природнього інтелекту за допомогою комп'ютерних алгоритмів.

Нейронні мережі не програмуються у звичному змісті цього слова, вони навчаються. Можливість навчання – одне з головних переваг нейронних мереж перед традиційними алгоритмами. Технічно навчання полягає в знаходженні коефіцієнтів зв'язків між нейронами. У процесі навчання нейронна мережа здатна виявляти складні залежності між вхідними даними й вихідними, а також виконувати узагальнення. Це значить, що, у випадку успішного навчання, мережа зможе повернути вірний результат на підставі даних, які були відсутні в навчальній вибірці, а також неповних і/або «зашумлених», частково перекручених даних.

Персептрон.Перші персептрони були створені Ф.Розенблатом у 60-х роках і викликали великий інтерес. Первинна ейфорія змінилася розчаруванням, коли виявилося, що персептрони не здатні навчитися рішенню ряду простих задач. М.Мінський строго проаналізував цю проблему і показав, що є жорсткі обмеження на те, що можуть виконувати одношарові персептрони, і, отже, на те, чому вони можуть навчатися. Оскільки у той час методи навчання багатошарових мереж не були відомі, дослідники перейшли в більш багатообіцяючі області, і дослідження у області нейронних мереж прийшли в занепад. Відкриття методів навчання багатошарових мереж більшою мірою, ніж який-небудь інший чинник, вплинуло на відродження інтересу і дослідницьких зусиль.

Навчання персептрона:

- Ініціалізація вагових матриць W (випадкові значення)

- Подати вхід X і обчислити вихід Y для цільового вектора YT
- Якщо вихід правильний – перейти на крок 4; інакше обчислити різницю D = YT – Y; модифікувати ваги за формулою:

wij(e+1) = wij(e) + α D Xi,

де wij(e) – значеннявагивіднейрона і донейрона j доналагодження, wij(e+1) – значеннявагипісляналагодження, α – коефіцієнтшвидкостінавчання, Xi – вхіднейрона i, e – номерепохи (ітераціїпідчаснавчання).

- Виконувати цикл з кроку 2, поки мережа не перестане помилятися. На другому кроці у випадковому порядку пред'являються всі вхідні вектори.

Один з найпесимістичніших результатів Мінського показує, що одношаровий персептрон не може відтворити таку просту функцію, як ВИКЛЮЧАЄ АБО. Це функція від двох аргументів, кожний з яких може бути нулем або одиницею. Вона приймає значення одиниці, коли один з аргументів рівний одиниці (але не обидва).

Багатошарові нейронні мережі.Багатошарові (1986 р.) мережі володіють значно більшими можливостями, ніж одношарові. Проте багатошарові мережі можуть привести до збільшення обчислювальної потужності в порівнянні з одношаровими лише в тому випадку, якщо активаційна функція між шарами буде нелінійною. Обчислення виходу шару полягає в множенні вхідного вектора на першу вагову матрицю з подальшим множенням (якщо відсутня нелінійна активаційна функція) результуючого вектора на другу вагову матрицю (XW1)W2. Оскільки множення матриць асоціативне, то X(W1W2). Це показує, що двошарова лінійна мережа еквівалентна одному шару з ваговою матрицею, рівною добутку двох вагових матриць. Отже, для лінійної активіаційної функції будь-яка багатошарова лінійна мережа може бути замінена еквівалентною одношаровою мережею.

2.1.4. Метод опорних векторів

Метод опорних векторів (англ. SVM,) належить до групи граничних методів; визначає класи за допомогою меж просторів. Опорними векторами вважаються об’єкти множини, що лежать на цих межах. Класифікація вважається вдалою, якщо простір між межами – пустий.

Метод опорних векторів – це набір схожих алгоритмів виду «навчання із вчителем». Ці алгоритми зазвичай використовуються для задач класифікації та регресійного аналізу. Метод належить до розряду лінійних класифікаторів. Також може розглядатись як особливий випадок регуляризації за А. Н. Тихоновим. Особливою властивістю методу опорних векторів є безперервне зменшення емпіричної помилки класифікації та збільшення проміжку. Тому цей метод також відомий як метод класифікатора з максимальним проміжком. Основна ідея методу опорних векторів – перевід вихідних векторів у простір більш високої розмірності та пошук роздільної гіперплощини з максимальним проміжком у цьому просторі. Дві паралельні гіперплощини будуються по обидва боки гіперплощини, що розділяє наші класи. Роздільною гіперплощиною буде та, що максимізує відстань до двох паралельних гіперплощин. Алгоритм працює у припущенні, що чим більша різниця або відстань між цими паралельними гіперплощинами, тим меншою буде середня помилка класифікатора.

2.2. Навчання моделей

Навчання моделей здійснюється на основі вибірки даних. Після закінчення навчання отриману модель потрібно перевірити на те, як добре вона виконує прогнозування.

Щоб навчити модель та перевірити її потрібно всю вибірку даних розбити на навчальну та перевірочну і в залежності від методу перевірки оцінити модель.

Для навчання моделі та її перевірки використовують крос-валідацію. Вона включає розбиття набору даних на частини, потім будуються моделі на

одній частині (тренувальна вибірка), та валідації моделі на іншій частині (тестова вибірка). Щоб зменшити розкид результатів, різні цикли крос-валідації яхпроводяться на різних розбитті, а результати валідації усереднюються по всіх циклах.

Крос-валідація важлива для захисту від гіпотез, нав'язаних даними («помилки третього роду»), особливо коли отримання додаткових даних важко або неможливо.

Припустимо, що ми навчаємо модель на основі тренувальних даних. В результаті отримали модель з оптимізованими параметрами під тренувальну вибірку. Вона дозволяє дуже чітко спрогнозувати результат для даних з цієї вибірки, але якщо взяти дані на яких вона не тренувалась вона спрогнозує їх гірше. Ця проблема називається перенавчанням (overfitting),і особливо часто зустрічається в ситуаціях, коли розмір тренувального набору невеликий, або коли число параметрів в моделі велике. Крос-валідація це спосіб оцінити здатність моделі працювати на гіпотетичному тестовому наборі, коли такий набір в явному вигляді отримати неможливо.

Існують наступні типи крос-валідації.

Крос-валідація по К блоках (K-foldcross-validation)

Візуально її можна описати так (рис. 2.3.1)

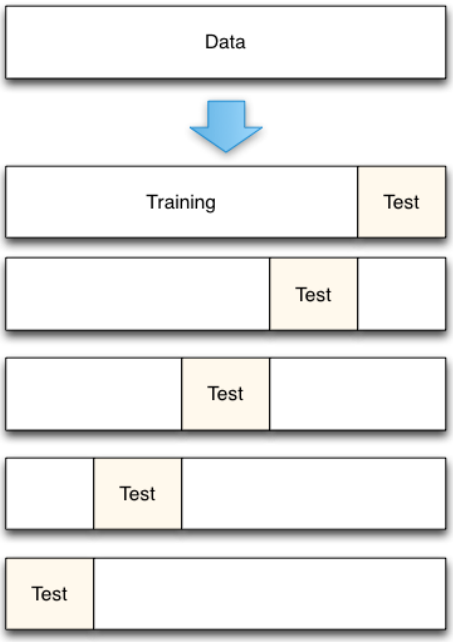

Рис. 2.3.1. Крос-валідація по K блоках

У цьому випадку вихідний набір даних розбивається на K однакових за розміром блоків. З K блоків один залишається для тестування моделі, а решта K-1 блоків використовуються як тренувальний набір. Процес повторюється K разів, і кожен з блоків використовується один раз як тестовий набір. Отримуємо K результатів, по одному на кожен блок, вони усереднюються або комбінуються яким-небудь іншим способом, і дають одну оцінку. Перевага такого способу перед випадковим розбиттям (random subsampling) в тому, що всі спостереження використовуються і для тренування, і для тестування моделі, і кожне спостереження використовується для тестування в точності один раз. Часто використовується крос-валідація на 10 блоках, але якихось певних рекомендацій щодо вибору числа блоків немає.

Валідація послідовним випадковим розбиттям (random subsampling)

Цей метод випадковим чином розбиває набір даних на тренувальний і тестовий набори. Для кожного такого розбиття, модель підганяється під тренувальні дані, а точність передбачення оцінюється на тестовому наборі. Результати потім усереднюються по всьому розбиття. Перевага такого методу перед крос-валідацією на K блоках в тому, що пропорції тренувального та тестового набору не залежать від числа повторень (блоків). Недолік методу в тому, що деякі спостереження можуть жодного разу не потрапити в тестовий набір, тоді як інші можуть потрапити в нього більше, ніж один раз. Іншими словами, тестові набори можуть перекриватися. Крім того, оскільки розбиття проводяться випадково, результати будуть відрізнятися в разі повторного аналізу.

Поелементна крос-валідація (Leave-one-out, LOO)

Тут окреме спостереження використовується в якості тестового набору даних, а решта спостережень з вихідного набору – в якості тренувального. Цикл повторюється, поки кожне спостереження не буде використане один раз в якості тестового. Це те ж саме, що і K-блокова крос-валідація, де K дорівнює числу спостережень у вихідному наборі даних.

2.3. Порівняння моделей

Для аналізу якості моделей і встановлення найкращої моделі для вирішення певної задачі використовують кілька критеріїв для оцінки адекватності моделей:

1.Загальна точність моделі.

2.Помилки I-го і II-го роду.

3.ROC-крива і індекс GINI.

Загальна точність моделі (CA – commonaccuracy) визначається як:

$$CA = \frac{CorrectForecast}{N}, \quad (2.3.1.)$$

де CorrectForecast – кількість вірно прогнозованих випадків, N – загальна кількість випадків. Загальна точність моделі є дещо суб'єктивною оцінкою, оскільки вона залежить від долі дефолтів в моделі, а також від порога відсікання. Для різних значень порога точність моделі також буде приймати різні значення.

ROC-крива (ReceiverOperationCharacteristic – робоча характеристика приймача) показує залежність кількості вірно класифікованих позитивних прикладів від кількості невірно класифікованих негативних прикладів. У термінології ROC-аналізу перші називаються істинно позитивним, другі – помилково негативним безліччю. При цьому передбачається, що у класифікатора є деякий параметр, варіюючи який, ми будемо отримувати те чи інше розбиття на два класи. Цей параметр часто називають порогом, або точкою відсікання (cut-offvalue). Залежно від нього будуть виходити різні величини помилок I і II роду.

Розглянемо таблицюспряженості2.2.1 (confusionmatrix), яка будується на основі результатів класифікації моделлю і фактичної (об'єктивної) приналежністю прикладів до класів.

	Прогноз моделі. Позитивний	Прогноз моделі. Негативний
Хворий	Вірно класифіковані (TP)	Помилки II-го роду (FP)
Здоровий	Помилки I-го роду (FN)	Вірно класифіковані (TN)

Таблиця 2.2.1. Таблиця спряженості

- TP (TruePositives) – вірно класифіковані позитивні приклади (так звані істинно позитивні випадки);
- TN (TrueNegatives) – вірно класифіковані негативні приклади (істинно негативні випадки);
- FN (FalseNegatives) – позитивні приклади, класифіковані як негативні (помилка I роду). Це так званий «помилковий пропуск» – коли цікавить нас помилково не виявляється (хибно негативні приклади);
- FP (FalsePositives) – негативні приклади, класифіковані як позитивні (помилка II роду); Це помилкове виявлення, тому що за відсутності події помилково виноситься рішення про його присутність (хибно позитивні випадки).

Для аналізу властивостей моделі найчастіше використовують такі відносні показники у відсотках:

- частка істинно позитивних прикладів (TruePositivesRate):

$$TPR = \frac{TP}{TP + FN}; \quad (2.3.2.)$$

- частка хибно позитивних прикладів (FalsePositivesRate):

$$FPR = \frac{FP}{TN + FP}. \quad (2.3.3.)$$

Зазвичай для моделей визначають ще дві характеристики: чутливість, специфічність і точність.

Чутливість моделі – це частка істинно позитивних випадків, тобто

$$Se = TPR = \frac{TP}{TP + FN}. \quad (2.3.4.)$$

Специфічність моделі – це частка істинно негативних випадків, які були правильно класифіковані моделлю:

$$Sp = \frac{TN}{TN + FP}. \quad (2.3.5.)$$

Точність моделі (Accuracy) – це відношення вірно класифікованих випадків до всіх.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.3.6.)$$

Повнота моделі (Recall) – показує скільки від загального числа реальних позитивних об'єктів було прогнозовано як позитивний клас.

$$Recall = \frac{TP}{TP + FN} \quad (2.3.7.)$$

Для побудови графіка ROC-кривої по осі Y відкладаються значення чутливості, а по осі X – 1 >Sp Графік ідеального класифікатора ROC-кривої проходить через верхній лівий кут, де частка істинно позитивних випадків становить 1 (ідеальна чутливість), а частка хибно позитивних прикладів дорівнює нулю. Тому чим ближче крива наближається до верхнього лівого кута, тим краще є здатність моделі пророкувати. Діагональна лінія відповідає класифікатору, який здатний розпізнати ці два класи. Оскільки візуальне порівняння ROC-кривих не завжди дозволяє визначити ефективну модель, застосовується оцінка площі під кривими. Чисельний показник площі під кривою AUC (AreaUnderCurve) обчислюється, наприклад, за методом трапецій:

$$AUC = \int f(x)dx = \sum_i \left[\frac{X_{i+1} + X_i}{2} \right] \cdot (Y_{i+1} - Y_i). \quad (2.3.8.)$$

Більш зрозумілим і частіше згадуємим в літературі параметром оцінки якості моделі є індекс GINI, який тісно пов'язаний з чисельним показником площі під ROC-кривої.

Діапазон значень індексу GINI становить 0 <G< 1, а моделі з найбільшою роздільною здатністю, тобто моделі, які виконують високоякісне сортування схильних до дефолту клієнтів та клієнтів, які не схильні до дефолту, отримують найвищі коефіцієнти. На практиці оцінка якості моделі значно залежить від даних, на яких вона будується. Для застосування на практиці скорингу (оцінки фінансового становища існуючих клієнтів) індекс GINI в основному приймає значення вище 70%. Приведено шкалу значень індексу (таблиця 2.2.2).

Інтервал AUC	Індекс GINI	Якість моделі
0.9-1.0	0.8-1.0	Відмінне
0.8-0.9	0.6-0.8	Дуже висока
0.7-0.8	0.4-0.6	Прийнятна
0.6-0.7	0.2-0.4	Середня
0.5-0.6	0-0.2	Незадовільна

Таблиця 2.2.2. Оцінка якості моделі за площею AUC і за індексом GINI

Значення точок ROC-кривої можуть бути використані для знаходження оптимального порогу відсікання – компромісу між чутливістю і специфічністю моделі. Критеріями вибору порогу відсікання можуть бути:

1.Вимога мінімальної величини чутливості (специфічності) моделі.

2.Вимога максимальної сумарної чутливості і специфічності моделі, тобто.

$$cut-off = \max_k (Se_k + Sp_k); \quad (2.3.9.)$$

3.вимога балансу між чутливістю і специфічністю, коли Sp>Se:

$$cut-off = \min_k |Se_k - Sp_k|. \quad (2.3.10.)$$

2.4.Висновок до розділу

Для обробки даних та побудові на їх основі моделі використовуватимуться методи машинного навчання:

- Логістична регресія
- Дерево рішень
- Метод опорних векторів
- Нейронна мережа

Для навчання моделей та перевірці її якості використовуватиметься метод крос-валідації К блоками та валідація випадковим послідовним розбиттям.

Для порівняння моделей між собою використовуватиметься критерій AUC.

Використання вищеописаних методів є загальним підходом в вирішенні подібних задач.

3.ОБРОБКА ДАНИХ

Як було сказано вище новизною даного дослідження є обробка пар чоловік-дружина а не чоловіка та дружини окремо. Тобто один логічний запис представляється набором однакових атрибутів для чоловіка та жінки. Отримані дані з інституту НМАПО ім. П. Л. Шупіка були представлені в наступному вигляді рис. 3.1.

id	Стать	PAI	MTHFR_C	MTHFR_A	FV	FII	MTRR_A6	FGB_C148	FGB_455C	ITGA2a2	ITGB3b1	Result
	ч	5G/5G	CT	AC	GG	GG	AG	CT	GA	CC	TT	1
1	ж	5G/4G	CC	AC	GG	GG	AG	CC	GG	CC	TT	1
	ч	5G/5G	CC	AC	GG	GG	AA	CC	GG	CC	TC	1
2	ж	5G/5G	CC	AC	GG	GG	AA	CC	GA	CC	TT	1
	ч	5G/4G	CC	AC	GG	GG	AA	CC	GA	CC	TC	0
3	ж	5G/4G	CT	AA	GG	GG	AA	CC	GG	CC	TC	0

Рис. 3.1. Генетичні дані пар чоловік-дружина.

На рисунку3.1"id" – унікальний номер пари;"Result" – результат який вказує на наявність (1) чи відсутність безпліддя (0);PAI, MTHFR_C, MTHFR_A, FV та інші – генотипи.

Якщо використовувати таке представлення даних ми будемо досліджувати чоловіка та дружину окремо, а це не є метою даної роботи. У такому випадку потрібно перейти до іншого вигляду. Нижче приведений приклад переходу від незалежних даних по чоловіку та жінки до вигляду у парі (рис. 3.2.).

Стать	PAI	MTH	Result
ч	5G/5G	CT	1
ж	5G/4G	CC	1
ч	5G/5G	CC	0
ж	5G/5G	CC	0

M_PAI	M_MTH	F_PAI	F_MTH	Result
5G/5G	CT	5G/4G	CC	1
5G/5G	CC	5G/5G	CC	0

Рис. 3.2. Приклад конвертації даних

Таким чином кількість незалежних між собою атрибутів зросла в два рази.

Конвертована вибірка представлена записами з 230 подружніх пар.

Для дослідження та оцінки вибірок використовувався сервіс MicrosoftAzureMachineLearning.

Microsoft Azure Machine Learning візуальне середовище розробки, яке дозволяє будувати, тестувати і розгортати розроблені рішення.Сервіс є хмарною технологією, тобто всі збережені дані та обчислювальні ресурси знаходяться не на локальній машині клієнта, а на сервері до якого здійснюється доступ через браузер. Цей сервіс використовує останні алгоритми машинного навчання, тому хвилюватись за їх якість не потрібно.

Середовище розробки Machine Learning має весь необхідний функціонал для завантаження, збереження, маніпуляції даними, алгоритми навчання та можливості використання зовнішніх скриптів написаних на мовах програмування Pythonта R.

Даний сервіс поставляє наступні алгоритми, частину з яких будемо використовувати в дослідженні (рис.3.3.)

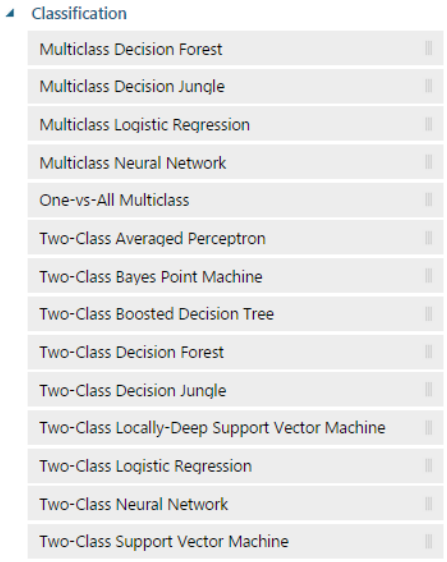

Рис. 3.3. Класифікаційні алгоритми Machine Learning

3.1. Обробка генетичних даних.

Генетичні дані складаються з 20 факторів. 10 з них належать чоловіку та інші 10 – жінці. Кожен фактор це генотип, що представляєтьсякатегоріальним типом даних і може приймати три різних значення. Всього 230 записів.

Для дослідження на основі генотипів будемо використовувати наступні алгоритми:

- Логістична регресія (Logistic Regression)
- Ліс рішень (Decision Forest)
- Нейронна мережа (Neural Network)
- Метод опорних векторів (Support Vector Machine)

Логістична регресія

Для навчання та оцінки моделі використаємо два методи валідації моделі:

- Крос-валідація по К блоках
- Валідація послідовним випадковим розбиттям

Весь процес з моменту взяття даних, їх обробки, навчання моделі та перевірки кості можна виконати в середовищі розробки MicrosoftAzureMachineLearningi виглядатиме це наступним чином для крос-валідації по К блоках (рис. 3.1.1).

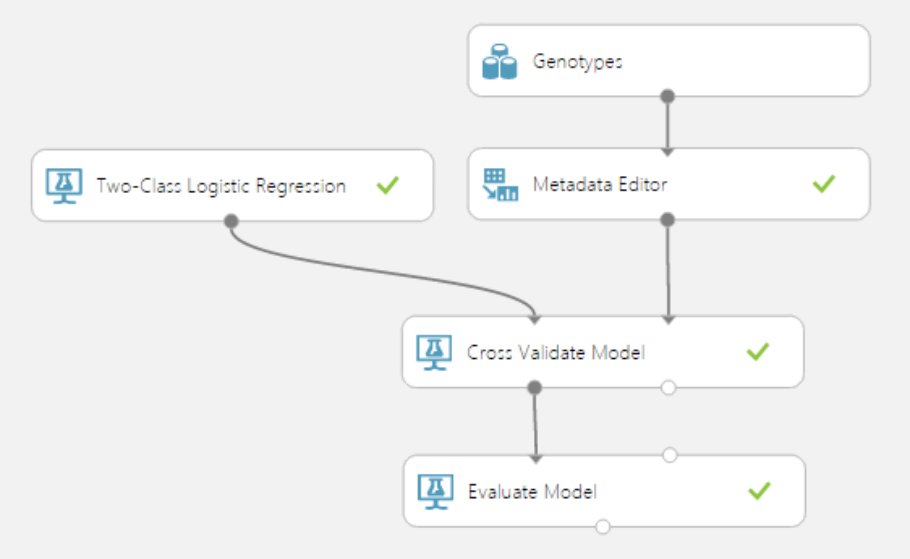

Рис. 3.1.1. Схема навчання та перевірки моделі

Результатом виконання схеми є отримання матриці спряженості, яка показує вірність класифікації моделі(таб. 3.1.1.)

	Прогноз моделі. Позитивний	Прогноз моделі. Негативний
Безпліддя	122	9
Не безпліддя	11	89

Таблиця.3.1.1. Матриця спряженості результатів класифікації

Отже в результаті класифікації з використанням навченої моделі отримали 122 вірно класифікованих пар з безпліддям (TP – TruePositive). 89 – вірно класифікованих пар з відсутністю безпліддя (можуть завагітніти) (TN – TrueNegative). 9 – пар, які помилково класифіковані, як пари з безпліддям. (FP – FalsePositive). 11 – пар, що не можуть завагітніти були класифіковані як ті, що можуть (FalseNegative).

На основі цих даних визначається чутливість та точність моделі. Чутливість (SE)моделі дорівнює 0.931, повнота(Recall)дорівнює 0.931, точність (ACC)дорівнює 0.913. Потім будується ROC-крива (рис. 3.1.2.) та визначається AUCкритерій.

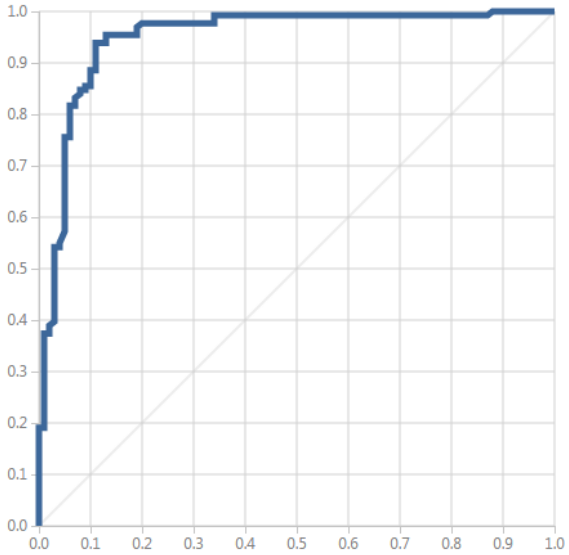

Рис. 3.1.2. ROC-крива оцінки моделі

AUCдорівнює 0.949, а отже якість моделі "відмінна" за шкалою оцінювання.

Крос-валідація по К блоках виконує 10 циклів. Тобто 10 разів вибірка розбивалась на тренувальну та тестову, навчалась і перевірялась. При цьому кожні 10 разів тестова вибірка включала в себе дані, що не повторювались в інших тестових вибірках. Для кожного циклу визначалися оцінки моделі (чутливість, точність та критерій AUC). В таблиці 3.1.2. наведені номера моделей та критерій AUC побудованих на тестових даних.

Номер моделі	Критерій AUC
1	0.987
2	0.992
3	0.892
4	0.968
5	0.892
6	0.969
7	0.991
8	0.911
9	0.953
10	0.876

Таблиця.3.1.2. Точність моделей

Середнє значення критерію AUCдорівнює 0.945.

Більше уваги при оцінюванні моделі потрібно приділяти критеріям оцінки отриманих на основі тестових даних, а не на основі навчальних і тестовихданих разом, тому що тестові дані, як і дані що поступатимуть на вхід цієї моделі при її експлуатації, не входять до навчальної вибірки. І основуючись на цьому критерію можна бути певним, що нові подані дані на вхід моделі будуть оцінені з точністю як тестові.

Якщо використовувати валідацію послідовним випадковим розбиттям схема матиме наступний вигляд (рис. 3.1.3).

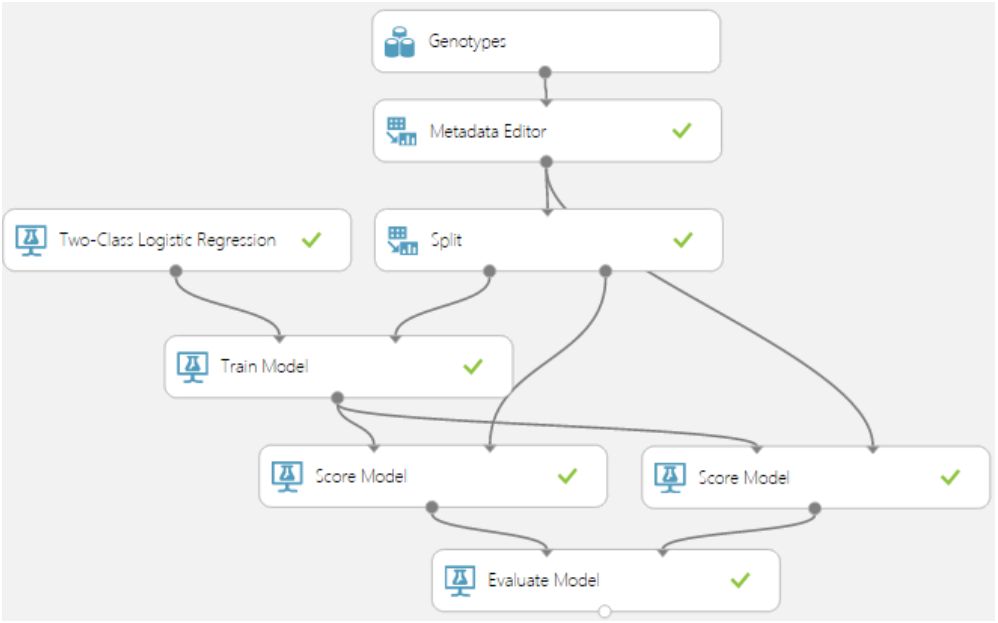

Рис. 3.1.3. Схема валідації послідовним випадковим розбиттям

На схемі представлений блок "Split", він розбиває вибірку на навчальну та тестову. В блоці вказується відсоток тренувальної вибірки, потім метод випадково обирає записи.

Для перевірки здійснимо чотири розбивання, тобто виконаємо чотири цикли навчання. Спочатку для тренувальної вибірки виділимо 60% записів, для другого, третього та четвертого 70%, 80% та 90% відповідно.

Після розбиття модель навчається на тренувальній вибірці. Якість моделі порівнюватимемо на основі тестової вибірки та всієї вибірки загалом.

Якість моделей наведена в таблиці 3.1.3.

Частина тренувальної вибірки	**Критерій AUC тестової вибірки**	**Критерій AUC всієї вибірки**
60%	0.934	0.964
70%	0.919	0.966
80%	0.893	0.966
90%	0.902	0.972

Таблиця.3.1.3. Якість моделей при різному розбитті вибірки

Найкращий результат отриманий при розбитті вибірки на 60% для тренувальної та 40% для тестової. Помітно, що зі зростанням частки тренувальної вибірки зростає і якість оцінки моделі по всім даним і водночас супроводжується падінням якості основаної на тестовій вибірці. Це спричинено перенавчанням моделі на тренувальних даних, тобто модель добре класифікує записи на яких навчалася, але погано класифікує нові.

Ліс рішень

Для цього методу також використаємо два види валідації.

Крос-валідація по К блоках.

Якість моделей при крос-валідації оцінюватимемо за допомогою критеріїв AUCотриманий по всій вибірці та по тестових. Також за допомогою характеристик: чутливість, повнота, точність.

Критерії AUCотримані при оцінці моделі на тестових вибірках наведені у таблиці 3.1.4.

Номер моделі	Критерій AUC
1	0,988
2	0.976
3	0.988
4	0.908
5	0.977
6	0.958
7	0.982
8	0.977
9	0.926
10	0.803

Таблиця.3.1.4. Точність моделей

Критерій AUCпри оцінці по всій вибірці дорівнює 0.945. Середня значення критерію AUCпо тестових вибірках дорівнює 0.952. Чутливість (SE)моделі дорівнює 0.925, повнота(Recall)дорівнює 0.939, точність (ACC)дорівнює 0.922.

Валідація послідовним випадковим розбиттям.

Розіб'ємо вибірку на тренувальну та тестову чотири рази. Тренувальна вибірка складатиме 60%, 70%, 80%, 90% від всіх записів.

Порівняння якості моделей приведена в таблиці 3.1.5.

Частина тренувальної вибірки	Критерій AUC тестової вибірки	Критерій AUC всієї вибірки
60%	0.919	0.968
70%	0.932	0.986
80%	0.877	0.977
90%	0.924	0.994

Таблиця.3.1.5. Якість моделей при різному розбитті вибірки

Найкращим розбиття вибірки для цього методу стало відношення 70% для тренувальної вибірки та 30% для тестової.

Метод опорних векторів

Для цього методу також використаємо два види валідації.

Крос-валідація по К блоках.

Якість моделей при крос-валідації оцінюватимемо за допомогою критеріїв AUCотриманий по всій вибірці та по тестових. Також за допомогою характеристик: чутливість, повнота, точність.

Критерії AUCотримані при оцінці моделі на тестових вибірках наведені у таблиці 3.1.6.

Номер моделі	**Критерій AUC**
1	0.986
2	0.984
3	0.85
4	0.944
5	0.883
6	0.969
7	0.978
8	0.911
9	0.930
10	0.869

Таблиця.3.1.6. Точність моделей

Критерій AUCпри оцінці по всій вибірці дорівнює 0.932. Середня значення критерію AUCпо тестових вибірках дорівнює 0.934. Чутливість (SE)моделі дорівнює 0.896, повнота(Recall)дорівнює 0.924, точність (ACC)дорівнює 0.896.

Валідація послідовним випадковим розбиттям.

Розіб'ємо вибірку на тренувальну та тестову чотири рази. Тренувальна вибірка складатиме 60%, 70%, 80%, 90% від всіх записів.

Порівняння якості моделей приведена в таблиці 3.1.7.

Частина тренувальної вибірки	**Критерій AUC тестової вибірки**	**Критерій AUC всієї вибірки**
60%	0.928	0.956
70%	0.925	0.957
80%	0.891	0.947
90%	0.864	0.96

Таблиця.3.1.7. Якість моделей при різному розбитті вибірки

Найкращим розбиття вибірки для цього методу стало відношення 60% для тренувальної вибірки та 40% для тестової.

Нейронна мережа

Для цього методу також використаємо два види валідації.

Нейронна мережа по замовчуванню містить наступні налаштування: один прихований шар зі 100 прихованими нейронами, крок навчання дорівнює 0.1,

100 ітерацій на навчання, для нормалізації входів використовується мінімаксна нормалізація.

Крос-валідація по К блоках.

Якість моделей при крос-валідації оцінюватимемо за допомогою критеріїв AUCотриманий по всій вибірці та по тестових. Також за допомогою характеристик: чутливість, повнота, точність.

Критерії AUCотримані при оцінці моделі на тестових вибірках наведені у таблиці 3.1.8.

Номер моделі	**Критерій AUC**
1	0.984
2	1
3	0.9
4	0.96
5	0.785
6	0.969
7	0.910
8	0.903
9	0.907
10	0.834

Таблиця.3.1.8. Точність моделей

Критерій AUCпри оцінці по всій вибірці дорівнює 0.902. Середня значення критерію AUCпо тестових вибірках дорівнює 0.915. Чутливість (SE)моделі дорівнює 0.869, повнота(Recall)дорівнює 0.863, точність (ACC)дорівнює 0.848.

Валідація послідовним випадковим розбиттям.

Розіб'ємо вибірку на тренувальну та тестову чотири рази. Тренувальна вибірка складатиме 60%, 70%, 80%, 90% від всіх записів.

Порівняння якості моделей приведена в таблиці 3.1.9.

Частина тренувальної вибірки	**Критерій AUC тестової вибірки**	**Критерій AUC всієї вибірки**
60%	0.851	0.937
70%	0.875	0.961
80%	0.837	0.964
90%	0.928	0.994

Таблиця.3.1.9. Якість моделей при різному розбитті вибірки

Найкращим розбиття вибірки для цього методу стало відношення 90% для тренувальної вибірки та 10% для тестової.

Порівняння якості моделей

Порівняємо отримані результати для двох видів валідації та всіх методів (таб. 3.1.10). Порівнюватимемо AUCпо всій вибірці, середнє AUCпо тестових вибірках отриманих при крос-валідації К блоками, середнім та максимальними значеннями AUCпо тестових вибірках при валідації послідовним випадковим розбиттям.

Метод	AUCвся вибірка К блоками	AUCсереднє К блоками	AUCсереднє послідовна валідація	AUCmax послідовна валідація
Логістична регресія	0.949	0.945	0.912	0.934
Ліс рішень	0.945	0.952	0.913	0.932
Метод опорних векторів	0.932	0.934	0.902	0.928
Нейронна мережа	0.902	0.915	0.872	0.928

Таблиця.3.1.10. Порівняння якості моделей

Порівнюючи методи валідації, крос-валідація К блоками дає меншу розбіжність між оцінкою моделі перевірену по всій вибірці та тестовими вибірками, коли при використанні послідовного випадкового розбиття вибірки при оцінці на всіх даних помітне перенавчання моделі. Однак метод послідовного випадкового розбиття при певному поділіна тренувальну та тестову вибірки дає непогану оцінку, що дозволяє її теж використовувати.

В подальшому дослідженні використовуватимемо крос-валідацію К блоками, тому що оцінка по всій вибірці дає майже той самий результат, що й при середньому значенні тестових.

Якщо порівнювати методи навчання моделі, то найкращими виявилися логістична регресія та ліс рішень, їхня оцінка майже однакова. Гіршим від них став метод опорних векторів та найгірший результат у нейронної мережі.

Однак за шкалою точності всі моделі мають "відмінну" якість, тому кожен метод впорався з побудовою якісної моделі.

Отримані моделі мають високий показник точності, а отже можна зосередитися тільки на їх використанні. Однак, як відомо, безпліддя це мультифакторне захворювання, а отже його причиною є не тільки генотипи вказані в дослідженні. Надзвичайно важливим є генеалогічна інформація поколінь. Наприклад, якщо в третьому та другому поколінні були проблеми з вагітністю, то є вірогідність того, що в цьому поколінні також будуть проблеми. Саме цей зв'язок потрібно досліджувати.

3.2. Обробка генеалогічних даних.

Медичнідані старших поколінь є надзвичайно важливою для наступних. Однак, щоб довести цю важливість потрібно їх значна кількість, іншими словами – велика кількість аналізів та обстежень кожного члена родини протягом їхнього життя. В такому випадку можна було б проводити дослідження основуючись тільки на даних родини і при цьому отримувати точні результати.

В цьому досліджені генеалогічні дані далеко від ідеальних. Це спричинено тим, що не має єдиної бази пацієнтів в якій би зберігалася інформація про сім'ї та її покоління.

Однак для аналізу були виділені дані з найважливішими факторами, які впливають на безпліддя. Тобто, у випадку, якщо хворів хтось з старших поколінь, то до запису пацієнта додавалась інформація про його недуги.

Набір генеалогічних даних представляють собою вибірку з22 незалежних факторів та одного результуючого. Результатом являється наявність чи відсутність безпліддя.

Якщо звернути увагу на дані, то стає помітна мультиколінеарність між деякими колонками (наприклад, інфаркти та число інфарктів в родині).

Мультиколінеарність – цеявище існування тісної лінійної залежності, або сильної кореляції, між двома або більше незалежними змінними.

В цьому випадку потрібно позбутися мультиколінеарності. Відомі наступні методи, що дозволяють це зробити:

- отримання додаткових даних чи нової вибірки;
- зміна специфікації моделі;
- виключення змінної з моделі;
- використання попередньої інформації про деякі параметри;
- перетворення змінних;
- метод покрокового включення;
- гребенева регресія (ridge).
- метод головних компонентів

В даному випадку використаємо виключення змінної з моделі.

Уважно розглянувши фактори стає зрозумілим, між якими саме існує мультиколінеарність. Наприклад, це спостерігається для пар факторів "інфаркти" та "число інфарктів в родині", "інсульти" та "число інсультів в родині" і ще для шести пар. Якщо подивитись на самі дані, то пара завжди заповнена одним значенням і тому одну з колонок потрібно виключити, щоб уникнути дубльованих значень.

Після виключення дубльованих факторів залишається 14. Можна виключити "Спадковість", тому що вона завжди дорівнює "так" якщо значення "Ступінь спорідненості" відрізняється від "немає" і дорівнює "ні" якщо "Ступінь спорідненості" дорівнює "немає".

Залишається ще одна мультиколінеарність. Це фактор "загальне число захворювань у родичів", який являє собою суму захворювань. Тому потрібно визначити, що залишити для аналізу "загальне число захворювань у родичів" і при цьому забрати фактори "інфаркти ", "інсульти", "тромбоемболії " та ін. чи залишити 8 факторів та опустити "загальне число захворювань у родичів". В даному випадку приберемо колонку "загальне число захворювань у родичів".

На цьому операції над вхідними даними закінчуються і виникає наступна проблема. Цезначна кількість відсутніх даних. Вони відсутні в колонках з числовим значенням, в яких записується число недуг. Наприклад, в колонці "інсульти" аж 407 пропущених значень, в інших колонках ще більше.

Існують різні методи заміни відсутніх даних: вставка середнього значення, медіани, моди чи якогось конкретного значення. В нашому випадку найкраще підійде заміна відсутнього значення нулем. Тому що при огляді пацієнта зазвичай вказуються лише хвороби на які хворіли, і не вказуються хвороби на які не хворіли. Замінивши всі відсутні значення отримаємо розріджену матрицю. Вона не дозволяє провести повноцінне статистичне дослідження і отримати адекватну статистичну оцінку впливу генеалогічної інформації поколінь. В такому випадку можливо отримати лише певну величину впливу спадковості на пари.

Використання таких методів якзвичайну логістичну регресію, дерево рішень чи метод опорних векторів не зможуть ефективно обробити вибірки з значною кількістю відсутніх даних. Тому потрібно використовувати алгоритми, що працюють з розрідженими матрицями. Саме для такої ситуації використаємо спеціальнийалгоритм логістичної регресії.

3.2.1. Дослідження чоловіка та жінки окремо.

Сенс дослідження пари окремо полягає в тому, щоб визначити наскільки персонально обтяжений родовід впливає на безпліддя пари.

Вибірка на якій буде проводитись дослідження це генеалогічна частина даних наданих лікарями. Оскільки попередньо досліджували пари, то вибірка містила 230 записів, тепер її число становить 460. Кількість незалежних змінних дорівнює 12.

Вибірка містить багато нулів отже потрібно застосувати методи навчання моделей, які здатні працювати з розрідженими матрицями.

Спочатку спробуємо за допомогою звичайної логістичної регресії, лісу рішень та методу опорних векторів побудувати модель. Для цього продовжимо використовувати сервіс MicrosoftAzureMachineLearning.

Побудована модель(рис. 3.2.1.1.) дозволяє отримати результат відразу по трьох методах. Для навчання та перевірки моделі використовується крос-валідація К блоками.

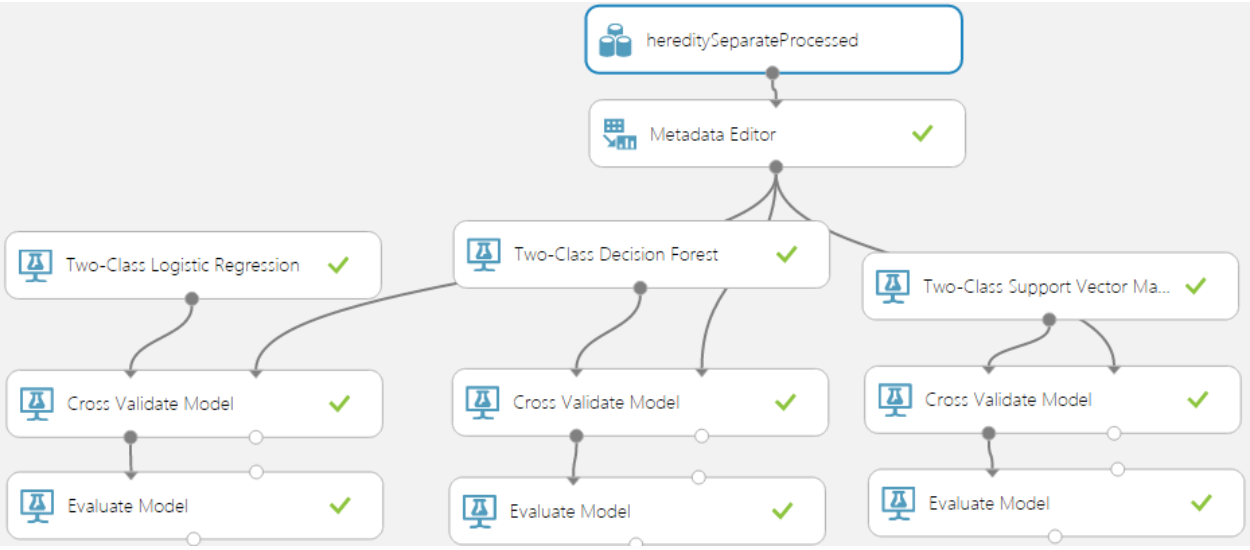

Рис. 3.2.1.1. Отримання оцінок якості моделей

В таблиці 3.2.1.1. наведені оцінки кожної моделі критерієм AUC

Назва методу	**Критерій AUC**
Логістична регресія	0.574
Ліс рішень	0.530
Метод опорних векторів	0.565

Таблиця.3.2.1.1. Оцінка моделей

Результат для всіх трьох моделей лежить в проміжку 0.5-0.6, що означає результат є незадовільним і в ніякому разі ці моделі не можуть бути використані для оцінки.

Це результати отримані внаслідок використання стандартних алгоритмів. Тепер використаємо лише один алгоритм логістичної регресі розроблений для обробки розріджених матриць.

Даний дослід потрібно виконати в програмному середовищі R попередньоінсталювати пакет gmlnet, що містить в собі алгоритми для роботи з розрідженими масивами.

В додатку Є приведений скрипт цієї програми.

Для навчання та перевірки цієї моделі також була використана крос-валідація К блоками. Власноруч було задано 10 циклів крос-валідації. Після навчання автоматично обирається найкраща модель і використовується для аналізу всіх даних.

В результаті використання алгоритму для роботи з розрідженими матрицями отримали критерій AUCрівний 0.649. Дана оцінка моделі вказує, що модель має середню прогнозуючу здатність.

Можна зробити висновок, що алгоритм для роботи з розрідженими матрицями набагато краще підходить для роботи з генеалогічними даними. Однак показник AUCрівний 0.649 незадовільний і на його основі не варто приймати рішення.

Такий низький показник критерію спричинений специфічністю даних, а більше її відсутністю. В вибірці є записи з однаковими значеннями факторів, але результативна ознака в них різна. Тобто існує значна кількість записів, в яких відсутні хвороби у старших поколінь, однак частина з них відноситься до класу з безпліддям, а інша ні.

3.2.2. Дослідження чоловіка та жінки як пари.

Набагато доцільніше перевіряти не окремо чоловіка-дружину, а разом. Тому що так враховується результуючий вплив двох родоводів.

Дослідження проводитиметься на вибірці даних з 230 записів. Загальне число ознак-факторів дорівнює 22. Спочатку проведемо дослідження з використанням звичайних методів навчання моделі і порівняємо їх з логістичною регресією для роботи з розрідженими матрицями.

Для навчання моделі використаємо MicrosoftAzureMachineLearning.

Схема навчання моделей залишається такою самою як зображено на рисунку 3.2.1.1. змінюються тільки вхідні дані.

В таблиці 3.2.2.1. наведені оцінки кожної моделі критерієм AUC.

Назва методу	**Критерій AUC**
Логістична регресія	0.681
Ліс рішень	0.696
Метод опорних векторів	0.688

Таблиця.3.2.2.1. Оцінка моделей

Порівнюючи з результатами отриманими після навчання на основі окремо чоловіка та жінки можна зробити висновок, що досліджувати їх як пару

є більш вірним. Це дозволяє врахувати наступний момент. Коли чоловік і жінка мають не значно обтяжений родовід і досліджуючись разом в них буде виявлено безпліддя, коли при дослідженні окремо вказуватиметься на йоговідсутність.

При утворенні пар отримали ще більш розріджену матрицю. Використаємо скрипт на мові R,що наведений в додатку Є для побудови моделі.

Отримано модель з AUCрівним 0.783. Ця модель краща чим побудовані моделі на основі стандартних алгоритмів. Якість цієї моделі оцінюється як прийнятна. Однак ця величина є не достатньо високою для того, щоб використовувати її для прогнозу.

Цей показник не високий з тої ж причини, що й при дослідженні пари окремо. Існують ідентичні записи в вибірці, що відносяться до різних класів. Наявність таких записів свідчить про те, що дана кількість факторів є недостатньою.

Використання отриманих результатів.

За допомогою алгоритму з розрідженими матрицями отримали найкращу модель. ЇЇ результати прогнозування можна використати для оцінки наявності безпліддя, але враховувати її потрібно тільки як певну величину впливу. Результат прогнозування цієї моделі можна представити у вигляді катеригоріальних даних, яка вказуватиме величину впливу генеалогічної обтяженості пари.

Формування критеріїв впливу генеалогічних даних.

В кращому випадку критерій впливу повинний описувати величину впливу на безпліддя, а також величину впливу на відсутність безпліддя. Однак враховуючи принцип класифікації моделі критерій буде описувати ступінь впливу на безпліддя і не буде описувати ступінь впливу на відсутність безпліддя. Щоб вірно йогосформувати потрібно розуміти як модель відносить запис до одного чи другого класу.

Коли модель отримує прогнозоване значення вона порівнює його з певним порогом і якщо значення вище цього порогу відносить його до класу очікуваної ознаки, якщо менше то до іншого класу.

Існує декілька методів визначення порогу відсічення. Використаємо поріг на основі максимальної сумарної чутливості і специфічності моделі.

Визначений поріг дорівнює 0.401. При цьому отримуємо наступну матрицю спряженості для моделі (таб. 3.2.2.2).

	Прогноз моделі. Позитивний	Прогноз моделі. Негативний
Безпліддя	78	52
Не безпліддя	10	90

Таблиця.3.2.2.2. Матриця спряженості моделі

Чутливість (SE)моделі дорівнює 0.886, повнота (Recall)дорівнює 0.6, точність (ACC)дорівнює 0.73.

Поріг відсікання ділить прогнозовані значення на дві частини. Розглянемо записи, що знаходяться під порогом. Ці записи отримали прогнозоване значення рівним або 0.3766, або 0.3886. Тобто значення коливаються в діапазоні від 0.3766 до 0.401. А розглянувши результат вищий порогу відсікання бачимо, що мінімальне та максимальне значення в цьому діапазоні дорівнюють 0.4186 та 0.9977 відповідно. Це дозволяє розбити позитивно класифіковані результати на групи, які вказують величину впливу обтяженого родоводу на прояв безпліддя. Через те, що негативно класифіковані результати коливаються в малому діапазоні не можна створити шкалу, що вказуватиме на величину шансів завагітніти.

3.2.3. Визначення критерію впливу та перевірка його якості

Щоб визначити найкраще розбиття для критерію впливу потрібно зробити декілька шкал і порівнювати моделі побудовані на генотипах з моделями навченими з критерієм. Також потрібно зробити порівняння якості моделей якщо їй напряму передавати прогнозовані значення обтяженої спадковості.

Всі значення класифіковані як негативні будемо відносити до класу не впливовий (NoEffect). Позитивно прогнозовані значення віднесемо до певного критерії. Критерії будуть визначатися за рахунок розбиття діапазону в якому знаходяться позитивно прогнозовані значення. Так будемо ділити інтервал [0.401; 1] на певну кількість частин і знаходитимемо його інтервали.

Розбиття на 2 значення

Спочатку виконаємо найпростіше розбиття. Це покаже на скільки сильно впливає на безпліддя наявність обтяженого генотипу.

Якщо прогнозоване значення менше порогу відсікання замінюємо його на NoEffect, якщо більше на MediumEffect. Тепер додаємо цей фактор до вибірки з генотипами.

Наступним етапом є побудова та порівняння моделей. Будемо використовувати MicrosoftAzureMachineLearningi наступні методи: Логістична регресія, Метод опорних векторів, Ліс рішень та нейронну мережу. Оскільки тип даних досліджуваної вибірки для всіх факторів є категоріальний, то можна використати логістичну регресію для роботи з розрідженою матрицею (GMLNET). Це спричинено тим, що всі категоріальні дані представляються в вибірці за допомогою штучних змінних (значеннями 0 та 1), що в свою чергу утворюють розріджену матрицю.

Для порівняння моделей була побудована схема представлена на рис. 3.2.3.1

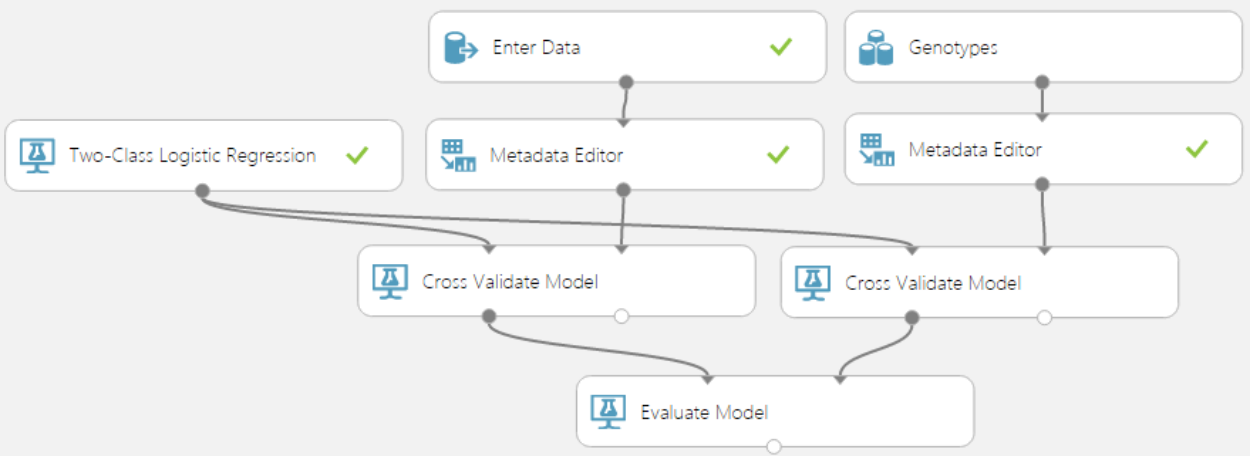

Рис. 3.2.3.1 Схема порівняння моделей

Будемо порівнювати моделі побудовані тільки на основі генотипів з моделями де крім генотипів береться критерій впливу розбитий на 2 частини.

Нижче приведена таблиця 3.2.3.1 порівняння моделей.

Метод	**AUCбез критерію**	**AUC**
Логістична регресія	0.949	0.957
Ліс рішень	0.945	0.935
Метод опорних векторів	0.932	0.933
Нейронна мережа	0.902	0.933
GMLNET	0.963	0.980

Таблиця.3.2.3.1 Порівняння моделей з розбиттям критерію впливу на 3 частини

При розбитті критерію на 2 частини отримали покращення прогнозуючої здатності для моделей основаних на логістичній регресії, нейронної мережі та GMLNET. Ліс рішень при такому розбитті критерію почав давати гіршу оцінку, а метод опорних векторів залишився майже не змінився.

Як і передбачалося GMLNET найкраще від усіх впоралось з прогнозуванням.

Розбиття на 3 значення

Інтервал розбиття: 1-0.401 = 0.599.

Величина зміщення кожного критерію: 0.599/2 = 0.2995

Отримані інтервали представлені в таблиці 3.2.3.2.

Назва інтервалу	**Інтервал**
NoEffect	менше 0.401
MediumEffect	[0.401;0.7005]
HighEffect	більше 0.7005

Таблиця.3.2.3.2. Критерії впливу при розбитті на 3 частини

Тепер прогнозований результат моделі замінюємо на значення критеріїв.

Будемо порівнювати моделі побудовані тільки на основі генотипів з моделями де крім генотипів береться критерій впливу розбитий на 3 частини.

Нижче приведена таблиця 3.2.3.3 порівняння моделей.

Метод	**AUCбез критерію**	**AUC**
Логістична регресія	0.949	0.959
Ліс рішень	0.945	0.947
Метод опорних векторів	0.932	0.934
Нейронна мережа	0.902	0.932
GMLNET	0.963	0.984

Таблиця.3.2.3.3. Порівняння моделей з розбиттям критерію впливу на 3 частини

При розбитті критерію на 3 частини отримали моделі з кращою прогнозуючою здатністю. Однак тільки для логістичної регресії, нейронної мережі та GMLNET вдалося відчутно покращити прогнозуючі властивості.

Розбиття на 4 значення

Інтервал розбиття: 1-0.401 = 0.599.

Величина зміщення кожного критерію: 0.599/3 = 0.1997

Отримані інтервали представлені в таблиці 3.2.3.4.

Назва інтервалу	**Інтервал**
NoEffect	менше 0.401
LowEffect	[0.401; 0.6007]
MediumEffect	(0.6007; 0.8004]
HighEffect	більше 0.8004

Таблиця.3.2.3.4. Критерії впливу при розбитті на 4 частини

Тепер замінимо прогнозований результат моделі на значення 4 критеріїв.

Будемо порівнювати моделі побудовані тільки на основі генотипів з моделями де крім генотипів береться критерій впливу розбитий на 4 частини.

Нижче приведена таблиця 3.2.3.5 порівняння моделей.

Метод	**AUCбез критерію**	**AUC**
Логістична регресія	0.949	0.958
Ліс рішень	0.945	0.942
Метод опорних векторів	0.932	0.933
Нейронна мережа	0.902	0.936
GMLNET	0.963	0.984

Таблиця.3.2.3.5 Порівняння моделей з розбиттям критерію впливу на 4 частини

Порівнюючи значення AUCпопередніх моделей та навчених з використанням критерію впливу спадковості отримали покращення прогнозування для всіх моделей окрім побудованої на основі лісу рішень. Додавання критерію впливу покращило прогнозуючі здатності для моделей навчених за допомогою: Логістичної регресії, Нейронної мережі та GMLNET.

Розбиття на 5 значення

Інтервал розбиття: 1-0.401 = 0.599.

Величина зміщення кожного критерію: 0.599/4 = 0.14975

Отримані інтервали представлені в таблиці 3.2.3.6.

Назва інтервалу	**Інтервал**
NoEffect	менше 0.401
LowEffect	[0.401;0.55075]
MediumEffect	(0.55075;0.7005]
HighEffect	(0.7005;0.85025]
GreatestEffect	більше 0.85025

Таблиця.3.2.3.6. Критерії впливу при розбитті на 4 частини

Тепер замінимо прогнозований результат моделі на значення 5 критеріїв.

Будемо порівнювати моделі побудовані тільки на основі генотипів з моделями де крім генотипів береться критерій впливу розбитий на 5 частин.

Нижче приведена таблиця 3.2.3.7 порівняння моделей.

Метод	**AUCбез критерію**	**AUC**
Логістична регресія	0.949	0.957
Ліс рішень	0.945	0.936
Метод опорних векторів	0.932	0.932
Нейронна мережа	0.902	0.937
GMLNET	0.963	0.983

Таблиця.3.2.3.7 Порівняння моделей з розбиттям критерію впливу на 5 частини

При розбитті критерію на 5 частин отримали кращу прогнозуючою здатність для наступних моделей: Логістична регресія, Нейронна мережа та GMLNET. Ліс рішень при такому розбитті критерію почав давати гіршу оцінку, а метод опорних векторів залишивсь ніяк не змінивсь.

Розбиття на 6 значення

Інтервал розбиття: 1-0.401 = 0.599.

Величина зміщення кожного критерію: 0.599/5 = 0.1198

Отримані інтервали представлені в таблиці 3.2.3.8.

Назва інтервалу	**Інтервал**
NoEffect	менше 0.401
LowestEffect	[0.401;0.5208]
LowEffect	(0.5208;0.6406]
MediumEffect	(0.6406;0.7604]
HighEffect	(0.7604;0.8802]
GreatestEffect	більше 0.8802

Таблиця.3.2.3.8. Критерії впливу при розбитті на 4 частини

Тепер замінимо прогнозований результат моделі на значення 6 критеріїв.

Будемо порівнювати моделі побудовані тільки на основі генотипів з моделями де крім генотипів береться критерій впливу розбитий на 6 частин.

Нижче приведена таблиця 3.2.3.9 порівняння моделей.

Метод	**AUCбез критерію**	**AUC**
Логістична регресія	0.949	0.957
Ліс рішень	0.945	0.940
Метод опорних векторів	0.932	0.929
Нейронна мережа	0.902	0.936
GMLNET	0.963	0.985

Таблиця.3.2.3.9 Порівняння моделей з розбиттям критерію впливу на 5 частини

При розбитті критерію на 6 частин отримали кращу прогнозуючою здатність для наступних моделей: Логістична регресія, Нейронна мережа та GMLNET. Ліс рішень та Метод опорних векторів при такому розбитті критерію почали давати гіршу оцінку.

Використання прогнозованих значень

В цьому досліді замість критерію впливу використаємо прогнозовані значення моделі.

В таблиці 3.2.3.10 приведені оцінки якості моделей побудованих на виключно генотипах та моделях з переданим результатом прогнозування спадковості.

Метод	**AUCбез критерію**	**AUC**
Логістична регресія	0.949	0.957
Ліс рішень	0.945	0.936
Метод опорних векторів	0.932	0.930
Нейронна мережа	0.902	0.933
GMLNET	0.963	0.982

Таблиця.3.2.3.10. Порівняння моделей без розбиття параметру

Якщо до генотипів додати колонку з прогнозованими значеннями, то отримуємо кращі оцінки для методів: Логістична регресія, нейрона мережа та GMLNET. При використанні лісу рішень або методу опорних векторів отримуємо гірший результат.

Вибір оптимального розбиття

Для порівняння результатів якості моделей з критерієм впливу приведена таблиця 3.2.3.11.

Метод	**AUC2**	**AUC 3**	**AUC 4**	**AUC 5**	**AUC 6**	**AUC без розбиття**
Логістична регресія	0.957	0.959	0.958	0.957	0.957	0.957
Ліс рішень	0.935	0.947	0.942	0.936	0.940	0.936
Метод опорних векторів	0.933	0.934	0.933	0.932	0.929	0.930
Нейронна мережа	0.933	0.932	0.936	0.937	0.936	0.933
GMLNET	0.980	0.984	0.984	0.983	0.985	0.982
Середнє знач.	**0.9476**	**0.9512**	**0.9506**	**0.949**	**0.9494**	**0.9476**

Таблиця.3.2.3.11. Якість моделей з критерієм впливу

Порівнюючи якості моделей з критерієм розбитим на різні величини можна зробити висновок, що найкращим є розбиття на три частини. При такому поділі використовуючи ліс рішень отримали покращення якості моделі в порівнянні з навченою тільки на основі генотипів. Інші розбиття для цього методу тільки погіршили свою характеристику. Найкращої якості вдалося досягти при використанні методу GMLNET та розбиття критерію на 6 частин, однак воно є не настільки значним, щоб надати йому перевагу.

Найгірший результат показали моделі навчені при розбитті критерію на 2 частини та вразі передачі прогнозованих значень генеалогічною моделлю в вибірку.

Оцінюючи результати можна використовувати розбиття критерію на 3, 4, 5 та 6 частин. В подальшому дослідженні використовуватиметься розбиття на 3 частини.

Поліпшення оптимального результату

Щоб покращити оцінку моделей при розбитті критерію впливу на 3 частини, потрібно змінити інтервали для позитивно класифікованих результатів.

Будуть наступний розподіл критерію.

1) Одна третина від інтервалу. Межу знаходимо взявши 1/3 від довжини діапазону позитивних значень. MediumEffect = [0.401; 0.6007] HighEffect = більше 0.6007.

2) Середина інтервалу (був вже використаний). MediumEffect = [0.401; 0.7005] HighEffect = більше 0.7005.

3) Середнє значення позитивно класифікованих результатів. Межею поділу стає середнє значення всіх позитивно класифікованих результатів. Отримуємо значення для інтервалів: MediumEffect = [0.401; 0.785] HighEffect = більше 0.785.

Результати порівняння приведені у таблиці 3.2.3.12.

Метод	**AUC 1/3**	**AUC середина інтервалу**	**AUCсереднє значення**
Логістична регресія	0.961	0.959	0.958
Ліс рішень	0.944	0.947	0.948
Метод опорних векторів	0.939	0.934	0.937
Нейронна мережа	0.937	0.932	0.932
GMLNET	0.982	0.984	0.982
Середнє значення	**0.9526**	**0.9512**	**0.9514**

Таблиця.3.2.3.12. Якість моделей з модифікованим критерієм впливу

Оцінюючи результат можна стверджувати, що зміна межі відсікання для позитивно класифікованих результатів не сильно впливає, а отже можна використовувати будь-який з них. Хоча найкращим серед трьох є критерій з межею відсікання рівний одній третій інтервалу позитивно класифікованих. Якби не його гірші показники при використанні лісу рішень він був би беззаперечним лідером.

3.3. Висновок до розділу

В цьому розділіпроведені дослідження на основі генотипів пар чоловік-жінка та визначення найкращої моделі прогнозування.

Наступним досліджувавсяокремий вплив обтяженої спадковості родоводів чоловіка та жінки на безпліддя. Після отримання оцінки якості моделей досліджувався сукупний вплив родоводів. Показники якості моделей, що враховували пару є вищими, а отже їх використовували для визначення величини впливу обтяженості родоводу.

Щоб врахувати вплив спадковості, до генотипів додається фактор-ознака представлена у вигляді критерію. Критерій береться зі шкали, яку отримано дослідницьким шляхом на основі результатів пронозу моделей спадковості.

Моделі генотипів з врахуванням спадковості мають кращі прогнозуючі властивості чим без спадковості.

4. ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ

Проведене дослідження показало, що є сенс в медицині використовувати даний підхід побудови моделі і з її допомогою оцінювати можливі проблеми безпліддя у подружніх парах. Це дозволить лікарям призначати профілактичні заходи та частково визначати проблеми безпліддя. Подібна дослідження дозволить підтвердити здогадки лікарів стосовно безпліддя пари.

Лікарі можуть подібним чином створювати інші моделі на інших даних та проводити подібні дослідження. Однак цей процес вимагає фахових навиків і звичайному лікарю генетику без підготовки не вдасться це зробити. Наприклад, встановлено, що для побудови моделі найкраще використати метод GMLNET написаний на мовіR, але викликом самої функції не вдається обмежитись, потрібно писати ще команди взяття даних з файлу, обробки даних, формування навчальної та тестової вибірки і проводити оцінку результатів. Це велика кількість роботи для лікаря і часто вона не під силу, тому що займає багато часу. Можна використовувати програми зі зручним інтерфейсом та гнучкою функціональністю. До таких можна віднести RapidMiner, Orange. Однак для лікаря вони теж не є інтуїтивно зрозумілими, тому що мають велику кількість елементів, які потрібно налаштовувати, а також ці програми реалізують функціональність, яку лікарі не будуть використовувати.

Тому є необхідність в створенні вузько направленого програмного продукту, що дозволить оцінювати безпліддя пар, реалізовуватиме тільки необхідну функціональність та матиме інтуїтивно зрозумілий інтерфейс.

4.1. Опис програми

Для розробки програмного забезпечення використовувалася мова програмування C#.

Надзвичайно важливо використовувати вірно написані методи для навчання моделі, томубуло прийнято рішення використовувати вже готові алгоритми. Безкоштовна платформа Accord.netвід Microsoft надає необхідні методи, тому будемо використовувати її. Як показало дослідження найкраща

модель прогнозування генотипів була отримана в результаті використання методу написаного на мові R. Тому є сенс також включити цей алгоритм в функціональність додатку. Для цього потрібно завантажити бібліотеку, що імітує середовище виконання мови R.

Для навчання моделі використовуватиметься крос-валідація К блоками, що також постачається платформою Accord.net.

Дослідження показало, що найкращими методами для навчання є логістична регресія, тому в даному додатку доступні два види моделі: логістична регресія та логістична регресія для роботи з розрідженими матрицями.

4.2. Робота з програмою

При завантаженні програми стає доступне головне вікно (рис. 4.2.1)

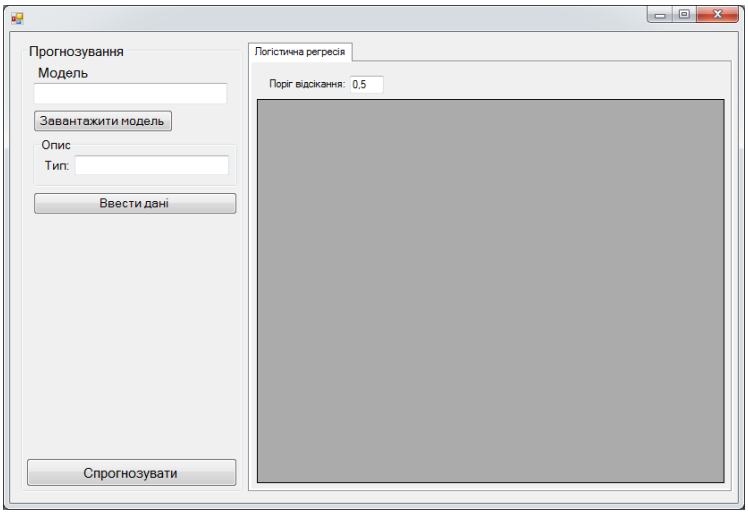

Рис. 4.2.1. Головне вікно програми

Перш ніж почати будь-яку роботу потрібно завантажити модель. На формі для завантаження (рис. 4.2.2) доступні дві функції: вибір з існуючої моделі та створення нової моделі.

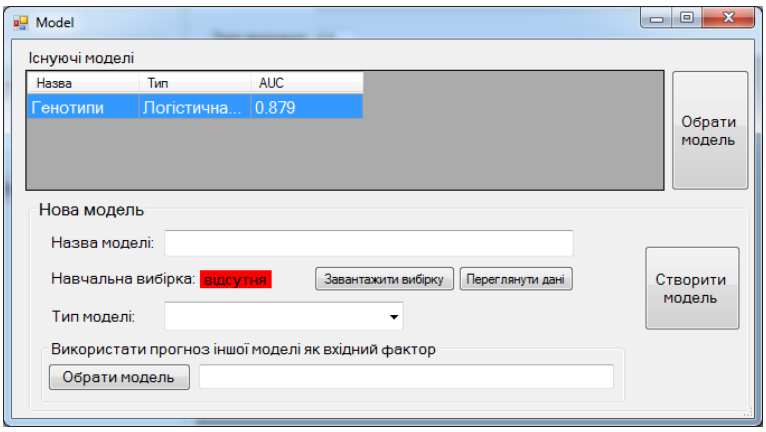

Рис. 4.2.2. Форма завантаження моделі

З списку існуючих моделей обирається та з якою будуть працювати. Щоб створити нову модель потрібно спочатку завантажити та налаштувати дані.

Програмне забезпечення дозволяє завантажувати дані з Exelтаблиці. Відразу після завантаження потрібно вказати тип даних (лише два типи: "факторні", "числові") для кожного з факторів, а також вказати колонку результативної ознаки. Форма вибору типів даних зображена на рисунку 4.2.3.

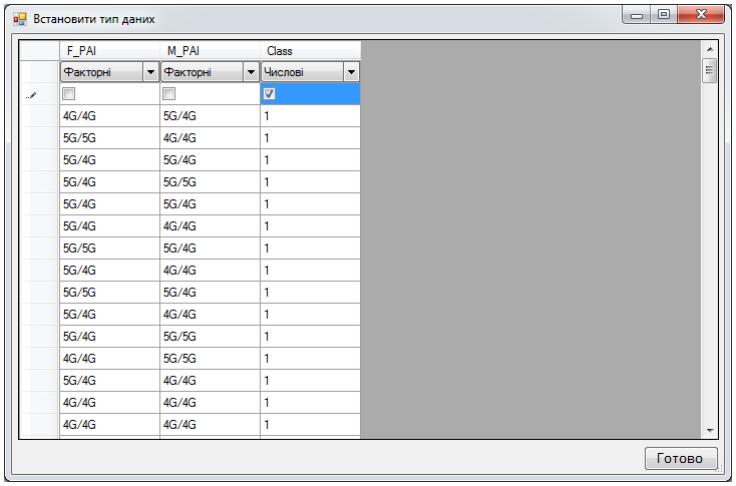

Рис. 4.2.3. Форма вибору типів даних

Після налаштування даних потрібно натиснути кнопку "Створити модель" і вона додасться в список доступних моделей.

Щоб використати модель для прогнозу потрібно на головній формі натиснути "Ввести дані". Відкриється вікно (рис. 4.2.4) для вводу даних.

Рис. 4.2.4. Вікно вводу даних

Програма сама генерує табличку для вводу даних основуючись на навченій моделі. Якщо користувачем було введено не вірні дані, які модель не може використати для прогнозування, то виведеться помилка (рис. 4.2.5)

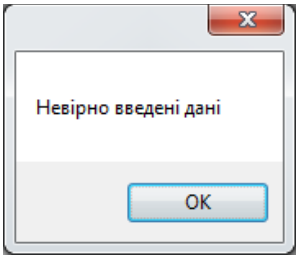

Рис. 4.2.5. Вікно помилки

Після проходження валідації дані потрапляють в модель і по них дається прогноз. Результат прогнозу виводиться на головну форму і зображується в вікні результату (рис. 4.2.6).

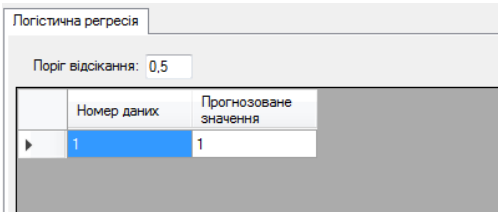

Рис. 4.2.6. Вікно результат прогнозу

Щоб використати прогноз іншої моделі як вхідний критерій потрібно на формі вибору моделей натиснути "Обрати модель". Відкриється вікно з доступними моделями (рис. 4.2.7).

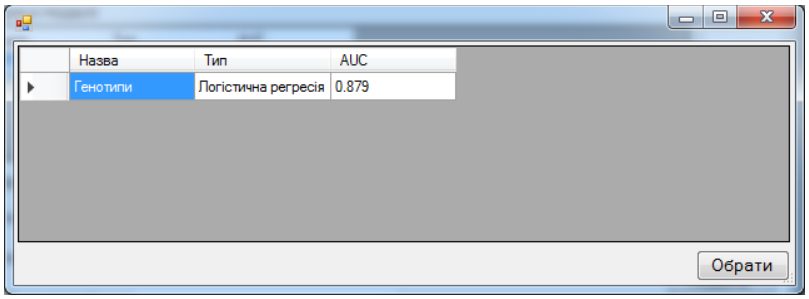

Рис. 4.2.7. Форма вибору моделей

Якщо використовувати таку модель для прогнозування в вікні вводу даних будуть відображатися колонки з обох моделей.

4.3.Висновок до розділу

Розроблений програмний додаток використовує сучасні алгоритми машинного навчання, що постачаються платформою Accord.netta R. Він містить в собі достатню функціональність, щоб відтворити проведене дослідження та виконувати подібні.

Додаток простий в користуванні, а отже лікарям не складно буде в ньому працювати. Завдяки прогнозованим значенням моделі лікар зможе використовувати цю оцінку для встановлення діагнозу чи призначення профілактики, лікування. Додаток дозволяє економити час лікаря.

5. РОЗДІЛ ОХОРОНИ ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

Для аналізу було обрано робоче місце з персональним комп'ютером де відбувалася розробка дипломної роботи. Схема приміщення у масштабі із зазначеними на ній меблями, обладнанням та устаткуванням буде наведена в наступному розділі. Робоче місце розраховано на одну особу. Ми маємо розглянути шкідливі чинники, їх чисельні показники. Приміщення буде розглянуто з точки зору охорони праці і буде з'ясовано наскільки приміщення відповідає стандартам та ДСТУ що існують та застосовуються на даний момент, в разі необхідності наведемо рекомендації щодо усунення порушень.

5.1. Загальна характеристика приміщення

Розглянемо приклад приміщення в якому була виконана дипломна робота.

Приміщення розташоване на 2 поверсі 5-ти поверхової офісної будівлі. Це задовольняє НПАОП, що не дозволяє розміщувати приміщення де використовується обчислювальна техніка на цокольних поверхах та у підвальних приміщеннях.

Характеристики приміщення наведемо в таблиці 5.1.1:

Довжина	5 м
Ширина	3 м
Висота	3 м
Загальна площа	15 $м^2$
Загальний об'єм	45$м^3$
Кількість робочих місць	1
Площа на робоче місце	15 $м^2$
Об'єм на робоче місце	45 $м^3$
Кількість вікон	1
Кількість батарей центрального водного опалення	1

Таблиця 5.1.1 – Характеристики робочого приміщення

Для приміщень, обладнаних комп'ютерами, згідно [5] площа на одне робоче місце має становити не менше ніж 6,0 $м^2$, а об'єм не менше ніж 20,0 $м^3$. В нашому випадку на одне робоче місце припадає площа 15 $м^2$ та об'єм 45 $м^3$, що відповідає встановленим нормам.

Фарбування приміщень і меблів повинна сприяти створенню сприятливих умов для зорового сприйняття, гарного настрою. Оскільки вікно приміщення орієнтоване на північ, рекомендується наступне фарбування стін і підлоги:

стіни світло-жовтогарячий або оранжево-жовтий кольори, підлога – червонясто-жовтогаряча.

Джерела світла, такі як світильники й вікна, які дають відбиття від поверхні екрана, значно погіршують точність знаків і спричиняють перешкоди фізіологічного характеру, які можуть виразитися в значній напрузі, особливо при тривалій роботі. Відбиття, включаючи відбиття від вторинних джерел світла, повинне бути зведене до мінімуму. Для захисту від надлишкової яскравості вікон застосовані вертикальні жалюзі.

Приміщення оснащене кондиціонером та системою опалення.

В кімнаті є 2 стола, один, основний, глибиною 800 мм, шириною 1200мм, висота стола повинна бути обрана з урахуванням можливості сидіти вільно, у зручній позі, при необхідності опираючись на підлокітники, тому рекомендована висота 680-760мм (в даному випадку становить 680 мм), інший стіл з глибиною 600мм., шириною 1000мм. для розміщення додаткових периферійних пристроїв.

Висота поверхні, на яку встановлюється клавіатура, 650мм. Конструкція основного стола містить 3 ящики, для зберігання документації. Біля столу стоїть крісло з висотою сидіння над рівнем підлоги 520 мм (з можливістю регулювання висоти). Поверхня сидіння м'яка, передній край закруглений, а кут нахилу спинки – регульований. На столі міститься LCD монітор на відстані 550мм від краю столу, клавіатура на висувній підставці, оптична мишка. Системний блок розташований поряд зі столом. Поряд, на додатковому столі встановлений принтер з Wi-Fi. В робочому приміщенні розташовані 2 шафи: одна для документів та інших паперів (глибиною 600мм), інша для верхнього одягу та ін. (глибиною 900мм).

Столи та стільці відповідають вимогам [5].

Параметри приміщення задовольняють вимогам.

Робоче місце включає в себе: системний блок, монітор, принтер та інші периферійні пристрої, які необхідні.

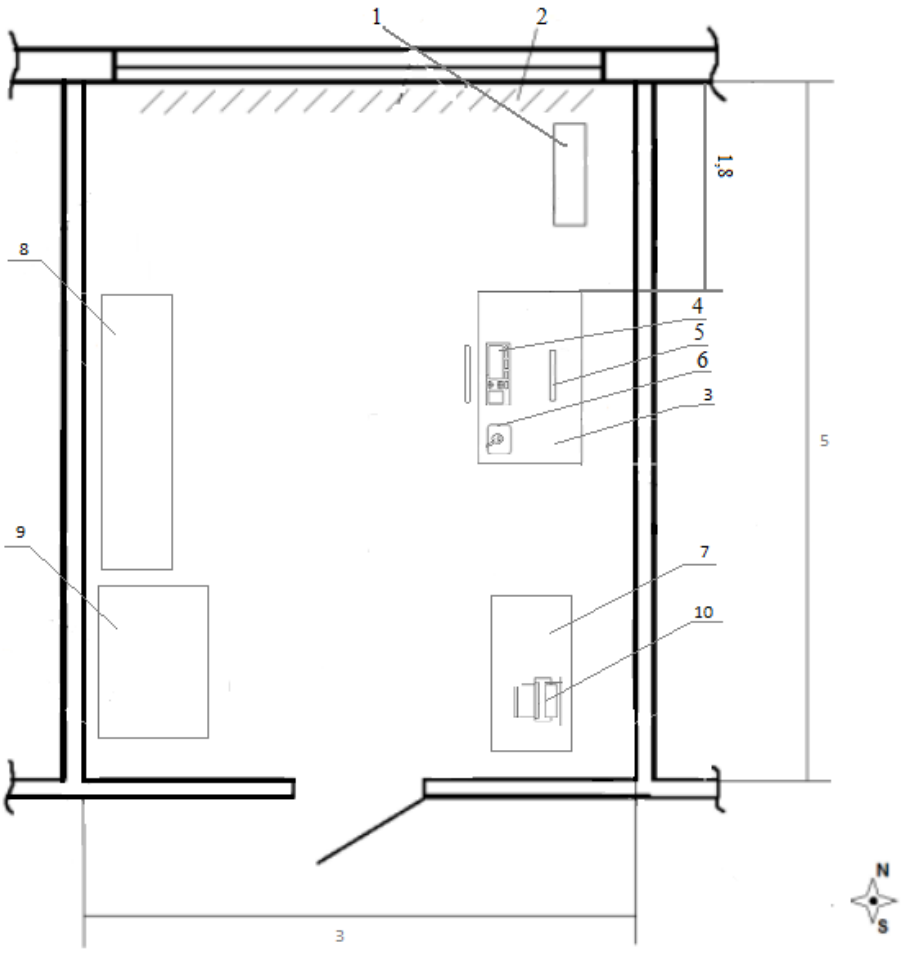

Рис. 5.1.1Схемаприміщення зкомп’ютеризованим робочим місцем:

1–кондиціонер; 2 – сонцезахисні жалюзі; 3 – комп’ютеризоване робоче місцепрацівника;4 – клавіатура;5–монітор; 6– коврик з оптичною мишкою;7–стіл для периферійних пристроїв;8–шафа для паперу та документів;9– шафа для верхнього одягу; 10 – принтер.

5.2. Аналіз шкідливих і небезпечних факторів

До основних небезпечних і шкідливих факторів належать мікроклімат, освітлення, шум, випромінювання, психофізіологічні фактори, електробезпека і пожежна безпека. В даному приміщенні слід окремо виділити такі шкідливі фактори як електромагнітне випромінювання (комп’ютерна техніка) і шум(комп’ютерна техніка, периферія, кондиціонування).

5.2.1. Мікроклімат приміщення

Згідно з [6]категорія робіт, здійснюваних в даному приміщенні, легка 1а (роботи, що виконуються сидячи і не потребують фізичного напруження). За санітарними нормами мікроклімат виробничих приміщень для категорії робіт легка 1а описаний в табл.5.1 та табл.5.2 :

Оптимальні та допустимі значення параметрів мікроклімату в робочій зоні наведені у таблиці 5.2.1.

Період року	Норми	Температура повітря, °С	Відносна вологість повітря, %	Швидкість руху повітря, м/с
Холодний	Опт.	22-24	40 – 60	0,1
	Доп.	21-25	75	0,1
Теплий	Опт.	23-25	40 – 60	0,1
	Доп.	22-28	55	0,1

Таблиця 5.2.1. – Мікроклімат приміщення

Система опалення централізована, використовується тільки в опалювальний сезон, в холодну пору року. Приміщення не має централізованої системи кондиціонування, вентиляція приміщення – природного типу. Місцева система кондиціонування – настінний кондиціонер (внутрішній блок спліт-системи) складається з автономного кондиціонеру «LG S09LHPT» (з площею охолодження до 26 м^2 та потужністю 2.7 кВ/год.), що встановлений біля вікна. Швидкість руху повітря у приміщенні приблизно 0.1 м/с, вологість -50%, температура повітря коливається від 21 ˚с до 23 ˚с. Шкідливі речовини у повітрі відсутні. Виходячи з вищенаведеного і враховуючи, що робоче місце постійне, можна зробити висновок, що згідно [6] мікрокліматичні умови знаходяться в оптимальних діапазонах.

5.2.2. Освітлення приміщення

Освітлення у приміщенні змішане, складається з бокового (вікно) та штучного загального освітлення. Необхідно провести розрахунок освітлення для кімнати площею 15 м ^{2,} ширина якої 3 м, довжина 5 м, висота – 3,5 м. Визначимо показники методом світлового потоку.

З табл. 1 [7] розряд зорової роботи – III (0.3-0.5 мм.) та під розряд – г. Система природного освітлення: бокове одностороннє освітлення, вікно виходить на північ, коефіцієнт світлового клімату m=0.9. КПО, ен, % = 2. Вікна в приміщенні задовольняють вимоги для природного освітлення.

Для визначення кількості світильників визначимо світловий потік, падаючий на поверхню за формулою:

$$F_л = \frac{E \cdot k \cdot s \cdot z}{\eta} \quad (5.2.1.1)$$

Де Fл – світловий потік, лм;

E – нормована мінімальна освітленість;

s – площа освітлюваного приміщення;

z – відношення середньої освітленості до мінімальної ;

к – коефіцієнт запасу, враховує зменшення світлового потоку лампи в результаті забруднення світильників у процесі експлуатації ;

n – коефіцієнт використання, (виражається відношенням світлового потоку, що падає на розрахункову поверхню, до сумарного потоку всіх ламп і обчислюється в частках одиниці; залежить від характеристик світильника, розмірів приміщення, фарбування стін і стелі, які характеризуються коефіцієнтами відображення від стін ($р_с$) і стелі ($р_п$)), значення коефіцієнтів $р_с$ і $р_п$: $р_с = 50\%$, $р_п = 70\%$. Значення n визначимо по таблиці коефіцієнтів використання різних світильників. Для цього обчислимо індекс приміщення по формулі :

$$i = \frac{s}{h \cdot (A + B)}, \quad (5.2.1.2)$$

Де s – площа приміщення;

Н – розрахункова висота підвісу світильників над робочою поверхнею

а – ширина приміщення;

в – довжина приміщення;

$h = H - h_p - h_c$

Де h – висота приміщення;

$h_p = 0{,}7$м – висота робочої поверхні від підлоги;

$h_c = 0{,}6$м – відстань від ламп до перекриття;

Таким чином,

$h = 3\text{-}0{,}7\text{-}0{,}6 = 1{,}7$м

Підставивши значення отримаємо:

$$i = \frac{15}{1.7 \cdot (5 + 3)} = 1{,}1$$

Знаючи індекс приміщення і, знаходимо n=0.43.

Підставимо значення у формулу для визначення світлового потоку:

$$\text{F}_\text{л} = \frac{300 \cdot 1{,}5 \cdot 15 \cdot 1{,}1}{0{,}43} = 17267 \ Лм$$

Для освітлення використані люмінесцентні лампи типу ЛБ 40-1, світловий потік яких *F = 4320 Лм.* Розрахуємо необхідну кількість ламп у світильниках за формулою:

$$N = \frac{F_л}{F} \quad (5.2.1.3)$$

Де N – обумовлене число ламп;

$F_л$ – світловий потік, $\text{F}_\text{л} = 17267$ Лм;

F – світловий потік лампи, *F = 4320 Лм*.

$$N = \frac{17267}{4320} \approx 4$$

Число ламп – 4

В приміщенні використовуються світильники типу ОД. Кожен світильник комплектується двома лампами. Тобто необхідно використовувати 2 світильника із 4 працюючими лампами в них. Схема розташування світильників зображена на рис. 5.2.2.1

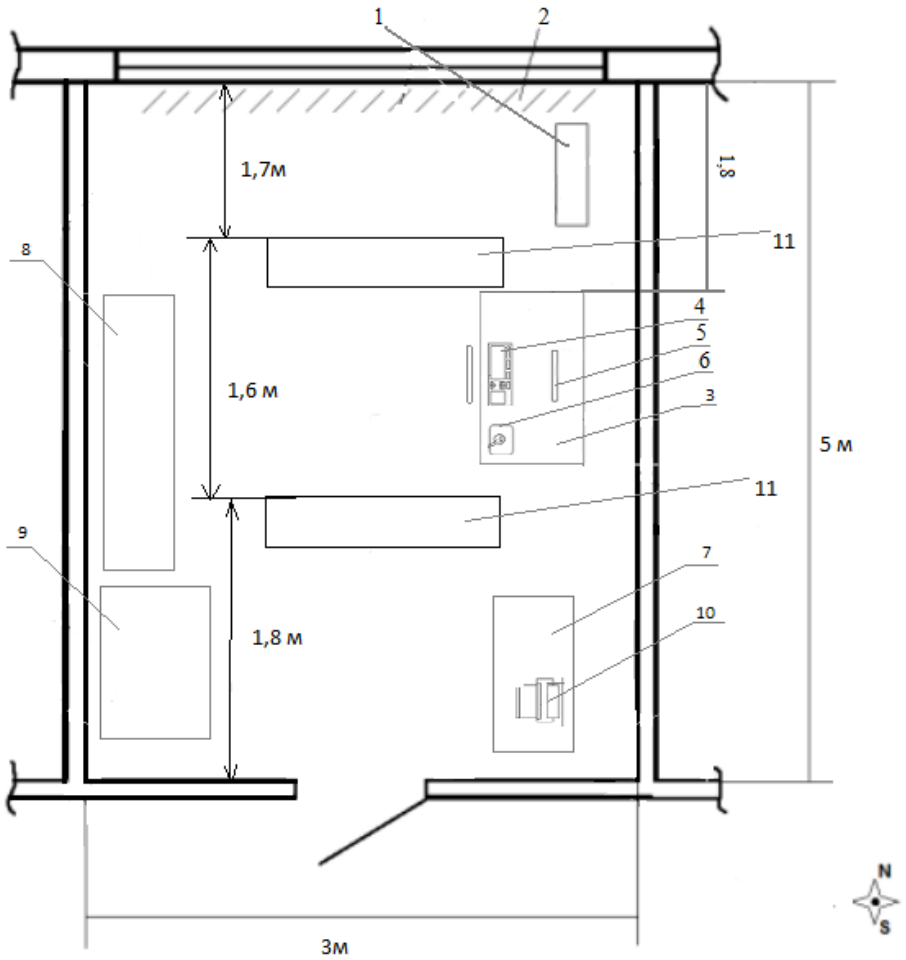

Рис. 5.2.2.1Схемаосвітлення приміщення: 11 – світильник (2 лампи)

Тип штучного освітлення в приміщенні – загальне. Штучне освітлення застосовується при роботі в темний час доби й удень, коли не вдається забезпечити нормовані значення коефіцієнта природного освітлення (похмура погода, короткий світловий день). В приміщенні розташовані дволампові світильники 2 шт., паралельно вікну. Штучне освітлення забезпечують лампи ЛБ40-1 довжиною 1.2м потужністю 40 Вт, Fл = 4320 лм.

5.2.3. Аналіз шуму

Рівень шуму на робочому місці програмістів і операторів відеоматеріалів не повинен перевищувати 50дба, а в залах обробки інформації на обчислювальних машинах – 65дба. Для зниження рівня шуму стіни і стеля приміщень, де встановлені комп'ютери, облицьовуються звукопоглинальними матеріалами. Рівень вібрації в приміщеннях обчислювальних центрів знижується шляхом встановлення обладнання на спеціальні віброізолятори.

В даному приміщенні застосовується кондиціонер LG S09LHPT, який створює шум в діапазоні від 26 до 33 дБа за рахунок роботи самого пристрою та потоків повітря, які виходять з нього на невеликій швидкості.

Також шум буде створювати системний блок ПК, так як в ньому знаходиться система охолодження, рівень шуму якої становить 22 дБа. Діє постійно весь робочий час. Також в приміщенні є принтер з рівнем шуму 34 дБа.

Рівні звукового тиску джерел шуму, що діють на оператора на його робочому місці представлені в таблиці 5.2.3.1.

Джерело шуму	Рівень шуму, дба
Системний блок	22
Монітор	6
Клавіатура	12
Принтер	34
Кондиціонер	33

Таблиця 5.2.3.1. – рівні звукового тиску різних джерел.

Рівні звуку та еквівалентні рівні звуку на робочих місцях, обладнаних ВДТ і ЕОМ, мають відповідати вимогам [8] та не перевищувати 50 дБа. В даному приміщенні рівень шуму становить:

$$L = 10\log\left(10^{0.1\times22} + 10^{0.1\times6} + 10^{0.1\times12} + 10^{0.1\times34} + 10^{0.1\times33}\right) = 84.52 \text{ дБа}$$

Отримане значення перевищує допустимий рівень шуму для робочого місця оператора, рівний 65 дб, але якщо врахувати, що навряд чи такі периферійні пристрої як сканер і принтер будуть використовуватися одночасно, то ця цифра буде ще нижчою. Крім того, при роботі принтера безпосередню присутність оператора необов'язково.

5.2.4. Аналіз електромагнітного та іонізуючого випромінювання

Вимоги до рівнів електричних та магнітних полів в виробничому приміщенні описані в [9].

Комп’ютер, що використовуються в приміщенні, пройшов відповідну сертифікацію і відповідає вимогам цих нормативних документів. Яскравість екрану складає 250-300 кд/кв.м., що більше ніж мінімальні нормативні значення 100-120 кд/кв. Відношення яскравості екрану до яскравості навколишніх поверхонь не перевищує 3:1.

Відстань від очей оператора до екрану складає 70 см в нормальному робочому положенні.

Тривалість зосередженого спостереження для працюючого за ПЭВМ не перевищує 80 % усього часу спостереження.

Оператор ЕВМ робить перерви 10-15 хвилин кожні 2 години одночасно з провітрюванням приміщення.

Допустимі значення параметрів неіонізуючих електромагнітних випромінювань від монітора комп'ютера представлені в таблиці 5.5. Максимальний рівень рентгенівського випромінювання на робочому місці оператора комп'ютера звичайно не перевищує 10мкбер / ч, а інтенсивність ультрафіолетового і інфрачервоного випромінювань від екрану монітора лежить в межах 10 ... 100мвт / м $^{2.}$

Найменування параметра	Допустимі значення
Напруженість електричної складової електромагнітного поля на відстані 50см від поверхні відео монітора	10в / м
Напруженість магнітної складової електромагнітного поля на відстані 50см від поверхні відео монітора	0,3 а / м
Напруженість електростатичного поля не повинна перевищувати: для дорослих користувачів для дітей дошкільних установ і що вчаться середніх спеціальних і вищих навчальних закладів	20кв / м 15кв / м

Таблиця 5.2.4.1. – допустимі значення параметрів неіонізуючих електромагнітних випромінювань

Усі перераховані параметри відповідають нормам.

5.3. Електробезпека

Приміщення відноситься до класу приміщень без підвищеної небезпеки для здоров’я працівників. Відносна вологість повітря не перевищує 60%, температура не перевищує 25°С. В приміщені присутні споживачі електроенергії: освітлювальні прилади, комп’ютер, кондиціонер.

В приміщені використовується прихована проводка, що виключає можливість дотику до оголених проводів. Проводка прокладена кабелем марки ВВП 3*1,5 з мідними струмопровідними жилами; ізоляція виготовлена з ПВХ; кабель плоский з розділеною основою.

Номінальна зміна напруги 220 В; частота до 50 Гц. $I_н$=15 А, $I_ф$=15 А. Вмикачі штучного освітлення виконані з діелектричних матеріалів.

Для забезпечення електробезпеки застосовується заземлення. Всі струмопровідні прилади ізольовані. В них використовується подвійне ізоляційне покриття (передня частина системного блоку виконана з пластику,

що виключає дотик користувача до металічних частин). Блок живлення в корпусі комп'ютера знаходиться в окремому захисному кожусі, що виключає можливість доступу до високої напруги.

Отже, в приміщенні використовуються наступні засоби захисту:

1) при нормальних умовах :
   - ізоляція;
   - недоступність токовивідних частин;
   - інформаційні позначення (використання маркувань окремих частин електричного обладнання, написи, попереджувальні знаки);
2) при аварійних умовах:
   - заземлення;
   - подвійна ізоляція;
   - запобіжні автомати.

Для попередження ураження електричним струмом використовується ряд організаційно-технічних дій:

- розташування проводів живлення поза зоною переміщення людей;
- всі співробітники ознайомленні з правилами техніки безпеки і в приміщенні знаходяться наступні пам'ятки для співробітників:
  - ✓ Перша допомога постраждалому від електричного струму;
  - ✓ Інструкція по техніці безпеки для працюючих в класах персональних ЕОМ.

Усі перераховані параметри відповідають ДСанПіН 3.3.6.096-2002.

5.4. Пожежна безпека

Приміщення відноситься до категорії В по вибухонебезпечній і пожежній небезпеці за ознакою перебування в ньому важкогорючих твердих й волокнистих речовин і матеріалів: ПК, монітор, папір, меблі тощо.

В приміщені є 2 вогнегасників ОУ-3 (об'єм 3 літрів, маса заряду – 2кг). Отже, на площі $15м^2$ є 2 вогнегасників, що задовольняє вимогам НАПБ В 03.002-2007 (на 20 квадратних метрів необхідно 2 вогнегасника).

Технічними засобами виявлення пожежі обраний димовий оптико-електронний автономний автоматичний сповіщувачий ДІП-43М встановлений на стелі приміщення на відстані.

На стіні в коридорі знаходиться план евакуації людей при пожежі. Двері в бік аварійного виходу відкриваються назовні. На шляху до дверей немає перешкод для руху. Всередині біля входу знаходиться вимикач електроенергії в межах всього робочого приміщення.

По пожежній безпеці приміщення відповідає вимогам НАПБ В 03.002-2007.

5.5. Інструкція з техніки безпеки

5.5.1. Під час роботи

Щоб уникнути пошкодження ізоляції проводів і виникнення коротких замикань не дозволяється:

- Вішати що-небудь на дроти,
- Зафарбовувати й білити шнури і дроти,
- Закладати дроти і шнури за газові та водопровідні труби, за батареї опалювальної системи,
- Висмикувати штепсельну вилку з розетки за шнур, зусилля повинно бути надано на корпус вилки.

Для виключення ураження електричним струмом забороняється:

- Часто вмикати і вимикати комп'ютер без необхідності,
- Торкатися до екрану і до тильної сторони блоків комп'ютера,
- Працювати на засобах обчислювальної техніки та периферійному обладнанні мокрими руками,
- Працювати на засобах обчислювальної техніки та периферійному обладнанні, що мають порушення цілісності корпусу, порушення ізоляції проводів, несправну індикацію включення живлення, з ознаками електричної напруги на корпусі,
- Класти на засоби обчислювальної техніки і периферійному обладнанні сторонні предмети.
- Під напругою очищати від пилу і забруднення електричне обладнання.

- Перевіряти працездатність електроустаткування в непристосованих для експлуатації приміщеннях з струмопровідними підлогами, сирих, які не дозволяють заземлити доступні металеві частини.
- Під напругою проводити ремонт засобів обчислювальної техніки і периферійного обладнання. Ремонт електроапаратури проводиться тільки фахівцями-техніками з дотриманням необхідних технічних вимог.

Щоб уникнути ураження електричним струмом, при користуванні електроприладами не можна торкатися одночасно будь-яких трубопроводів, батарей опалення, металевих конструкцій, з'єднаних із землею.

При користуванні електроенергією в сирих приміщеннях дотримуватися особливої обережності.

5.5.2. По закінченню роботи

По закінченні роботи оператор зобов'язаний дотримуватися такої послідовності вимикання обчислювальної техніки:

- зробити закриття всіх активних задач;
- переконатися, що в cd/dvd приводах немає дисків;
- виключити живлення системного блоку;
- виключити живлення всіх периферійних пристроїв;
- відключити блок живлення.

По закінченні роботи оператор зобов'язаний оглянути й упорядкувати робоче місце, вимити з милом руки й обличчя, перед виходом вимкнути всі прилади та світло в приміщенні.

5.6. Психофізіологічні небезпечні та шкідливі виробничі фактори

До небезпечних та шкідливих психофізіологічних виробничих чинників належать фізичні (статичні, динамічні та гіподинамічні) і нервово-психічні перевантаження (розумове, зорове, емоційне).

Праця науково-дослідних та інших установ, а також інших працівників невиробничої сфери характеризується тривалою багатогодинною (8 год і більше) працею в одноманітному напруженому положенні, малою руховою активністю при значних локальних динамічних навантаженнях. Робоче положення "сидячи" супроводжується статичним навантаженням значної кількості м'язів ніг, плечей, шиї та рук, що дуже втомлює. М'язи перебувають довгий час у скороченому стані і не розслабляються, що погіршує кровообіг. В

результаті виникають больові відчуття в руках, шиї, верхній частині ніг, спині та плечових суглобах.

Помірними гімнастичними вправами можна викликати активізацію обміну речовин в організмі.

Тривала робота на комп'ютеризованому робочому місці призводить до значного навантаження на всі елементи зорової системи і зумовлює втому та перевтому зорового аналізатора. Напружена зорова робота викликає "очні" (біль, печія та різь в очах, почервоніння повік та очей, ломота у надбрівній частині тощо) та "зорові" (пелена перед очима, подвоєння предметів, мерехтіння, швидка втома під час зорової роботи) порушення органів зору, що може викликати головний біль, посилення нервово-психічного напруження, зниження працездатності.

Як згадувалося раніше, регулярні перерви для виконання нескладних вправ кожні 15-20 хв. Допоможуть зняти втому та напруженість.

5.7. Безпека в надзвичайних ситуаціях

За останні роки різко збільшилась кількість надзвичайних ситуацій на території України. Значна кількість надзвичайних ситуацій виникла на об'єктах електроенергетики, більшість із них пов'язана з аваріями на електроенергетичних мережах, що призводило до масового відключення від електропостачання на тривалий період часу значної кількості населених пунктів. Практично всі ці надзвичайні ситуації мали природне походження (сильний снігопад, обледеніння, ураган та шквальний вітер) хоча, слід відзначити, що немало НС сталося техногенного характеру (пожежі, вибухи, пошкодження тощо) та надзвичайні ситуації іншого характеру (інфекційні захворювання, виявлення особливо небезпечних речовин та ін.).

За класифікацією надзвичайні ситуації можна розподілити на: надзвичайні ситуації техногенного, надзвичайні ситуації природного, медико-біологічного характеру, надзвичайні ситуації іншого характеру, внаслідок яких загинуло І постраждало багато людей. За масштабом дії НС розподілені так: надзвичайні ситуації загальнодержавного масштабу, регіонального масштабу, місцевого масштабу, об'єктового масштабу.

До надзвичайних ситуацій техногенного характеру відносяться:

- транспортні аварії;
- пожежі, вибухи;
- аварії з викидом (загрозою викиду) сильнодіючих отруйних речовин (СДОР);

- наявність в оточуючому середовищі шкідливих речовин понад гранично-допустиму кількість;
- аварії з викидом (загрозою викиду) радіоактивних речовин (РР);
- раптове зруйнування споруд;
- аварії на електроенергетичних системах;
- аварії на комунальних системах життєзабезпечення;
- аварії на очисних спорудах;
- аварії на системах життєзабезпечення;
- гідродинамічні аварії.

Надзвичайні ситуації природного характеру:
- геофізично небезпечні явища;
- геологічне небезпечні явища;
- метеорологічне та агрометеорологічне небезпечні явища;
- морські гідрологічне небезпечні явища;
- гідрологічне небезпечні явища;
- пожежі лісові та торф'яні;
- масова загибель диких тварин.

Надзвичайні ситуації медичного та біологічного характеру:
- інфекційні захворювання людей;
- отруєння людей;
- інфекційні захворювання сільськогосподарських тварин;
- ураження сільськогосподарських рослин хворобами та шкідниками;
- масові отруєння сільськогосподарських тварин.

Надзвичайні ситуації Іншого характеру:
- збройні напади, захоплення І утримання важливих об'єктів або реальна загроза здійснення таких акцій;
- встановлення вибухового пристрою в громадському місці, установі, організації, підприємстві, житловому секторі, на транспорті;
- зникнення або крадіжка з об'єктів зберігання, використання, переробки та при транспортуванні;

- виявлення або вилучення особливо небезпечних предметів та речовини;
- інші події (не класифіковані).

Так як в Україні знаходиться 15 атомних реакторів, є ризик руйнування ядерного об'єкту. Неподалік від місця проживання (місця роботи над дипломною роботою) знаходиться АЕС, припустимо що в 12-00 10.01.15 там сталося руйнування ядерного реактору, що спричинило викид радіоактивних речовин, а в 12-30 були виміряні рівні радіації. Припустимо що ми потрапили в групу ліквідаторів аварії, або нам доручили оцінити наслідки для такої групи. Тому оцінимо радіаційну обстановку для груп ліквідації наслідків аварії.

Радіаційна обстановка – це обстановка, яка утворюється на території адміністративного району, населеного пункту, господарського об'єкту в результаті радіаційного забруднення при аварії на АЕС та інших радіаційно – небезпечних об'эктах.

При оцінці радіаційної обстановки потрібно визначити:

1) Рівень радіації на 1 годину після аварії.
2) Дозу опромінення, яку отримано під час роботи.
3) Допустимий час роботи по встановленій дозі опромінення.
4) Робимо висновки про радіаційну обстановку, можливість проведення робіт в зоні, наслідки перебування та способи захисту.

Вихідні дані наступні:

- Виміряний рівень радіації на 12-30: $P_{вим}$ = 40 Р/год.
- Час початку роботи: 14-00.
- Необхідний час перебування в зоні зараження: 3.5год.
- Допустима доза радіації: 20 Р.
- Коефіцієнт ослаблення радіації: $К_{осл}$ = 2.
- Реактор: ВВЕР.

Так як всі розрахунки мають проводитися в відносному часі, то переведемо астрономічний час у відносний (рис. 5.7.1):

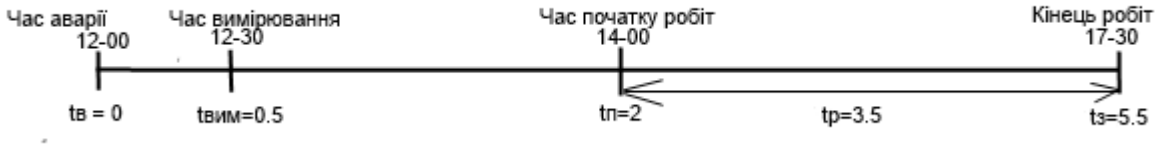

Рис. 5.7.1. Графік часу робіт

Де:

$t_в$ – початок відліку;

$t_{вим}$ – час вимірювання радіації;

$t_п$ – час початку робіт;

$t_р$ – тривалість робіт;

$t_з$ – час завершення робіт.

1) Визначимо рівень радіації через 1 годину після аварії за наступною формулою:

$$P_1 = P_{вим.} \cdot K_{t_{вим.}} = 40 \cdot 0,7 = 28P / c$$

Де K_t - коефіцієнт для перерахунку рівнів радіації на рівний час після аварії.

2) Визначимо дозу опромінення, отриману при роботі в зоні за формулою:

$$Д = \frac{P_{сер.} \cdot t_р}{K_{осл}} \quad (5.7.1)$$

Де:

$$P_{сер.} = \frac{P_п + P_к}{2} \quad (5.7.2)$$

В свою чергу:

$Р_п$ – рівень радіації на час початку робіт, визначається за наступною формулою:

$$P_п = \frac{P_1}{K_{t_п}} = \frac{28}{1.32} = 21.21P / год.$$

$Р_к$ – рівень радіації на час закінчення робіт, визначається за наступною формулою:

$$P_к = \frac{P_1}{K_{t_к}} = \frac{28}{1.98} = 14.14P / год.$$

Де $К_{tп}$ та $К_{tк}$ – коефіцієнти перерахунку рівнів радіації.

Тепер порахуємо середнє значення:

$$P_{сер.} = \frac{P_п + P_к}{2} = \frac{21.21+14.14}{2} = 17.675P / год.$$

Маючи всі необхідні дані визначимо дозу опромінення:

$$Д = \frac{P_{сер.} \cdot t_р}{K_{осл}} = \frac{17.675 \cdot 3.5}{2} = 30.93P$$

3) Визначимо допустимий час знаходження в небезпечній зоні по допустимій дозі опромінення:

$$a = \frac{P_1}{Д_{доп.} \cdot K_{осл}} = \frac{28}{20 \cdot 2} = 0.7$$

Де $Д_{доп.}$ = 20 Р – допустивидопустимий рівень радіації.

За величиною «а» і по часу початку роботи, можна визначити допустимий час перебування в зараженій зоні згідно настпного графіку (рис. 5.7.2.):

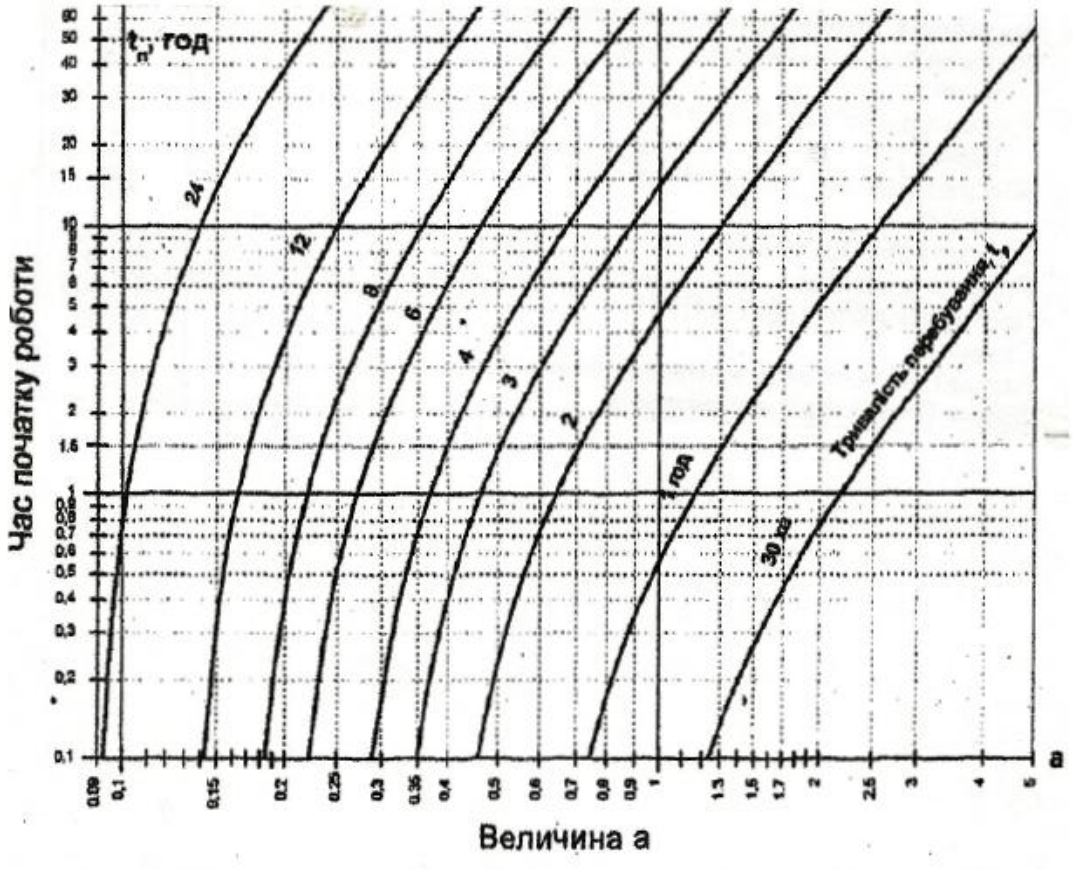

Рис 5.7.2. Графік визначення тривалості перебування в зоні радіоактивного зараження при аварії на АЕС з реактором ВВЕР-1000

Згідно графіку допустимий час перебування в зоні забруднення рівний 2 годинам без значних ушкоджень організму.

В будь-якому випадку, при перебуванні в небезпечній при/після виникненні вибуху ядерного реактору зоні необхідно дотримуватися наступних принципів захисту від іонізуючого випромінювання:

- Запобігання потраплянню радіоактивного пилу в середину організму людини. Для цього необхідно захистити органи дихання

респіратором або захисною пов'язкою, закрити волосся, шкіру, використати захисні окуляри для захисту очей, не користуватися кремами, косметикою, мазями, які сприяють налипанню радіоактивних часток на шкіру. Необхідно коротко постригти нігті, частіше приймати душ, промивати порожнину носа теплою водою, проводити вологе прибирання приміщень, захищати їх від потрапляння пилу.

- Врахування змін терміну, екрану, відстані. Чим менший термін опромінення людини, тим краще; чим важчі ядра атомів, з яких складається екран, що захищає людину, та чим він товстіший, тим краще; чим більша відстань від джерела опромінення людини, тим краще.

Захисну екрануючу дію мають такі матеріали, як парафін, графіт, вода, котрі затримують швидкі нейтрони. Сповільнені нейтрони також легко поглинаються бором, кадмієм, гадолінієм, індієм. При захисті від нейтронів використовується комбінація сповільнюючих і поглинаючих речовин. В якості захисного екрану від опромінення широко використовується бетон із спеціальними наповнювачами.

- Створення в організмі людини надлишку нерадіоактивного елемента, близького за властивостями до радіоактивного. Одним з таких елементів є йод – хімічний елемент VII групи періодичної системи. Радіоактивні ізотопи йоду набувають великого токсикологічного значення в перші двадцять днів після ядерного вибуху. Радіоактивний йод безперервно утворюється при руйнуванні ядерних реакторів АЕС. Йодна профілактика полягає у прийманні перорально (через рот) препаратів стабільного йоду – йодистого калію, або п'ятивідсоткового водно-спиртового розчину йоду. Йодистий калій слід приймати після їжі разом із чаєм або водою один раз на день протягом семи діб.

- Застосування індивідуальних засобів захисту. При роботі з відкритими радіоактивними речовинами та на зараженій місцевості застосовуються індивідуальні засоби захисту: протигази (респіратори), спеціальний одяг, захисні рукавиці. Радіоактивне забруднення спецодягу,

засобів захисту і шкірних покривів людини не повинно перевищувати допустимих рівнів.

- Виведення радіоактивних ізотопів з організму за допомогою промивання шлунку, використання активованого вугілля та інших способів, які застосовують при хімічних отруєннях.
- Правильне харчування. Найкращим джерелом цінного в біологічному відношенні білку є яйця, нежирна телятина, тріска, яловичина другої категорії, нежирна баранина, нежирний сир, просо, бобові. Білок цих продуктів відрізняється оптимальним вмістом незамінних амінокислот і добре засвоюється. Якщо в харчуванні людини, яка знаходиться на радіаційно забрудненій території, не вистачає білку, то відбувається накопичення в організмі значної кількості цезію-137. Раціони з повноцінним вмістом білку пришвидшують виведення цього радіонукліду з організму.
- Забезпечення чистої води є найважливішим завданням у період перебування на зараженій території. Додатковими способами очищення звичайної питної води може бути дистиляція (очищення від радіоактивних домішок випаровуванням), заморожування, фільтрація тощо. У період перебування в зоні не бажано обмежувати потребу у чистій питній воді. Не треба використовувати солоні продукти, які затримують воду в організмі.

При дотриманні вищеописаних рекомендій в разі виникнення надзвичайної ситуації можна зберегти собі життя, і отримати мінімальну шкоду для свого організму.

5. 8. Висновки до розділу

В даному розділі були розглянуті питання, що стосуються безпеки використання та правильності експлуатації комп’ютерної техніки. Мікроклімат приміщення, освітленість відповідають відповідним вимогам та нормам. Шум несуттєво перевищує задані норми, але не знаходиться на критичному рівні через специфічність техніки (принтер). Електробезпека та пожежна безпека також забезпечені в повній мірі. На підставі цього можна зробити висновок, що робочі місця в даному приміщенні задовольняють екологічним нормам і вимогам та створюють безпечні умови праці.

ВИСНОВОКИ

Предметною областю дослідження прояву спадкових хвороб стало безпліддя подружньої пари. Ця проблема є актуальною в Україні і вимагає постійних досліджень. Новизною є спосіб проведення. Раніше досліджували безпліддя окремо чоловіка та жінки, а в цій роботі розглядається безпліддя як захворювання пари.

Для аналізу установа НМАПО ім. П. Л. Шупика надала вибірку, що включає в себе інформацію генотипів подружньої пари та дані обтяженого родоводу. Дані використовувались з метою навчання моделі та використання її для прогнозу. Для побудови моделі використовувалися алгоритми машинного навчання: логістична регресія, дерево рішень, метод опорних векторів, нейронна мережа, логістична регресія для роботи з розрідженими матрицями.

Генотипи подружніх пар та обтяженість родоводу досліджувались окремо. Найкращі результати досягнуті при використанні логістичної регресії. Модель побудована на генотипах подружніх пар має кращі прогнозуючі властивості чим на основі обтяженості родоводу, тому для підвищення якості прогнозу безпліддя до генотипів додано критерій впливу спадковості. Через порівняння якостей моделі прогнозування з критерієм визначили найкращу шкалу значень для критерію.

На основі результатів дослідження розроблено програмний продукт, що дозволяє завантажувати дані з Exelдокументів, на їх основі будувати моделі, порівнювати їх продуктивність та використовувати для прогнозування. Додаток включає в себе функціональність передавання критерію впливу до вибірки даних.

Результати даного дослідження:

- Новий підхід аналізу хворих на безпліддя.
- Визначено найкращу модель для прогнозування безпліддя на основі генотипів та обробки генеалогічних даних.
- Розроблений програмний продукт реалізує всю необхідну функціональність для побудови та використання моделей прогнозування.

Містить інтуїтивний інтерфейс. Підвищує продуктивність лікаря, дозволяє підтвердити здогадки чи вплинути на заключення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. А. Т. Опря Статистика. — Київ: Центр учбової літератури, 2012. — 448 с. с. — ISBN 978-611-01-0266-7.

2. Kiss, F. Credit scoring processes from a knowledge Management perspective / F. Kiss // Periodica Polytechnica. Ser. Soc. Man. Scl. — 2003. — Vol. 11, № 1. — P. 95 — 110.

3. Кузнєцова, Н.В. Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування / Н.В. Кузнєцова, П.І. Бідюк // Наук. вісті НТУУ “ КПІ”. — 2010. — № 1. — С. 42 —53.

4. Г. Шилд – «C# 4.0: полное руководство». – М.: ООО «И.Д.Вильямс», 2011. – 1056 с.

5.НПАОП 0.00-1.28-10. Правила охорони праці під час експлуатації електронно-обчислювальних машин.

6. ДСН 3.3.6.042-99 Державнісанітарнінормимікрокліматувиробничихприміщень.

7. ДБНВ.2.5.-28-2006. Природне і штучне освітлення. Мінбуд України, 2006 – 76 с.

8. ДСН 3.3.6.037-99 Санітарнінормивиробничогошуму, ультразвукутаінфразвуку

9. ГОСТ 12.1.006-84(1999) Электромагнитные поля радиочастот. Допустимые уровни на рабочих местах и требования к проведению контроля.

10. В. С. БарановаГенетический паспорт – основа индивидуальной и предикативной медицины. —2009. — 528с.—ISBN: 5948690841.