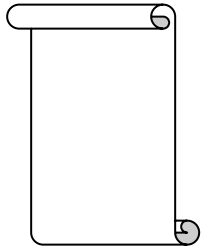
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. І. СІКОРСЬКОГО»

МЕТОДИ СКОРИНГОВОГО АНАЛІЗУ ДАНИХ ПОЗИЧАЛЬНИКІВ КРЕДИТІВ У ФІНАНСОВИХ УСТАНОВАХ

ДОПОВІДАЧ:
БОЙКО ІРИНА ОЛЕКСАНДРІВНА

НАУКОВИЙ КЕРІВНИК:

Д.Т.Н., ПРОФЕСОР БІДЮК П.І.



Мета, об'єкт, предмет та методи дослідження

- **Мета дослідження:** підвищення якості моделей для оцінювання кредитоспроможності шляхом удосконалення системної методології побудови моделей та комплексного застосування кількох методів інтелектуального аналізу даних.
- **Об'єкт дослідження:** кредитні історії позичальників, представлені статистичними даними та експертними оцінками.
- **Предмет дослідження:** математичні методи і моделі інтелектуального аналізу даних для опису характеристик позичальників кредитів у фінансових установах.
- **Методи дослідження:** логістична регресія, мережі Байєса та дерева рішень.

Завдання дослідження

Мета дослідження – підвищення якості моделей для оцінювання кредитоспроможності шляхом удосконалення системної методології побудови моделей та комплексного застосування кількох методів інтелектуального аналізу даних. Для досягнення мети потрібно вирішити такі завдання:

1. зібрати статистичні дані стосовно характеристик позичальників кредитів для виконання обчислювальних експериментів;

2. вибрати типи математичних моделей і методів інтелектуального аналізу даних для формального опису позичальників кредитів;

3. побудувати вибрані типи моделей за допомогою статистичних даних та оцінити якість цих моделей;

4. виконати порівняльний аналіз отриманих результатів.

Особливості фінансових процесів

Фінансові процеси – це динамічні стохастичні процеси, які залежать від великої кількості різноманітних факторів та впливів процесів різної природи; мають велику кількість обмежень, характеризуються різкою зміною факторів; складні в управлінні.

Проблеми аналізу фінансових даних

1) задачі аналізу фінансових процесів відносяться до важко формалізованих або слабо структурованих;
2) характерні неповнота, неоднозначність, невизначеність як вхідних даних, так і правил їх перетворення;
3) важко визначити єдиний формальний критерій вирішення задачі;
4) результати аналізу носять як кількісний, так і якісний характер;
5) немає чітких рекомендацій, які методи і підходи краще застосовувати для яких конкретних задач;
6) класичні методи і підходи виявляються неефективними для аналізу великих складних фінансових систем в умовах динамічних змін середовища.

Методи оцінювання кредитоспроможності позичальника

Експертний: знання, досвід спеціалістів-експертів в області кредитування; встановлення ваг та балів.

Скоринговий: побудова математичної моделі кредитоспроможності позичальників на основі кредитних історій банку та оцінюванні ймовірності дефолту потенційного позичальника.

Математичний апарат побудови скорингових функцій:

- Логістична регресія
- Дерева рішень
- Мережа Байєса
- Нейронні мережі
- Кластерний аналіз
- Інші методи

Перехресне тестування

Перехресне тестування (*cross-validation*) – група методів тестування статистичних моделей, що передбачає перебір різних однорідних варіантів розбиття на навчальну та тестову вибірки з подальшою побудовою множини статистичних моделей реалізованих на різних підмножинах з метою усереднення отриманих результатів по всіх тестових вибірках, що сприяє послабленню залежності отриманих результатів валідації від способу розбиття вибірки на дві підвибірки.

Оцінка якості моделей
1. Загальна точність моделі

Загальна точність моделі (CA – common accuracy) визначається як

$$CA = \frac{CorrectForecast}{N}$$

де *CorrectForecast* – кількість вірно спрогнозованих випадків,
N – загальна кількість випадків.

Загальна точність моделі є дещо суб’єктивною оцінкою, оскільки вона залежить від долі дефолтів в моделі, а також від порогу відсікання. Для різних значень порогу точність моделі також буде приймати різні значення

Оцінка якості моделей
2. Помилки I-го та II-го роду

	Прогноз моделі: Повернення кредиту (0)	**Прогноз моделі: Дефолт (1)**
Фактично: Повернення кредиту(0)	Вірно класифіковані (TP)	**Помилки II-го роду (FN)**
Фактично: Дефолт (1)	**Помилки I-го роду (FP)**	**Вірно класифіковані (TN)**

Для аналізу якостей моделі найчастіше використовують такі відносні показники у процентах:

- частка істинно позитивних прикладів (True Positives Rate): $TPR = \frac{TP}{TP + FN}$

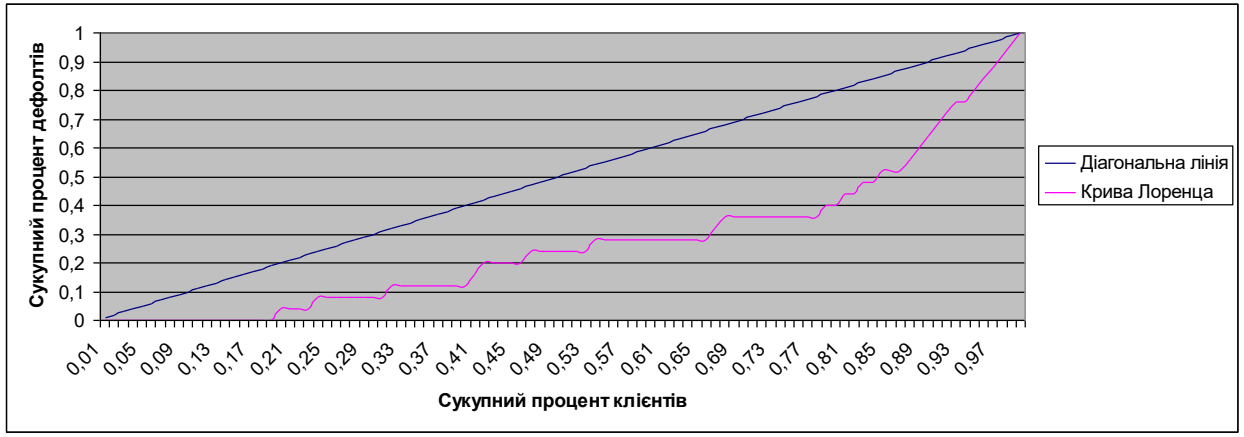
- частка хибно позитивних прикладів (False Positives Rate): $FPR = \frac{FP}{TN + FP}$

Оцінка якості моделей
3. ROC-крива та індекс Джіні

Для побудови графіка ROC-кривої по осі Y відкладається значення чутливості Se, а по осі X – долю хибно позитивних випадків FPR або 1-Sp, де:

$$Se = TPR = \frac{TP}{TP + FN} \qquad Sp = \frac{TN}{TN + FP} \qquad Sp = \frac{TN + FP - FP}{TN + FP} = 1 - \frac{FP}{TN + FP} = 1 - FPR$$

Індекс GINI – це площа області між діагоналлю і кривою Лоренца, поділена на площу усієї області під діагоналлю. Індекс GINI широко використовується для аналізу роздільної здатності системи оцінювання при управлінні кредитними ризиками, тобто оцінки здатності моделі розділяти клієнтів на схильних та несхильних до дефолту. Очевидно, що 80% дефолтів, що спостерігалися, зустрічаються саме серед останніх 20% клієнтів. Отже, чим більша область між діагоналлю та кривою, тим вища якість моделі.



Діагональна лінія – це лінія «випадкової моделі». На осі X клієнти впорядковані за заданою ймовірністю дефолту. Індекс GINI можна визначити через площу фігури, що знаходиться під ROC-кривою, таким чином:

$$GINI = 2 \cdot AUC - 1 \qquad 0 \le G \le 1$$

Скорингова функція на основі нелінійної регресії

У лінійній регресії припускається, що залежна змінна є лінійною функцією незалежних змінних, тобто: $y = b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + u$, де

y – залежна змінна (результат прийняття рішення); x_i – пояснююча змінна (критерій); b_i – вага пояснюючої змінної і; u – випадкова похибка.

У випадку застосування **бінарної логістичної регресії** ймовірність p входження клієнта у стан дефолту D пов'язана із змінними $(X_1,\ldots,X_n)$, таким чином:

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

Ймовірність дефолту клієнта представляється у зручнішій формі:

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

Мережі Байєса

Мережі Байєса дозволяють:

- поєднати графічне представлення певного процесу з його ймовірнісним характером;
- проаналізувати можливі варіанти розвитку процесу;
- відстежити правильність встановлення причинно-наслідкового зв’язку між окремими факторами;
- підвищити обґрунтованість рішень при аналізі складних проблемних ситуацій.

Етапи побудови і застосування мережі Байєса:

1) аналіз взаємозв’язків між змінними;
2) навчання структури мережі;
3) формування ймовірнісного висновку;
4) застосування мережі до реальних даних.

Застосування дерев рішень

Дерева рішень – метод автоматизованого аналізу даних, коли правила представляються у вигляді послідовної ієрархічної структури, в якій кожному вузлу відповідає єдиний вузол, що дає рішення. Правило – це конструкція виду: «**якщо** ... , **то**...».

Відомі алгоритми, що реалізують дерева рішень: CART,C4.5, CHAID, CN2, тощо. CART (Classification And Regression Tree) – алгоритм побудови бінарного дерева рішень, кожний вузол якого при розбитті має лише двох нащадків. Для вибору оптимального правила розбиття дерева використовується функція оцінки якості розбиття:

$$GINI(T) = 1 - \sum_{i=1}^{n} p_i^{\,2} \rightarrow \min$$

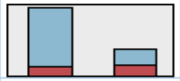
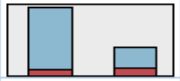
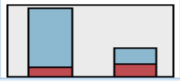
Побудова моделей

Для побудови моделі використана навчальна вибірка з 1600 записів, які містять інформацію про клієнтів (вік, кількість дітей, заробітна плата, місце роботи, сума кредиту, тощо) та інформацію – були кредити повернуті чи ні (0 – клієнт повернув кредит, 1 – дефолт). В навчальній вибірці 75% (1150 кредитів) були повернуті, а 25% (тобто 450 кредитів) – не були повернуті. Побудована модель перевіряється на вибірках з 500 записів, з яких: 90% – кредити повернуті, 10% – дефолти; 50% – кредити повернуті, 50% – дефолти . Для перевірочної вибірки по кожному клієнту обчислюємо ймовірність дефолту та порівнюємо спрогнозовані значення з фактичною інформацією, чи був повернутий кредит (0 – якщо кредит повернули, 1 – в іншому випадку). Незалежно від того, яка саме технологія моделювання використовується, результатом буде ймовірність несплати (дефолту), визначена для кожного клієнта у портфелі.

Для аналізу використаємо моделі оцінювання кредитоспроможні позичальника, побудовані на основі методів логістичної регресії, дерев рішень та мереж Байєса, обчислюється загальна точність кожної моделі, визначається рівень значущості характеристик клієнтів, індекси AUC та GINI, ROC-крива. Також використовуємо підхід перехресного тестування з вибіркою training 70%(90%) записів та test-вибіркою 30%(10%) записів.

На основі отриманих результатів робляться висновки стосовно доцільності і перспективності цих моделей при оцінюванні кредитоспроможності позичальників.

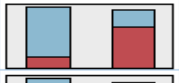
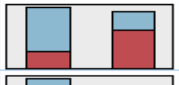
Загальна точність побудованих моделей

Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30...	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	Logistic regression 1	< 1	115,0	17	1,771	73,737	12	0,739
	C5 1	< 1	47,400	11	1,498	73,131	11	0,608
	Bayesian Network 1	< 1	70,0	21	1,798	71,111	12	0,725

Для вибірки незміненої структури

Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	C5 1	< 1	625,0	100	1,0	92,517	12	0,5
	Bayesian Network 1	< 1	635,0	98	1,049	87,755	12	0,777
	Logistic regression 1	< 1	635,0	98	1,081	87,075	12	0,779

Для вибірки зміненої структури (90%-повернень; 10%-дефолту)

Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	Logistic regression 1	< 1	205,0	43	1,752	77,124	12	0,829
	C5 1	< 1	152,456	49	1,438	70,588	6	0,703
	Bayesian Network 1	< 1	190,0	57	1,539	66,013	12	0,808

Для вибірки зміненої структури (50%-повернень; 50%-дефолту)

Порівняння отриманих результатів

Назва методу	Індекс Джіні	Значення AUC	Помилка 1-го роду	Помилка 2-го роду	Точність моделі	Якість моделі
Логістична регресія	0,595	0,796	0,07	0,16	77,5%	Прийнятна
Байєсівські мережі	0,736	0,868	0,08	0,11	81,56%	Дуже висока
Дерева рішень	0,51	0,755	0,05	0,016	79,81%	Прийнятна

- для вибірки незміненої структури

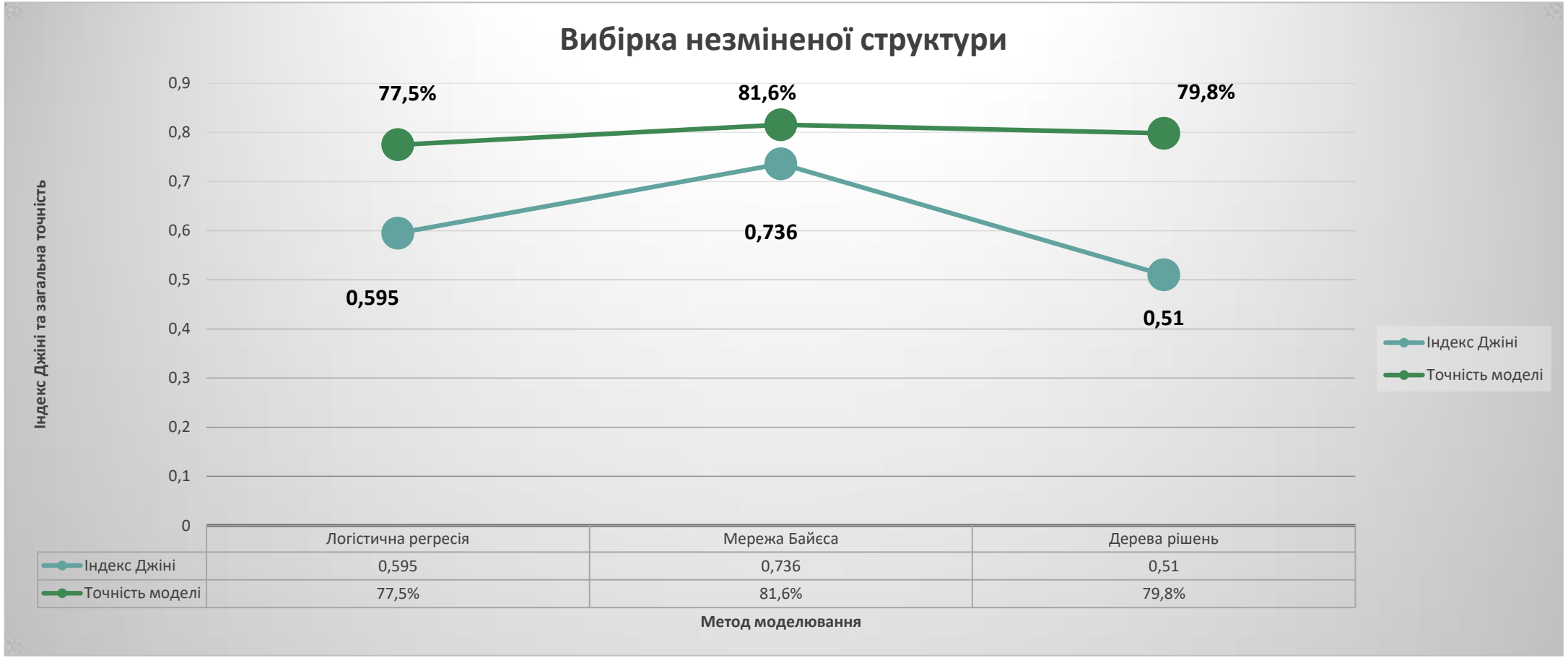
Назва методу	Індекс Джіні	Значення AUC	Помилка 1-го роду	Помилка 2-го роду	Точність моделі	Якість моделі
Логістична регресія	0,804	0,902	0,08	0,016	90,8%	Відмінна
Байєсівські мережі	0,928	0,964	0,03	0,016	94,6%	Відмінна
Дерева рішень	0	0,5	0,1	-	90%	Незадовільна

- для вибірки зміненої структури (90%-повернень; 10%-дефолту)

Назва методу	Індекс Джіні	Значення AUC	Помилка 1-го роду	Помилка 2-го роду	Точність моделі	Якість моделі
Логістична регресія	0,671	0,835	0,12	0,12	75,2%	Дуже висока
Байєсівські мережі	0,807	0,904	0,08	0,09	83%	Відмінна
Дерева рішень	0,784	0,892	0,09	0,07	84,4	Дуже висока

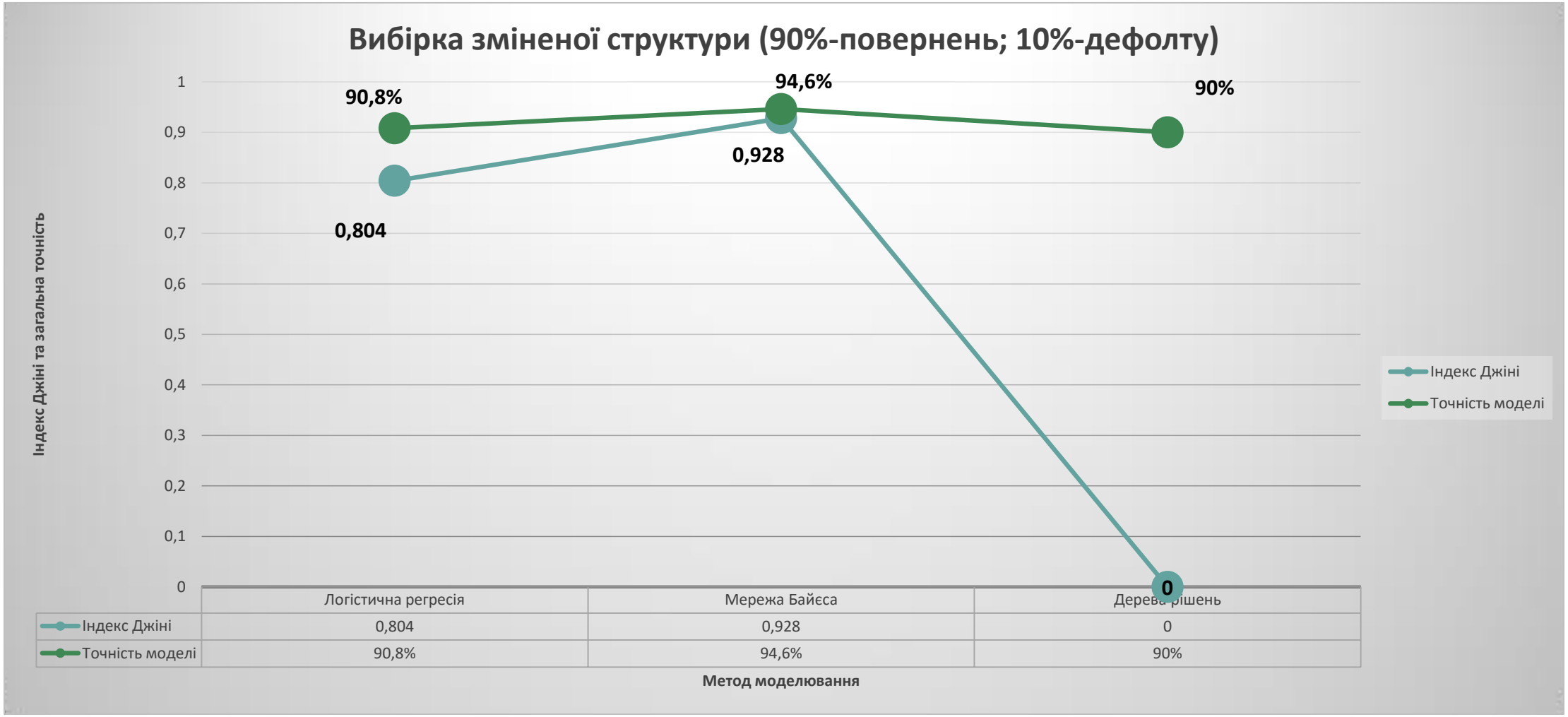
- для вибірки зміненої структури (50%-повернень; 50%-дефолту)

Індекс Джіні та загальна точність моделей



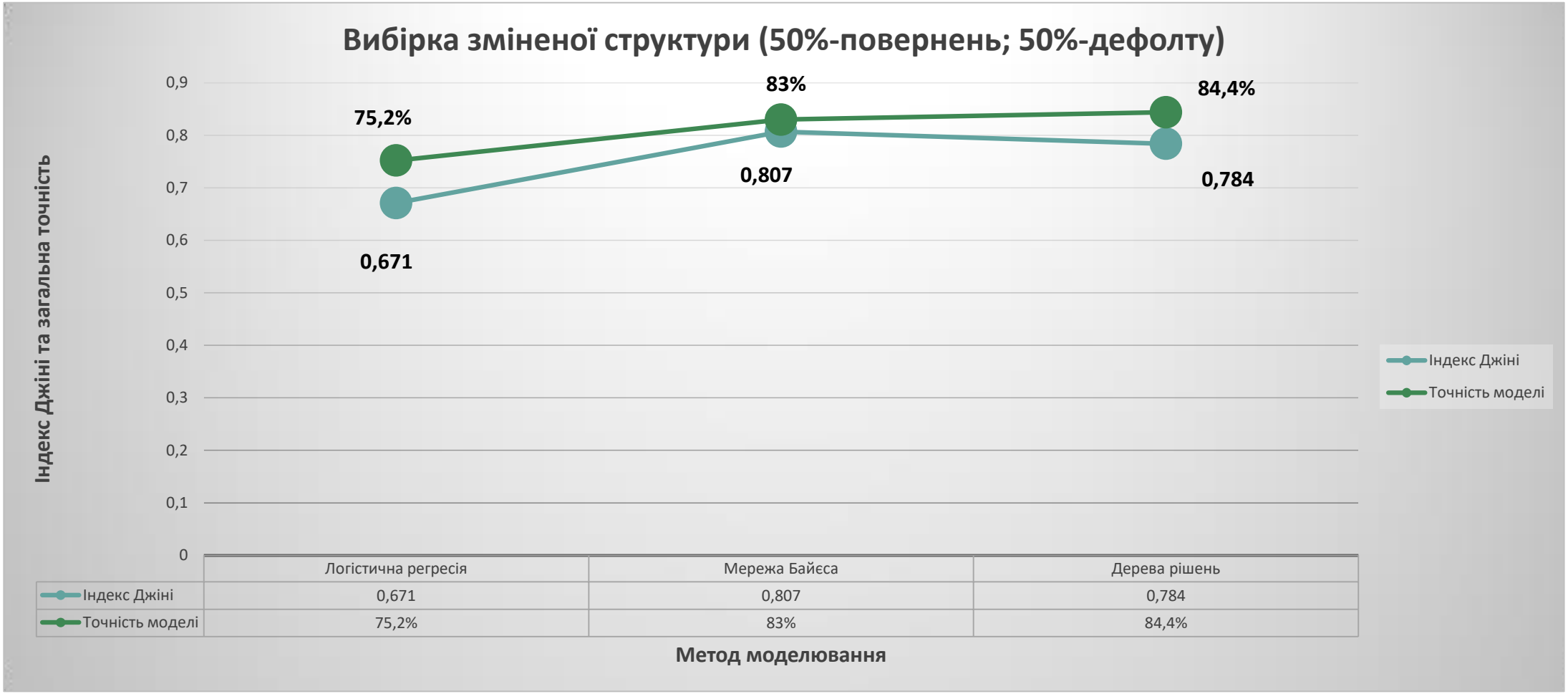
	Логістична регресія	Мережа Байєса	Дерева рішень
Індекс Джіні	0,595	0,736	0,51
Точність моделі	77,5%	81,6%	79,8%

Індекс Джіні та загальна точність моделей для вибірки зміненої структури



	Логістична регресія	Мережа Байєса	Дерева рішень
Індекс Джіні	0,804	0,928	0
Точність моделі	90,8%	94,6%	90%

Індекс Джіні та загальна точність моделей



Порівняння отриманих результатів з перехресним тестуванням (70%-навчальної; 30%-тестової вибірки)

Назва методу	Індекс GINI		Значення AUC		Помилка 1-го роду		Помилка 2-го роду		Точність моделі		Якість моделі
	train	test	train	test	train	test	train	test	train	test	
Логістична регресія	0,607	0,507	0,804	0,753	0,07	0,09	0,16	0,15	77,19%	75,56%	Дуже висока
Мережа Байєса	0,753	0,405	0,876	0,702	0,08	0,1	0,1	0,14	81,54%	71,11%	Дуже висока
Дерева рішень	0.412	0.203	0.706	0.601	0,02	0,07	0,18	0,2	79.82%	73.13%	Середня

- для вибірки незміненої структури

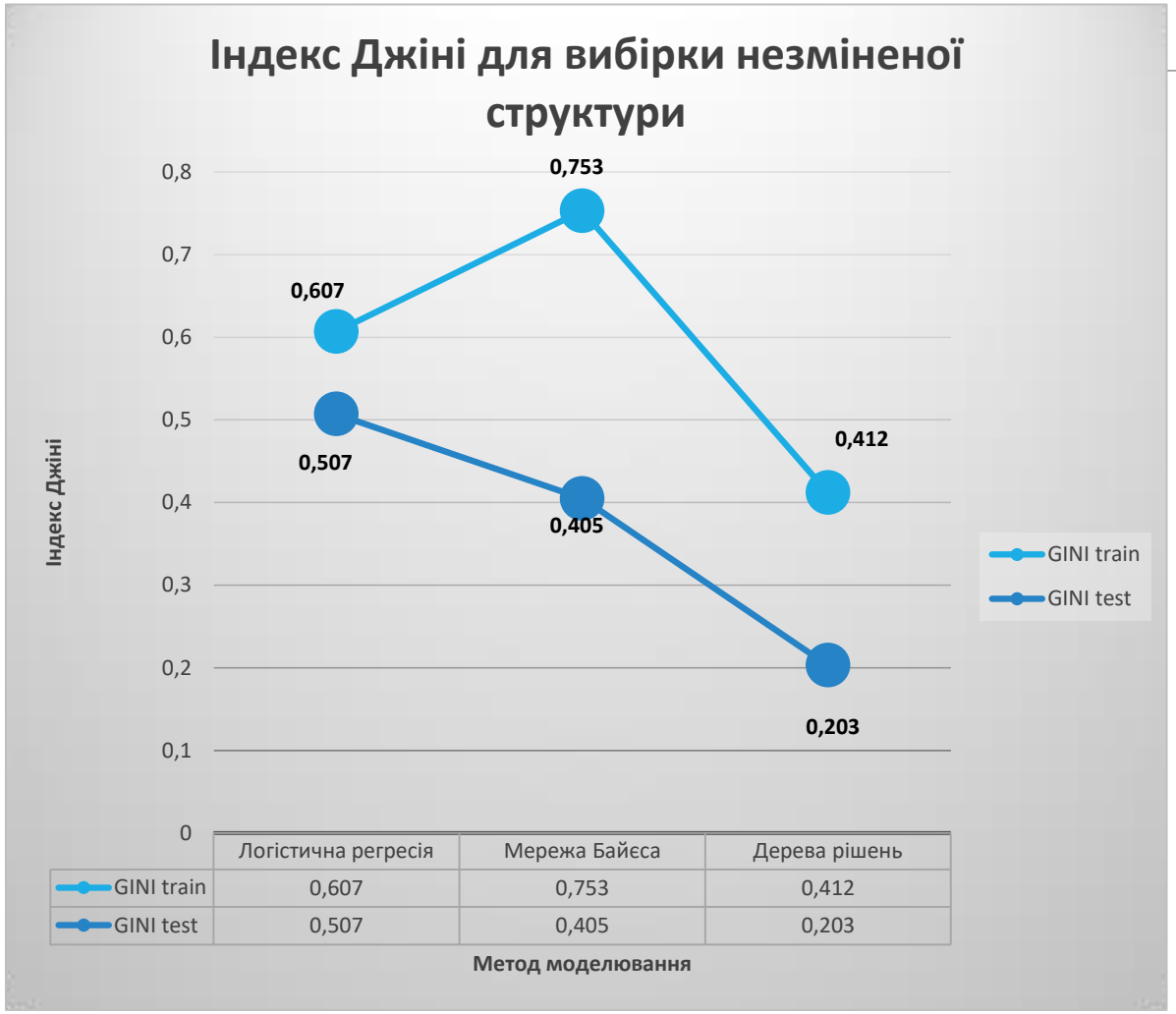
Назва методу	Індекс GINI		Значення AUC		Помилка 1-го роду		Помилка 2-го роду		Точність моделі		Якість моделі
	train	test	train	test	train	test	train	test	train	test	
Логістична регресія	0,881	0,41	0,94	0,705	0,05	0,05	0,017	0,06	93,2%	86,39%	Прийн.
Мережа Байєса	0,976	0,536	0,988	0,768	0,025	0,05	0,014	0,007	96,03%	87,76%	Дуже висока
Дерева рішень	0	0	0,5	0,5	0,11	0,07	-	-	88,95%	92,52%	Незадов.

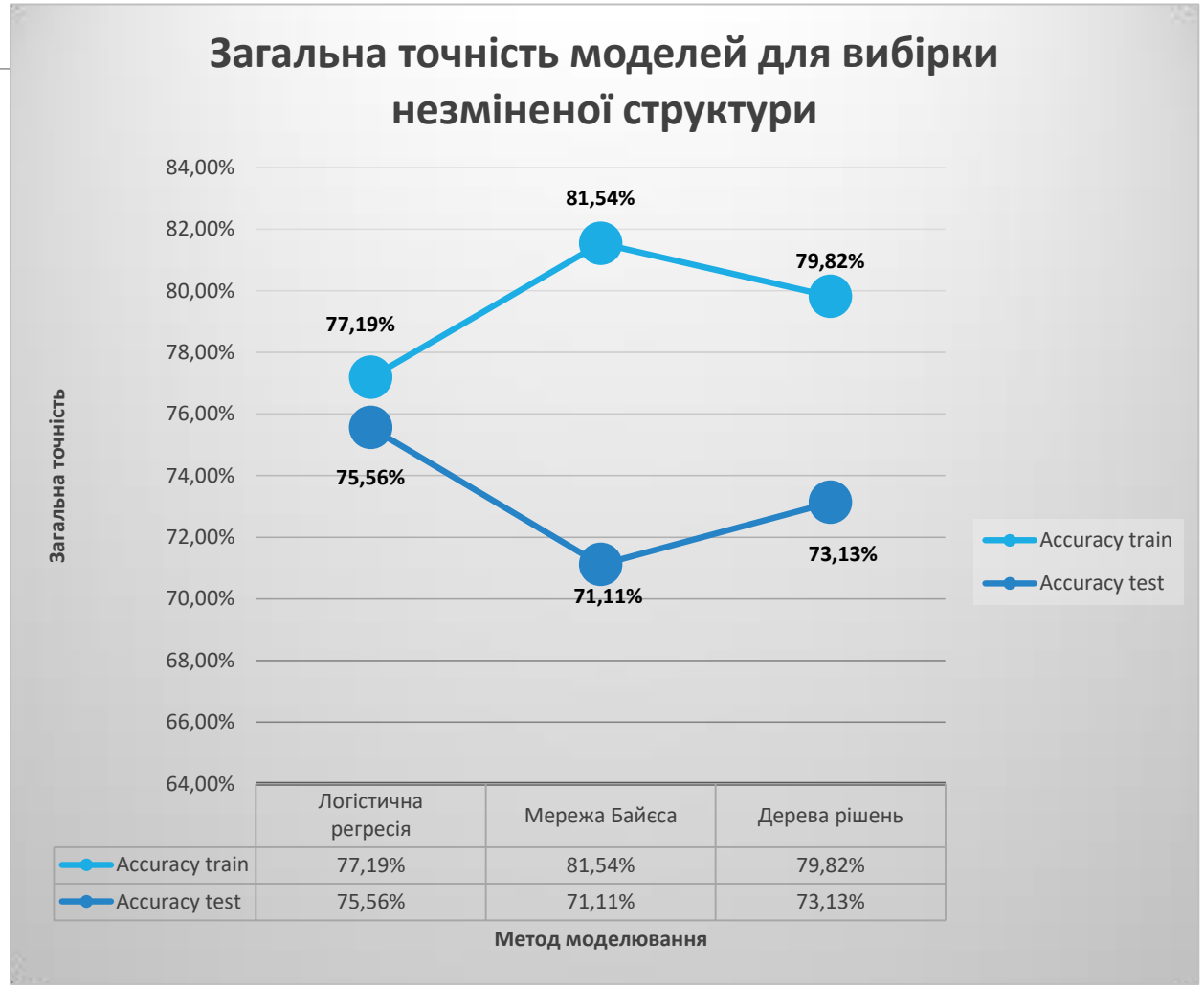
- для вибірки зміненої структури (90%-повернень; 10%-дефолту)

Назва методу	Індекс GINI		Значення AUC		Помилка 1-го роду		Помилка 2-го роду		Точність моделі		Якість моделі
	train	test	train	test	train	test	train	test	train	test	
Логістична регресія	0,612	0,718	0,806	0,859	0,14	0,07	0,14	0,12	71,76%	80,39%	Дуже висока
Мережа Байєса	0,799	0,449	0,899	0,725	0,1	0,11	0,09	0,1	81,56%	66,01%	Дуже висока
Дерева рішень	0,556	0,347	0,778	0,673	0,13	0,14	0,11	0,15	75,5%	70,59%	Середня

- для вибірки зміненої структури (50%-повернень; 50%-дефолту)

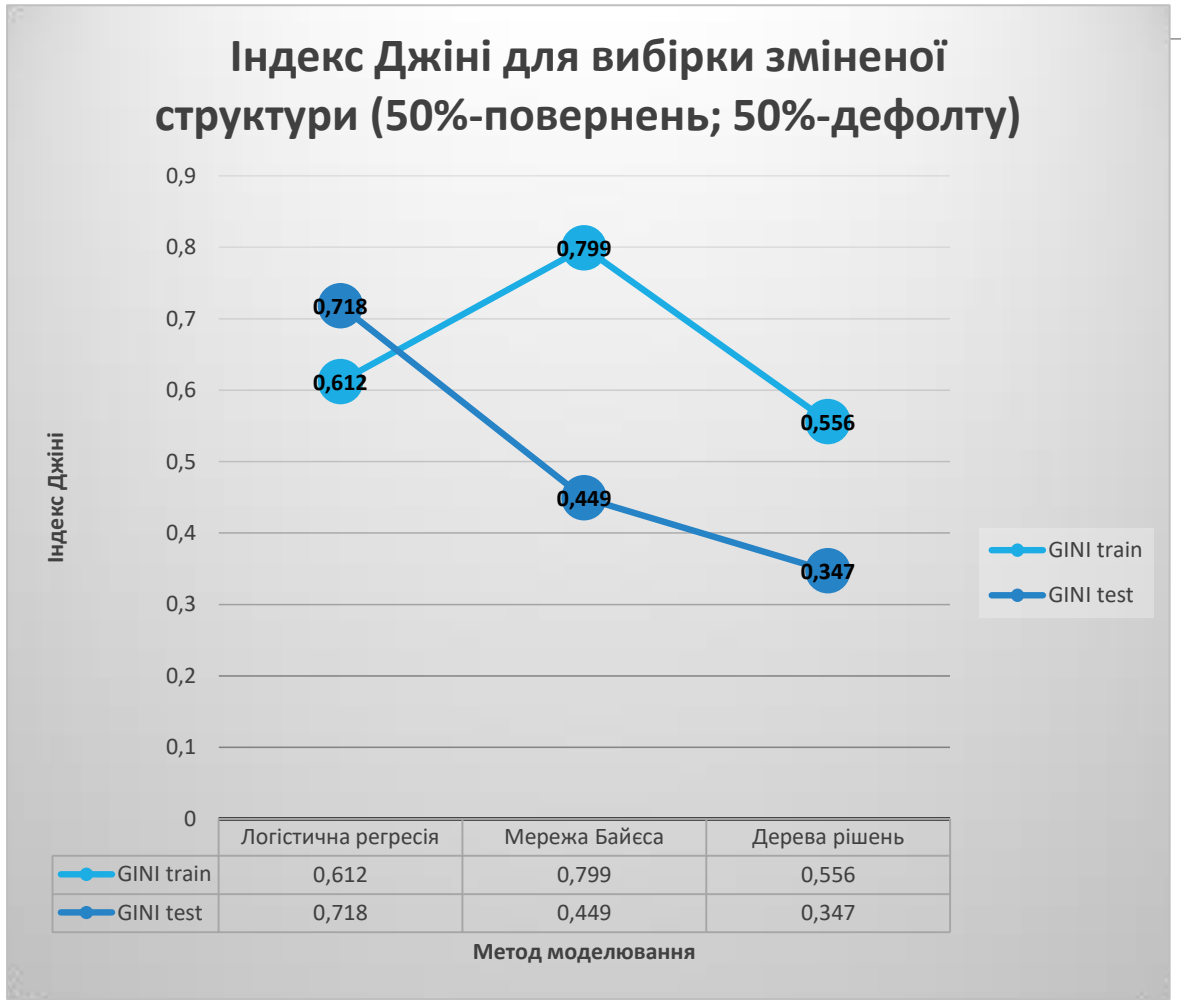
Порівняння індексу Джіні та загальної точності моделей з перехресним тестуванням (70%-навчальної; 30%-тестової)

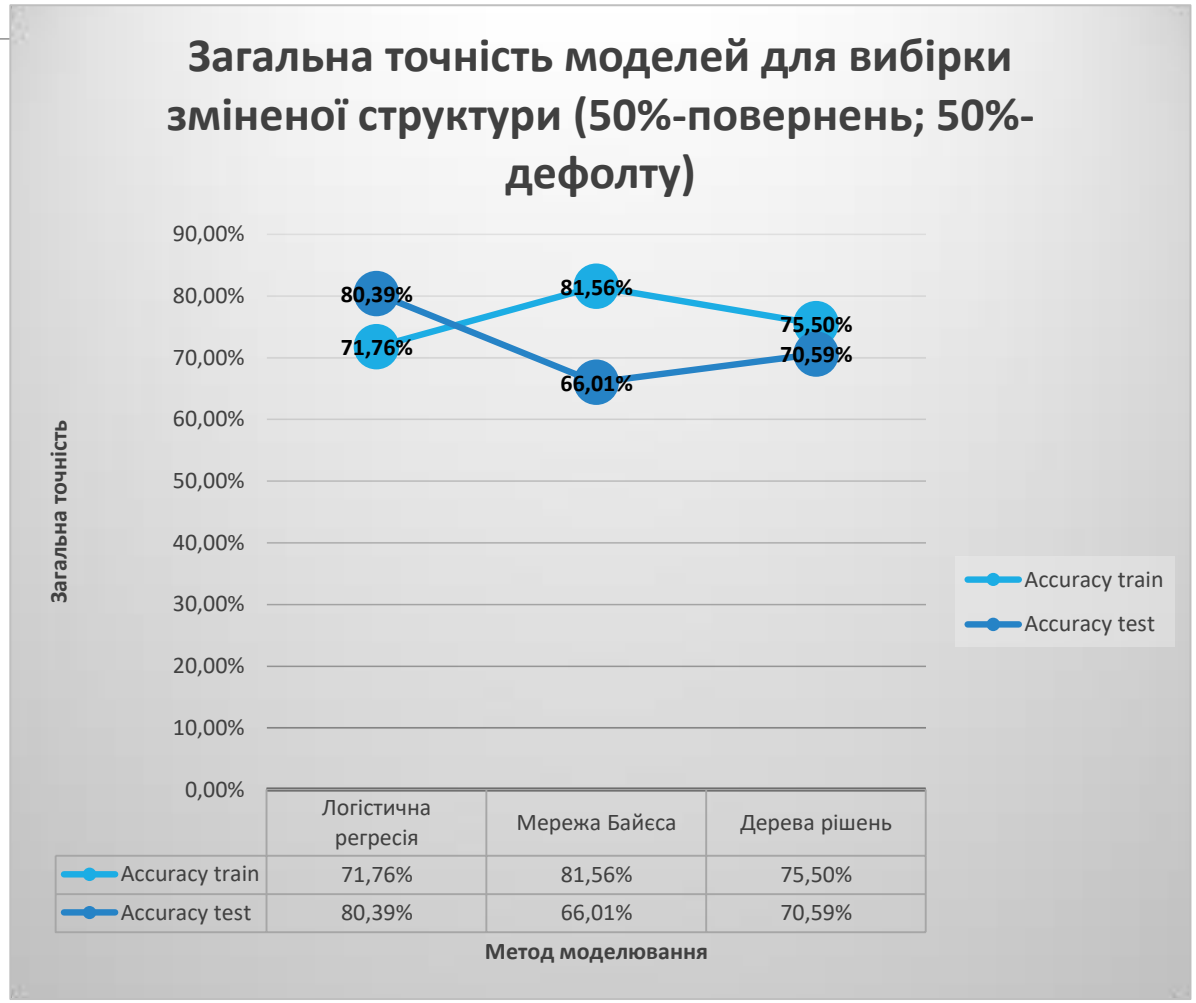




Порівняння індексу Джіні та загальної точності моделей з перехресним тестуванням (70%-навчальної; 30%-тестової)

Порівняння індексу Джіні та загальної точності моделей з перехресним тестуванням (70%-навчальної; 30%-тестової)





Порівняння отриманих результатів з перехресним тестуванням (70%-навчальної; 30%-тестової вибірки)

Назва методу	Індекс GINI		Значення AUC		Помилка 1-го роду		Помилка 2-го роду		Точність моделі		Якість моделі
	train	test	train	test	train	test	train	test	train	test	
Логістична регресія	0,599	0,528	0,799	0,764	0,06	0,06	0,16	0,18	77,59%	75%	Дуже висока
Мережа Байєса	0,747	0,43	0,873	0,715	0,07	0,12	0,1	0,14	81,86%	70,35%	Дуже висока
Дерева рішень	0.374	0.255	0.687	0.627	0,05	0,07	0,18	0,2	76.68%	72.67%	Середня

- для вибірки незміненої структури

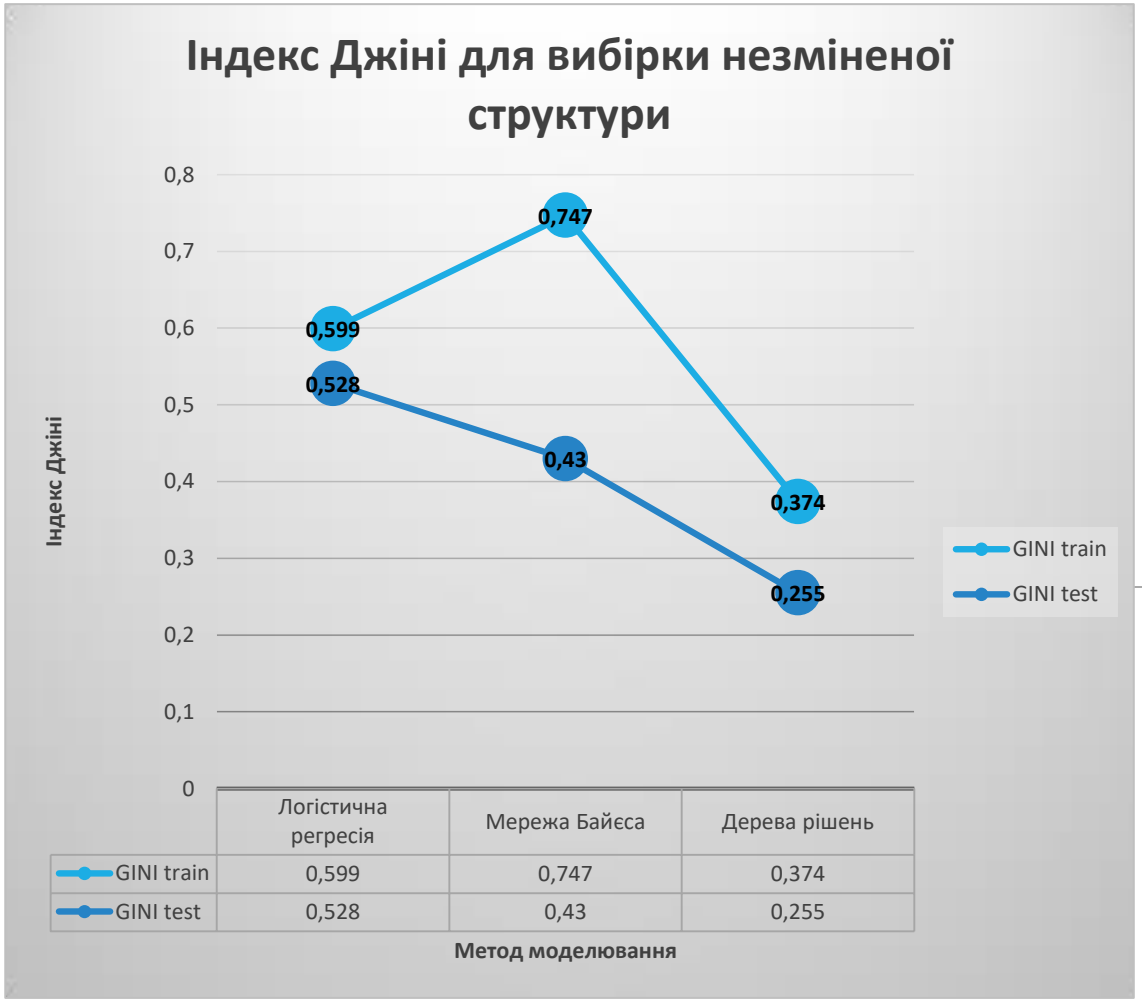
Назва методу	Індекс GINI		Значення AUC		Помилка 1-го роду		Помилка 2-го роду		Точність моделі		Якість моделі
	train	test	train	test	train	test	train	test	train	test	
Логістична регресія	0,808	0,648	0,904	0,824	0,07	0,09	0,01	0,05	91,46%	85,45%	Дуже висока
Мережа Байєса	0,924	0,048	0,962	0,524	0,02	0,07	0,02	-	94,38%	85,45%	Дуже висока
Дерева рішень	0	0	0,5	0,5	0,1	0,09	-	-	89,89%	90,91%	Незадов.

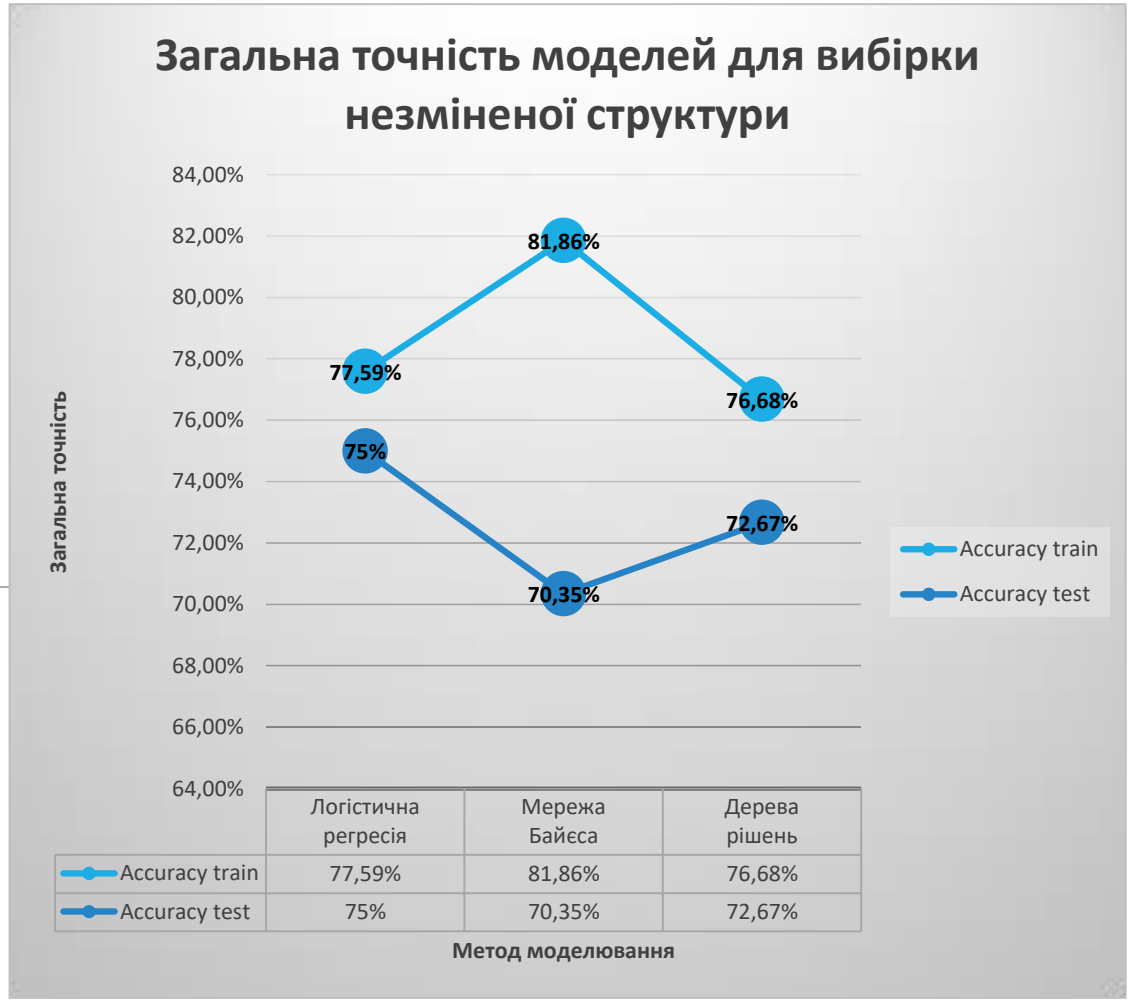
- для вибірки зміненої структури (90%-повернень; 10%-дефолту)

Назва методу	Індекс GINI		Значення AUC		Помилка 1-го роду		Помилка 2-го роду		Точність моделі		Якість моделі
	train	test	train	test	train	test	train	test	train	test	
Логістична регресія	0,644	0,722	0,822	0,861	0,12	0,14	0,12	0,07	75,3%	79,07%	Дуже висока
Мережа Байєса	0,803	0,671	0,901	0,836	0,07	0,09	0,09	0,09	83,81%	60,47%	Дуже висока
Дерева рішень	0,77	0,417	0,885	0,708	0,11	0,2	0,04	0,04	84,25%	74,42%	Дуже висока

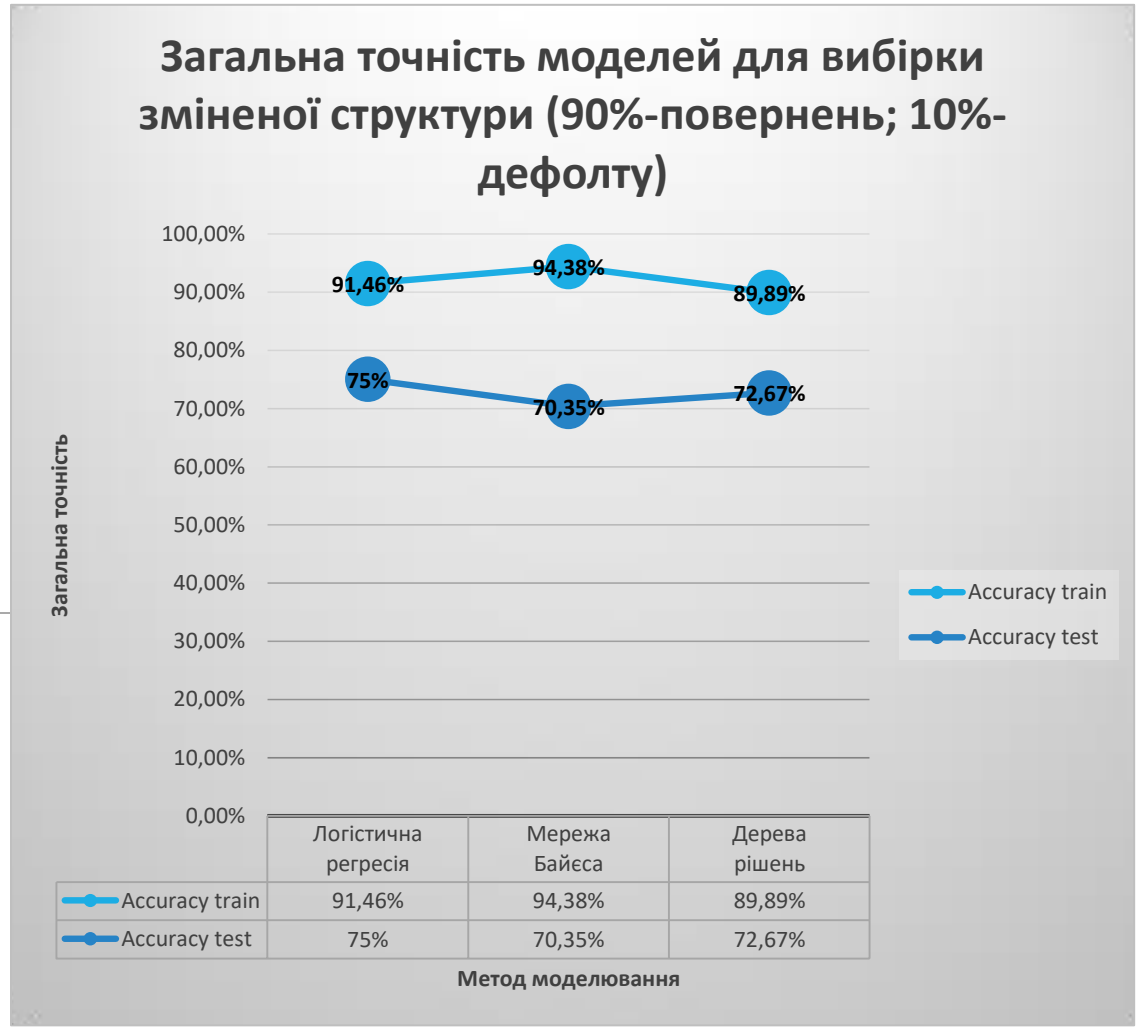
- для вибірки зміненої структури (50%-повернень; 50%-дефолту)

Порівняння індексу Джіні та загальної точності моделей з перехресним тестуванням (90%-навчальної; 10%-тестової)

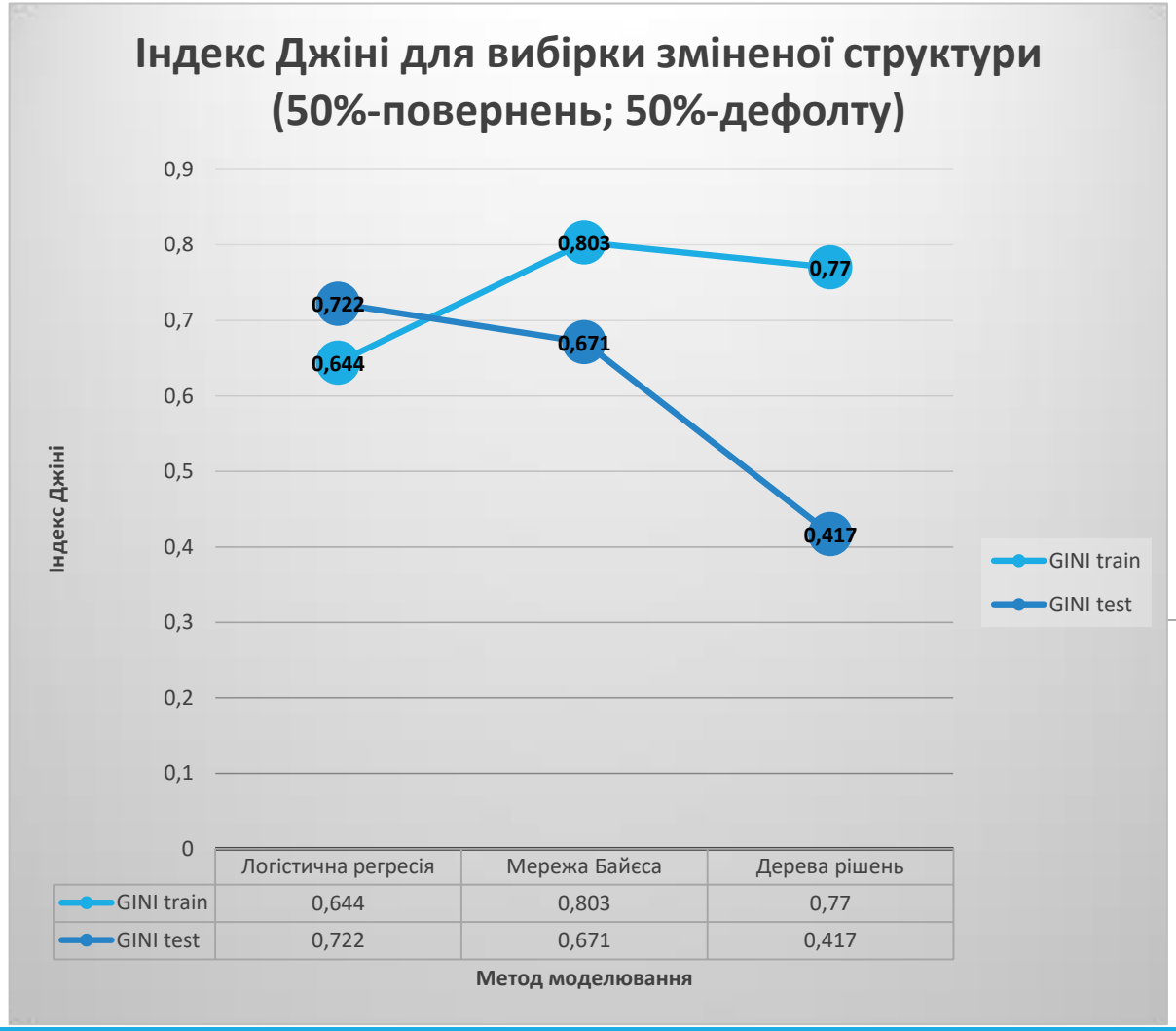


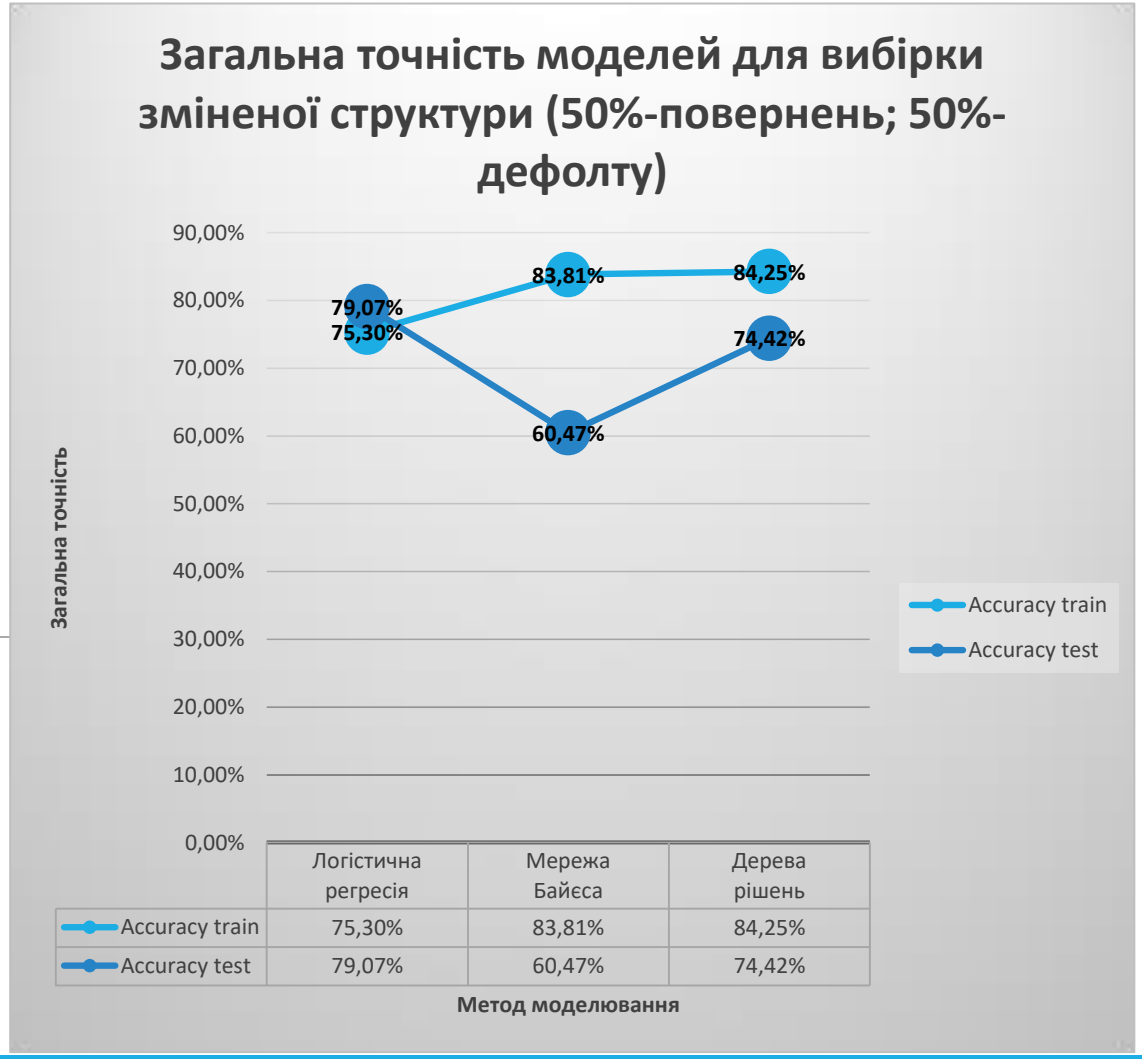


Порівняння індексу Джіні та загальної точності моделей з перехресним тестуванням (90%-навчальної; 10%-тестової)



Порівняння індексу Джіні та загальної точності моделей з перехресним тестуванням (90%-навчальної; 10%-тестової)





Аналіз отриманих результатів

Отримані результати свідчать, що найкращими моделями оцінки кредитоспроможності позичальників є ті, що побудовані за методами логістичної регресії та мереж Байєса. Також бачимо що кращі та більш точніші за використанням навчальної та тестової вибірки результати дає вибірка зі структурою 50%-повернень та 50%- дефолту. Ці результати ще раз підтверджують доцільність використання логістичної регресії та дерев рішень при оцінюванні кредитоспроможності позичальника.

Наукова новизна отриманих результатів:

- Удосконалено системну методологію побудови моделей оцінювання кредитоспроможності клієнтів фінансових установ шляхом корегування структури вибірки даних.
- Комплексне застосування кількох методів інтелектуального аналізу даних, яке дало можливість підвищити якість оцінювання (загальну точність) платоспроможності у середньому на 3-4% та індекс Джіні на 9,6%.
- Побудовані нові моделі для оцінювання кредитоспроможності у формі логістичної регресії, дерев рішень та мереж Байєса.

Публікації.

За матеріалами магістерського дослідження опубліковано 2 наукові праці:

- 1 стаття у студентському збірнику ІПСА «Системні науки і кібернетика»
- 1 стаття у вигляді тез у конференції SAIT 2017.

Дякую за увагу!