ПРОГНОЗУВАННЯ ПСИХІЧНОГО ЗДОРОВ'Я НА ОСНОВІ ДАНИХ ПРО МУЗИЧНІ ВПОДОБАННЯ

Виконала: студентка IV курсу, групи КА-93
Ланько Анна Анатоліївна

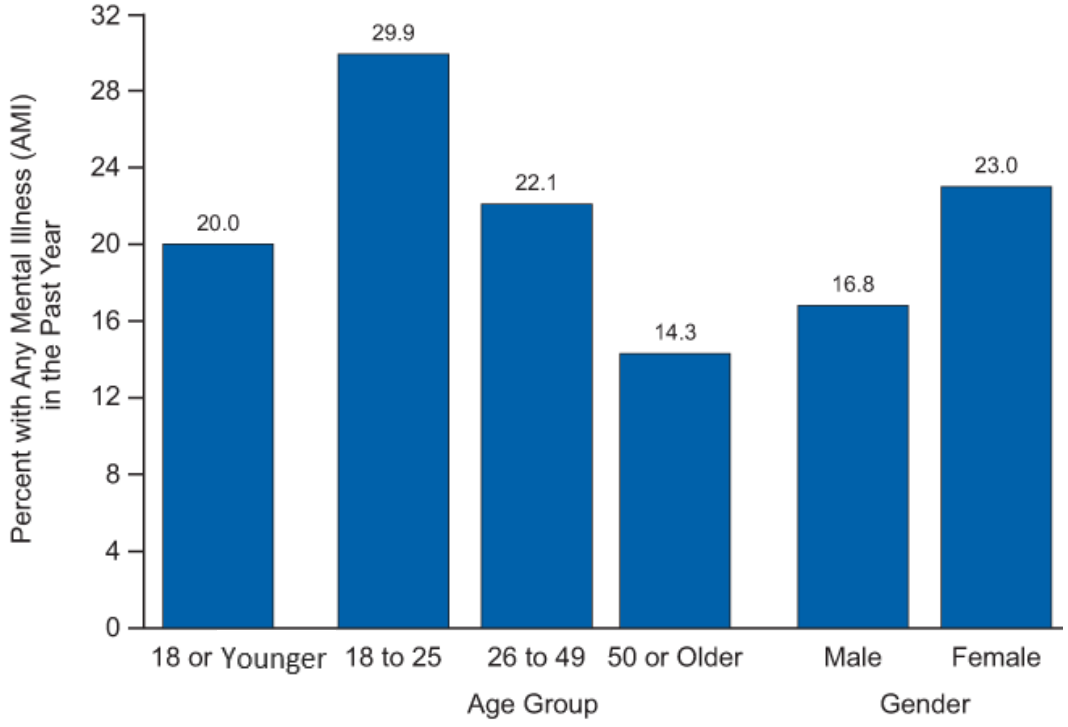
Керівник: д.т.н., доцент
Недашківська Надія Іванівна

АКТУАЛЬНІСТЬ ТЕМИ ДОСЛІДЖЕННЯ

Незначні ментальні порушення зазвичай виникають в умовах постійного стресу та ігноруються суспільством, у той час як **їх хронічна наявність послаблює нервову систему людини та поступово еволюціонує у більш серйозні розлади.**

Дане дослідження є інструментом попередження розвитку серйозних психічних аномалій шляхом своєчасного виявлення проблеми на основі загальнодоступних даних про музичні вподобання.

Вибір вхідних даних зумовлено науково обґрунтованим зв'язком між емоційним станом та характером музичних вподобань.

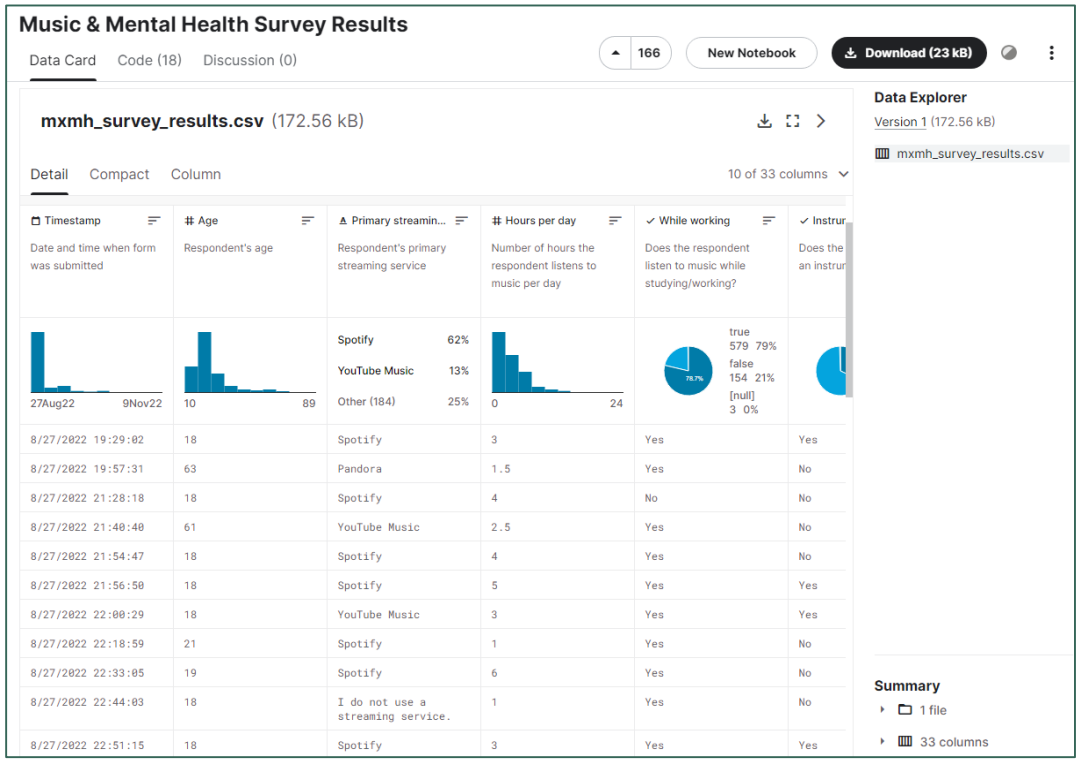


Відсоток людей з наявністю ментальних розладів за віковими та гендерними категоріями у 2022 р.

ВХІДНІ ДАНІ

Дані було зібрано користувачем відкритої платформи для обміну досвідом та матеріалами у галузі машинного навчання *Kaggle* двома способами:

1) **опитування у мережі** (посилання на гугл-форму в соціальних мережах, на форумах та Discord-серверах);
2) **«живе» опитування** (на вулицях, у громадських місцях).



ПОСТАНОВКА ЗАДАЧІ

Суть задачі полягає в **прогнозуванні ступеня наявності тривожності, безсоння, депресії та обсесивно-компульсивного розладу в певної особи у вигляді оцінки в діапазоні від 0 до 10 на основі 27 категоріальних та числових ознак.**

Формалізована задача:
Нехай X – множина ознак, а Y – множина цільових змінних. Існує невідома цільова залежність (відображення) $y^*: X \to Y$, значення якої відомі лише для об'єктів скінченної вибірки $X^m = \{(x_1; y_1), \dots, (x_m; y_m)\}$. Потрібно побудувати функцію $f: X \to Y$, яка допоможе прогнозувати оцінки станів для довільного зразка $x \in X$.

Оскільки прогнозується числове значення, а не належність до класу, **маємо задачу регресії**.

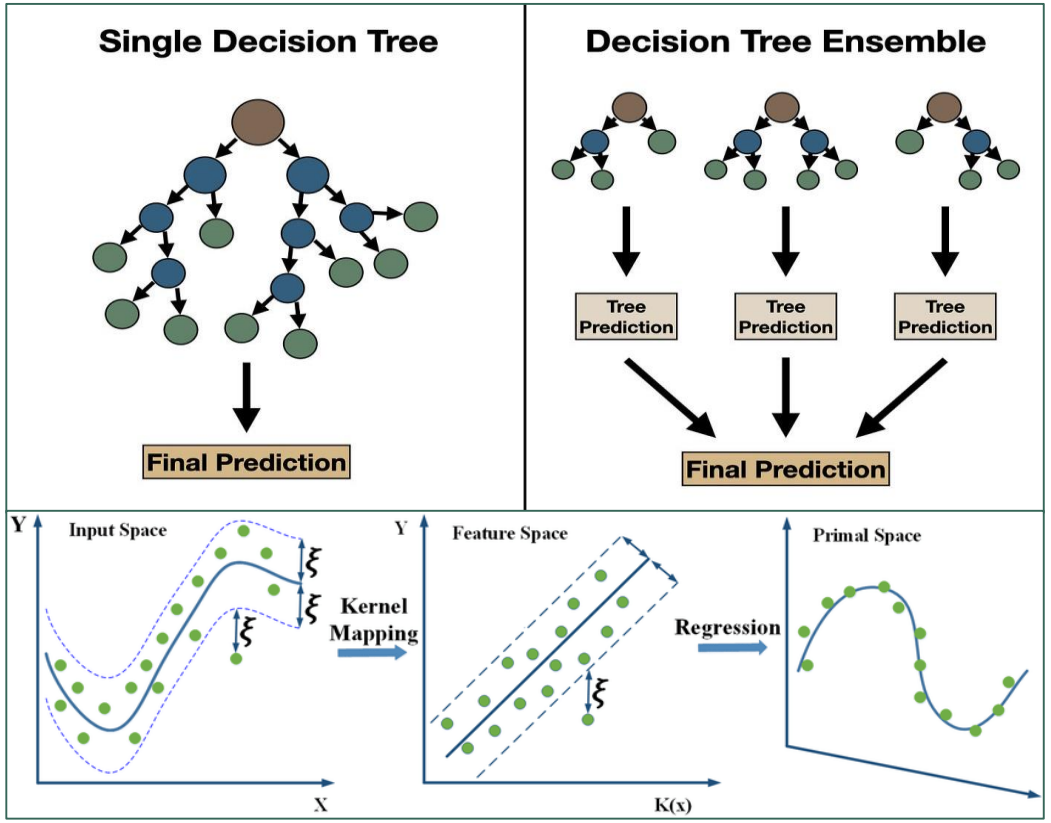
X			
№	«Age»	…	«Music effect»
1			
…			
736			

Y			
№	«Anxiety»	…	«OCD»
1			
…			
736			

ОГЛЯД МЕТОДІВ РОЗВ’ЯЗАННЯ ЗАДАЧІ

З огляду на складність вхідних даних, було розглянуто більш складні методи розв’язання задачі регресії:

1) **ансамблеві методи** – комбінація кількох дерев рішень (прогнозують усереднене значення за відповідним листовим вузлом) для покращення точності прогнозування та уникнення перенавчання;
2) **метод опорних векторів** – регресія (SVR) на основі знаходження гіперплощини, яка найкраще описує дані у просторі ознак, підходить для даних високої розмірності.



Концепція ансамблів (зверху) та SVR (знизу)

ПОПЕРЕДНЯ ОБРОБКА ДАНИХ

Попереднє форматування

↓

Видалення аномалій

↓

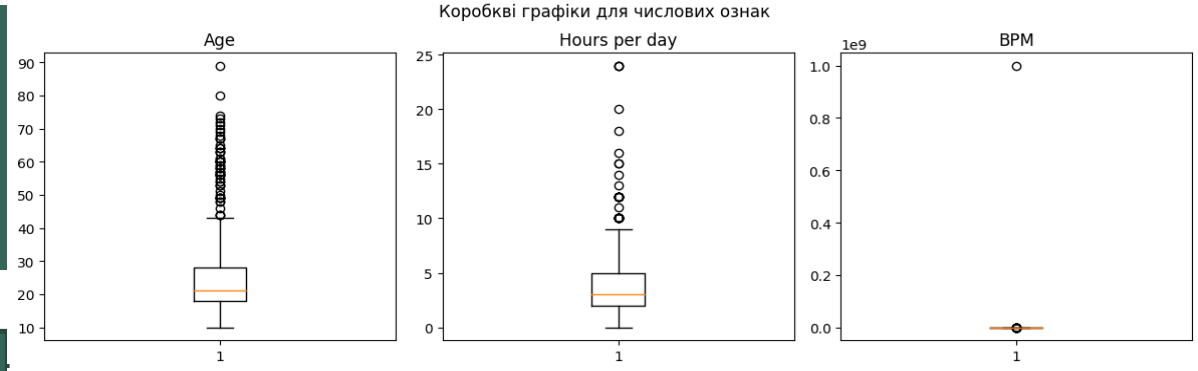
Заповнення пропусків

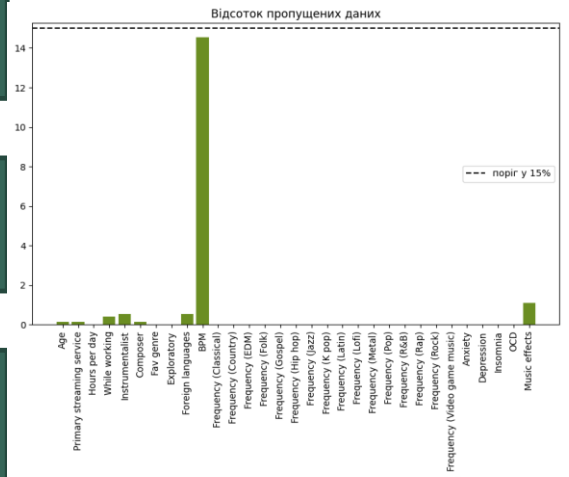
↓

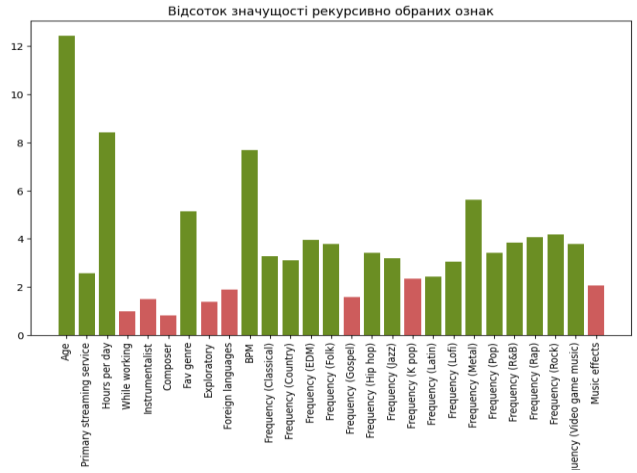
Вибір значущих ознак

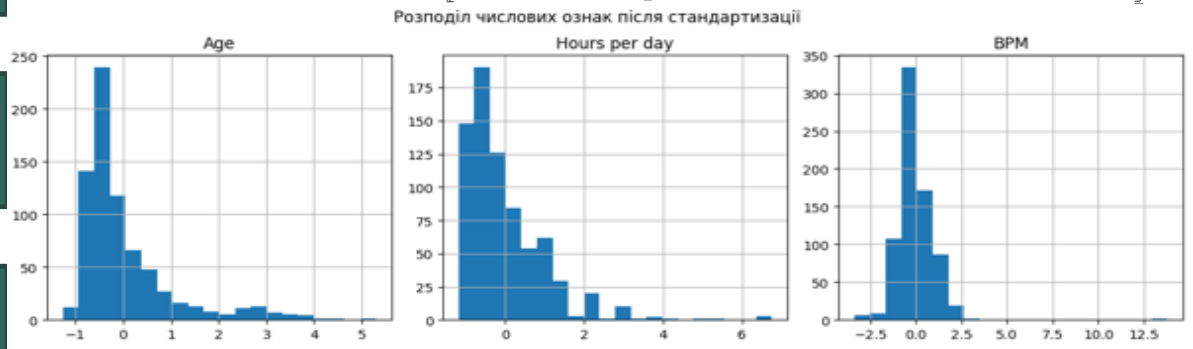
↓

Кодування та масштабування









Аналітичні візуалізації

НАЛАШТУВАННЯ ГІПЕРПАРАМЕТРІВ МОДЕЛЕЙ

1. З наявного переліку гіперпараметрів для кожної моделі було обрано найбільш суттєві.

2. Для кожного з них було визначено найдоцільніші варіанти значень.

3. Таким чином отримали множину комбінацій для кожної моделі.

Результат:
За найкращий набір гіперпараметрів для кожної моделі було обрано конфігурацію з мінімальною середньоквадратичною похибкою (MSE) на валідаційній вибірці.

Назва гіперпараметра	Розглянуті значення	Найкраще значення
n_estimators	50, 100, 150	50
max_samples	0.6, 0.8, 1.0	0.8
max_features	0.6, 0.8, 1.0	0.8
bootstrap	True, False	False
bootstrap_features	True, False	False

Приклад підбору для найкращої моделі – **BaggingRegressor**

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Найкраща модель – **BaggingRegressor**

Модель	Метрики якості						Час навчання, с
	RMSE		MAE		R2		
	train	test	train	test	train	test	
SVR	2,565	3,149	1,865	2,513	0,212	-0,173	0,03
ExtraTreesRegressor	2,715	2,905	2,313	2,458	0,154	-0,012	0,16
BaggingRegressor	2,737	2,894	2,331	2,448	0,141	-0,004	12,38
XGBRegressor	1,402	3,008	1,147	2,511	0,774	-0,008	2,05
LGBMRegressor	2,878	2,979	2,426	2,485	0,013	-0,03	0,02
CatBoostRegreesor	2,38	3,031	1,983	2,517	0,323	-0,068	0,04
StackingCVRegressor	2,788	2,975	2,347	2,48	0,067	-0,029	6,14

СПРОБИ ПОКРАЩЕННЯ МОДЕЛЕЙ

Можливі причини низької результативності:

1) використання некоректних методів розбиття даних;
2) неправильне застосування метрик;
3) неправильна інтерпретація задачі;
4) складність, слабка кореляція даних.

Висновок:
Результат експериментів довів, що низькі значення метрик викликані складністю вхідних даних та суб'єктивністю оцінок.

1. Перехресна перевірка
(результат мінімально покращився)

2. Округлення результатів регресії
(результат не змінився)

3. Класифікація оцінок
(низька результативність)

АЛЬТЕРНАТИВНІ ЗАДАЧІ

З метою покращення результативності дослідження було розглянуто наступні альтернативні задачі:

1. **Класифікація ментального стану за рівнями наявності розладів:**
   1) 0-3 – «Низький»;
   2) 4-7 – «Середній»;
   3) 8-10 – «Високий».
2. **Пошук асоціативних правил.**

Висновок:
Точність такої класифікації за всіма моделями варіюється в межах 50%±3% і є практично прийнятною для складних даних, побудовані асоціативні правила є слабкими за змістом.

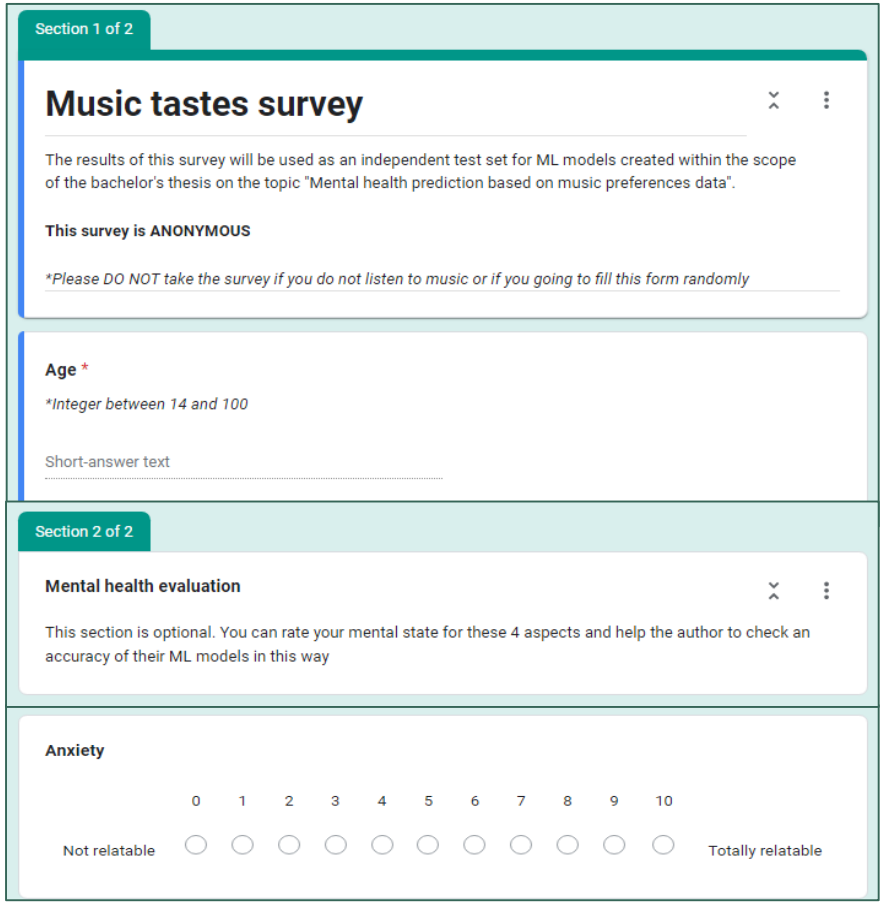
№	Набір чинників тривожності
1	«While working»: так, «Exploratory»: так, «Primary streaming service»: Spotify, «Music effects»: покращує
2	«While working»: так, «Primary streaming service»: Spotify, «Frequency (Gospel)»: ніколи, «Music effects»: покращує
3	«While working»: так, «Exploratory»: так, «Frequency (Gospel)»: ніколи, «Music effects»: покращує

Приклад найпоширеніших умов (X) для асоціативного правила $X \rightarrow Y$, де Y– тривожність

СТВОРЕННЯ ВЛАСНОГО ОПИТУВАЛЬНИКА

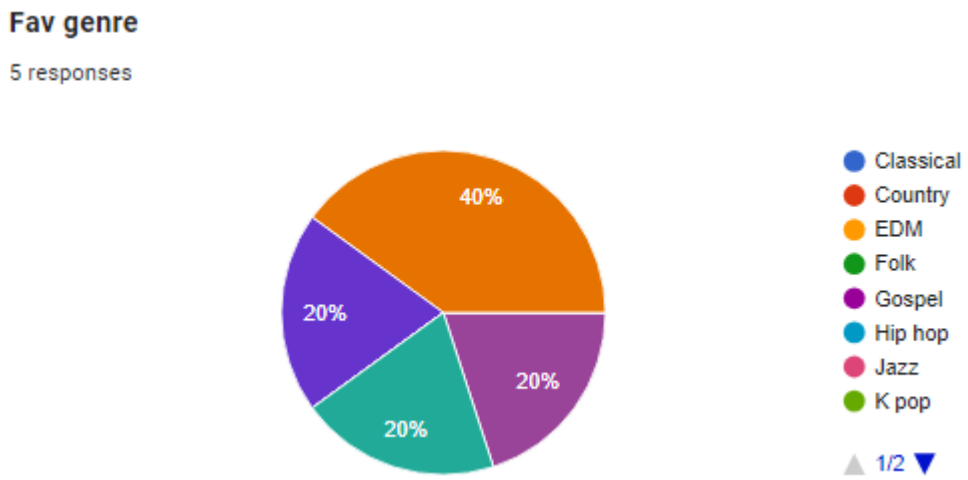
Анонімний опитувальник містить лише значущі для алгоритму питання та, на відміну від вхідних даних, має дві секції:

1) **обов'язкові питання** – ознаки, що будуть використані для прогнозу;
2) **вибіркові питання** – респондент за бажанням міг оцінити свій ментальний стан; ці відповіді допоможуть оцінити адекватність спрогнозованої оцінки.



РЕЗУЛЬТАТИ НА ВЛАСНИХ ЗНАЧЕННЯХ

На опитування дало відповідь 5 осіб, кожній особі присвоєний ID від 1 до 5.



Приклад відповідей на питання про улюблений жанр

ID	Тривожність		Депресія		Безсоння		ОКР	
	pred	true	pred	true	pred	true	pred	true
1	5,5	4	3	2	6	2	1	2
2	5,5	4,5	3	2,5	1	0	3	0
3	5,5	4	3	2	1	0	0	0
4	5,5	4,5	3,5	2,5	1	0	7	4
5	5	3,5	3	2	8	6	5	6

Регресія для найкращої моделі – **BaggingRegressor**

Висновок:
Алгоритм не є усередненим, відслідковує зміни значень. Результат є практично прийнятним

ВИСНОВКИ

1. **Найкращою моделлю є бегінг моделей особливо випадкового лісу.** Це пояснюється випадковістю розбиттів та незалежним навчанням дерев з подальшою агрегацією прогнозу, що попереджує перенавчання і допомагає знаходити складні нелінійні залежності між даними.
2. З огляду на гарний прогноз на власній вибірці та достатню якість модифікованої класифікації, **задачу можна вважати успішно розв'язаною, а розроблений алгоритм – практично прийнятним з точки зору результативності.**
3. **Недоліком розробленого алгоритму є нижчі за очікувані показники метрик,** що пояснюється складною залежністю всередині даних та суб'єктивністю визначення фактичних значень цільових змінних.
4. **Покращити його можна шляхом застосування нейронних мереж,** що не було реалізовано в даній роботі з огляду на замалий обсяг даних та масштаб предмету дослідження.

ПЕРСПЕКТИВИ ДОСЛІДЖЕННЯ

Дані про музичні вподобання легко зібрати у великому обсязі на основі персональної інформації користувачів стрімінгових платформ.

Отриманий прогноз на основі повідомлень, рекомендацій та обмежень може **допомогти конкретній особі звернути увагу на свій ментальний стан і тим самим попередити розвиток психічних порушень**.

Щодо вищезгаданих ресурсів, розроблений програмний продукт допоможе **налагодити рекомендаційні системи та створити більш здорове середовище у спільноті**.

ДЯКУЮ ЗА УВАГУ!