

ПРОГНОЗУВАННЯ КРЕДИТОСПРОМОЖНОСТІ ФІЗИЧНИХ ОСІБ ЗА МЕТОДАМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Виконала:
Студентка гр. КА-43м
Кириченко Вікторія
Науковий керівник:
д.т.н., професор Бідюк П. І.

Об'єкт дослідження: статистичні дані основних характеристик фізичних осіб-позичальників

Предмет: математичні моделі і методи інтелектуального аналізу даних (дерева рішень, байєсівські мережі, логістичні регресії)

Мета: побудувати класифікаційні моделі ймовірнісного типу щодо визначення кредитоспроможності фізичних осіб

Актуальність обраної теми

Сучасні тенденції розвитку банківського сектору – аналіз і управління кредитним ризиком з метою зниження його рівня.

Кредитні портфелі українських банків характеризуються високим рівнем простроченої, пролонгованої і безнадійної заборгованостей, надмірною концентрацією кредитного ризику.

У зв'язку з цим зростає актуальність питання дослідження сучасних методик та вдосконалення діючої методики оцінювання кредитоспроможності клієнтів банку.

Постановка задачі дослідження

- Провести огляд методів інтелектуального аналізу даних, що можуть бути використані для прогнозування кредитоспроможності фізичних осіб;
- Реалізувати комп'ютерну систему для визначення кредитоспроможності фізичних осіб за допомогою байєсівських мереж;
- Виконати апробацію запропонованих методів у вигляді комп'ютерної системи на реальних статистичних даних
- Порівняти ефективності описаних методів та мереж Байєса за допомогою аналітичної системи SAS Enterprise Miner;
- Проаналізувати результати моделювання кредитоспроможності фізичних осіб;

Кредитування фізичних осіб



Кредитоспроможність – наявність у позичальника банку передумов для отримання кредиту та його здатність повернути кредит і відсотки за ним у повному обсязі та обумовлені договором строки.

Банками висуваються особливі вимоги до позичальника – фізичної особи, що виявляється у пакеті документів для одержання кредиту та видів його забезпечення



Види кредитування



Проблема прогнозування кредитоспроможності фізичних осіб

- відсутність єдиного обґрунтованого визначення поняття кредитоспроможності
- відсутність єдиного обґрунтованого методологічного підходу
- відсутність єдиної методики оцінки кредитоспроможності позичальників.

Відомі розробники скорингових систем для оцінки кредитоспроможності

- Скоринг – це математична модель у вигляді зваженої суми певних характеристик, яка на основі кредитної історії попередніх клієнтів банку намагається з'ясувати ймовірність того, що конкретний позичальник поверне кредит у визначений термін

- «Fair Isaac Corporations»



- «Scorto Solutions»

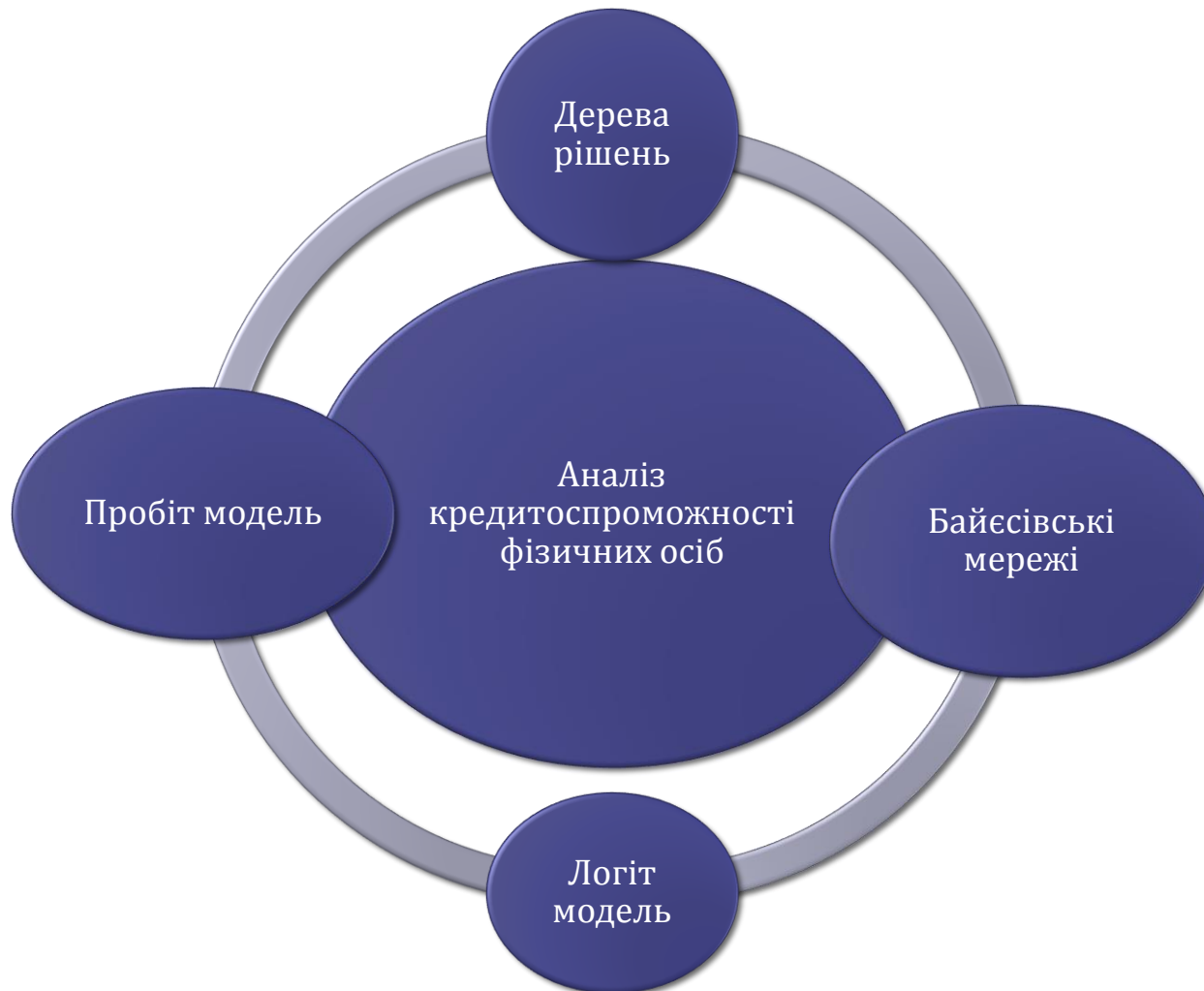
- «Creditinfo Group»



- «SAS»

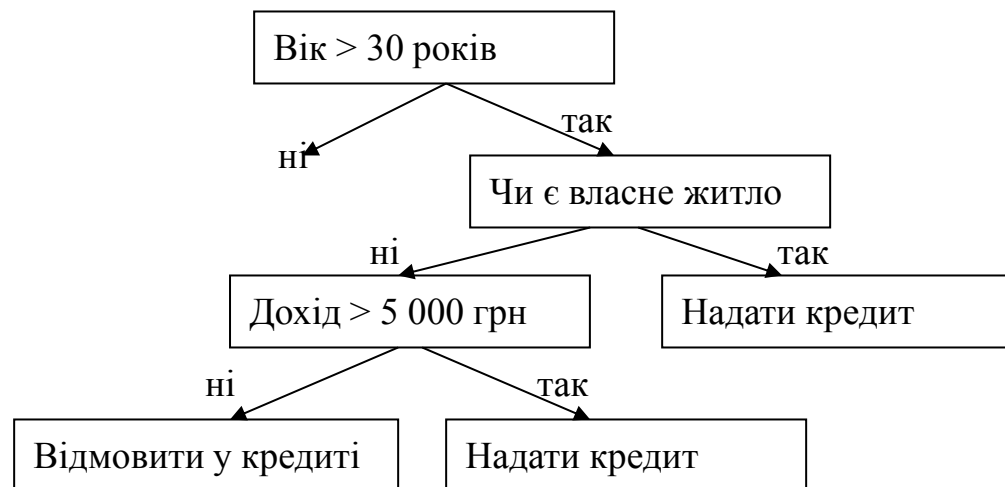


Обрані методи інтелектуального аналізу даних для прогнозування кредитоспроможності



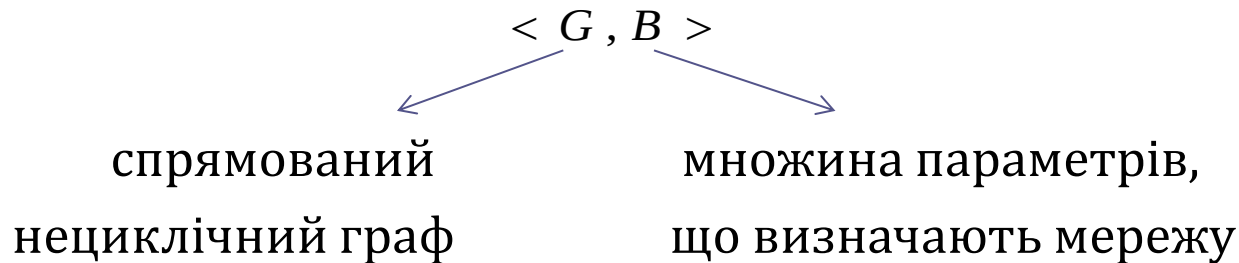
Дерева рішень

Дерево рішень - це графічне зображення послідовності рішень і станів середовища з зазначенням відповідних ймовірностей і виграшів для будь-яких комбінацій альтернатив і станів середовища.



Байєсівські мережі

Байєсівська мережа представляє собою пару $\langle G, V \rangle$



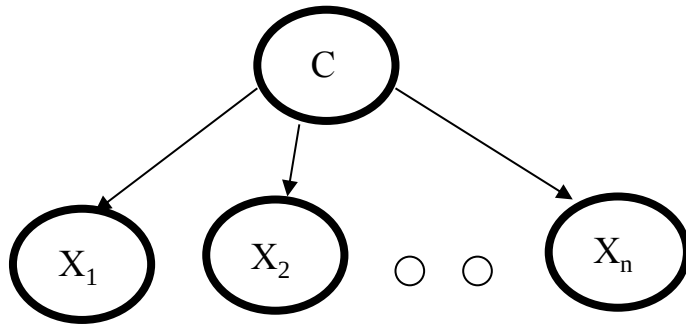
Спільний розподіл випадкових величин зводиться до рівняння:

$$P(X_1, X_2, \dots, X_n) = P(C) \cdot \prod_{i=1}^n P(X_i | C)$$

Структури мережі:

- Наївна мережа Байєса
- Доповнена деревом мережа
- Доповнена наївна мережа

Наївна мережа Байєса



Наївна мережа – особлива форма ймовірнісної моделі у формі байєсівської мережі, яка характеризується тим, що має сильні припущення стосовно незалежності змінних.

Де C являє собою цільову змінну;

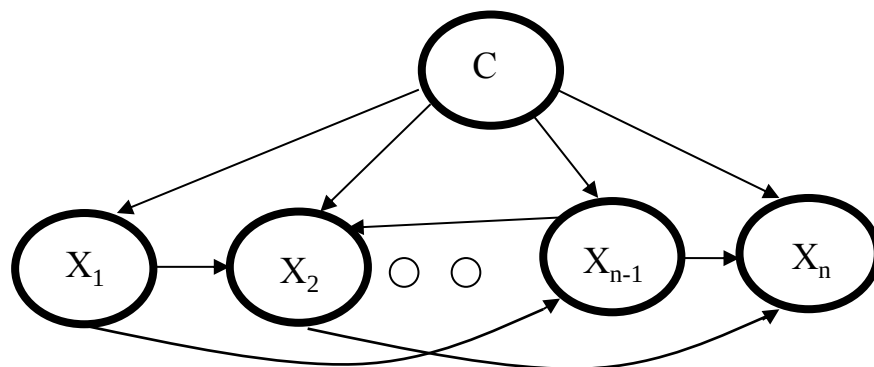
$X_1, \dots, X_n$ є вхідними змінними процесу, що можуть впливати на цільову.

З теореми Байєса маємо:

$$P(C \mid X_1, X_2, \dots, X_n) = \frac{P(X_1, X_2, \dots, X_n \mid C) \cdot P(C)}{P(X_1, X_2, \dots, X_n)}$$

Доповнена наївна мережа

Мережа, що підтримує структуру наївного байєсівського класифікатора і доповнює її шляхом додавання ребер між вершинами, що представляють змінні процесу, з метою відображення кореляції між параметрами.

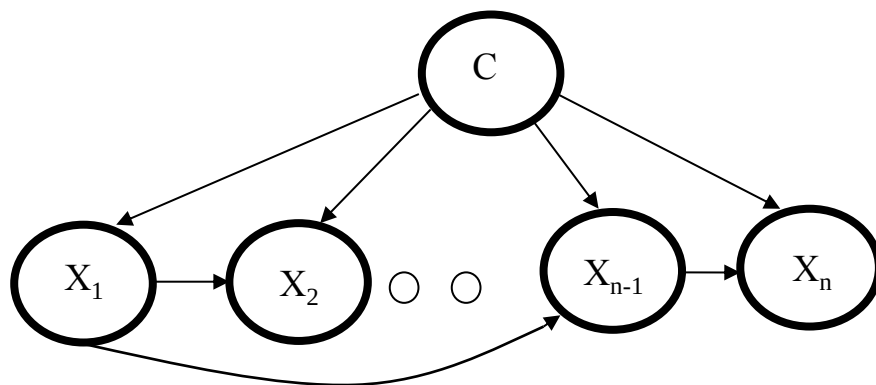


- Взаємна інформація між змінними:

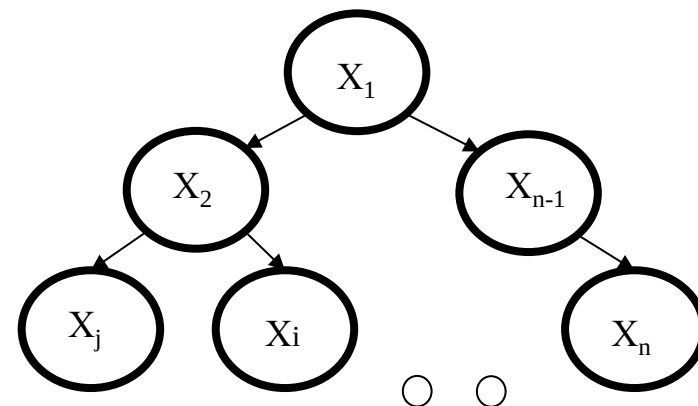
$$MI(X_i, X_j|C) = \sum_{x_1, x_2} P(x_1, x_2, c) \cdot \log \left(\frac{P(x_1, x_2|c)}{p(x_1|c)p(x_2|c)} \right)$$

Доповнена деревом мережа

Доповнена деревом мережа – це доповнена наївна мережа, що має обмеження на кількість батьківських змінних, не більше двох.



Доповнена деревом мережа



Деревоподібна мережа
вхідних змінних

Функція опису мінімальною довжиною

Значення функції ОМД довільної структури g при заданій послідовності з n спостережень $x^n = d_1 d_2 \dots d_n$ обчислюється за формулою:

$$L(g, x^n) = H(g, x^n) + \frac{k(g)}{2} \cdot \log(n)$$

де $H(g, x^n)$ - це значення емпіричної ентропії, а $k(g)$ – кількість незалежних умовних ймовірностей в структурі g .

$$H(g, x^n) = \sum_{j \in J} H(j, g, x^n) \quad k(g) = \sum_{j \in J} k(j, g)$$

ОМД j -ї вершини обчислюється за формулою:

$$L(j, g, x^n) = H(j, g, x^n) + \frac{k(j, g)}{2} \cdot \log(n)$$

а кількість незалежних умовних ймовірностей j -ї вершини:

$$k(j, g) = (\alpha^j - 1) \cdot \prod_{k \in \phi(j)} \alpha^k$$

Евристичний алгоритм побудови БМ

Вхідні дані: Множина навчальних даних $D = \{d_1, \dots, d_n\}$ $d_i = \{x_i^{(1)} x_i^{(2)} \dots x_i^{(N)}\}$

n – кількість спостережень; N – кількість вершин (змінних)

Перший етап. Для всіх пар вершин обчислюють значення взаємної інформації

$$Set_MI = \{MI(x^{s_1}, x^{s_2}), MI(x^{s_3}, x^{s_4}), MI(x^{s_5}, x^{s_6}), \dots\}$$

Другий етап.

Крок 1. На основі 1-го значення Set_MI будується множина моделей G :

$$\{(s_1 \rightarrow s_2); (s_1 \leftarrow s_2); (s_1 \text{ не залежить від } s_2)\}$$

Крок 2. Пошук серед моделей множини G . Оптимальна структура має найменше значення функції

$$1. g^* \leftarrow g_0 (\in G)$$

$$2. \text{ для } \forall g \in G - \{g_0\}, \text{ якщо } L(g, x^n) < L(g^*, x^n), \text{ то } g^* \leftarrow g_0$$

3. на вихід в якості рішення подається

Крок 3. Для наступного значення Set_MI будується множини моделей G :

$$\{(g^*; i_next \rightarrow j_next), (g^*; i_next \leftarrow j_next), (g^*; i_next \text{ не залежить від } j_next)\}$$

Перехід на крок 2.

Умова завершення алгоритму. Перегляд частини або всіх елементів Set_MI

Критерії якості моделі

- Загальна точність моделі:

$$CA = \frac{\text{кількість вірно спрогнозованих значень}}{\text{загальна кількість прогнозованих значень}}$$

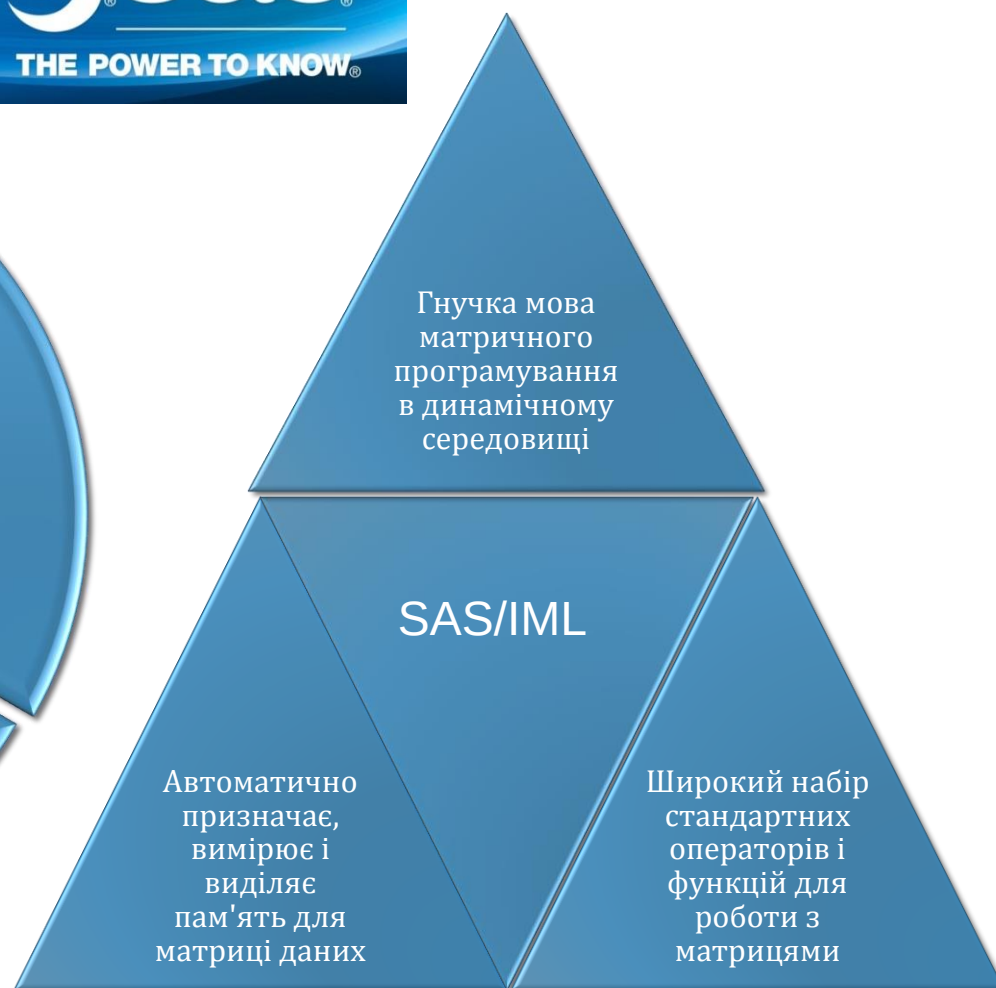
- Помилки I-го та II-го роду

Прогноз моделі:	Особа повертає кредит(0)	Особа не повертає кредит (1)
Фактично: Особа повертає кредит (0)	Вірно класифіковані (TP)	Помилки II-го роду (FN)
Фактично: Особа не повертає кредит (1)	Помилки I-го роду (FP)	Вірно класифіковані (TN)

- ROC-крива показує залежність кількості вірно класифікованих позитивних прикладів від кількості невірно класифікованих негативних прикладів.
- Чисельний показник площі під кривою AUC (Area Under Curve) і обчислюється, наприклад, за методом трапецій:

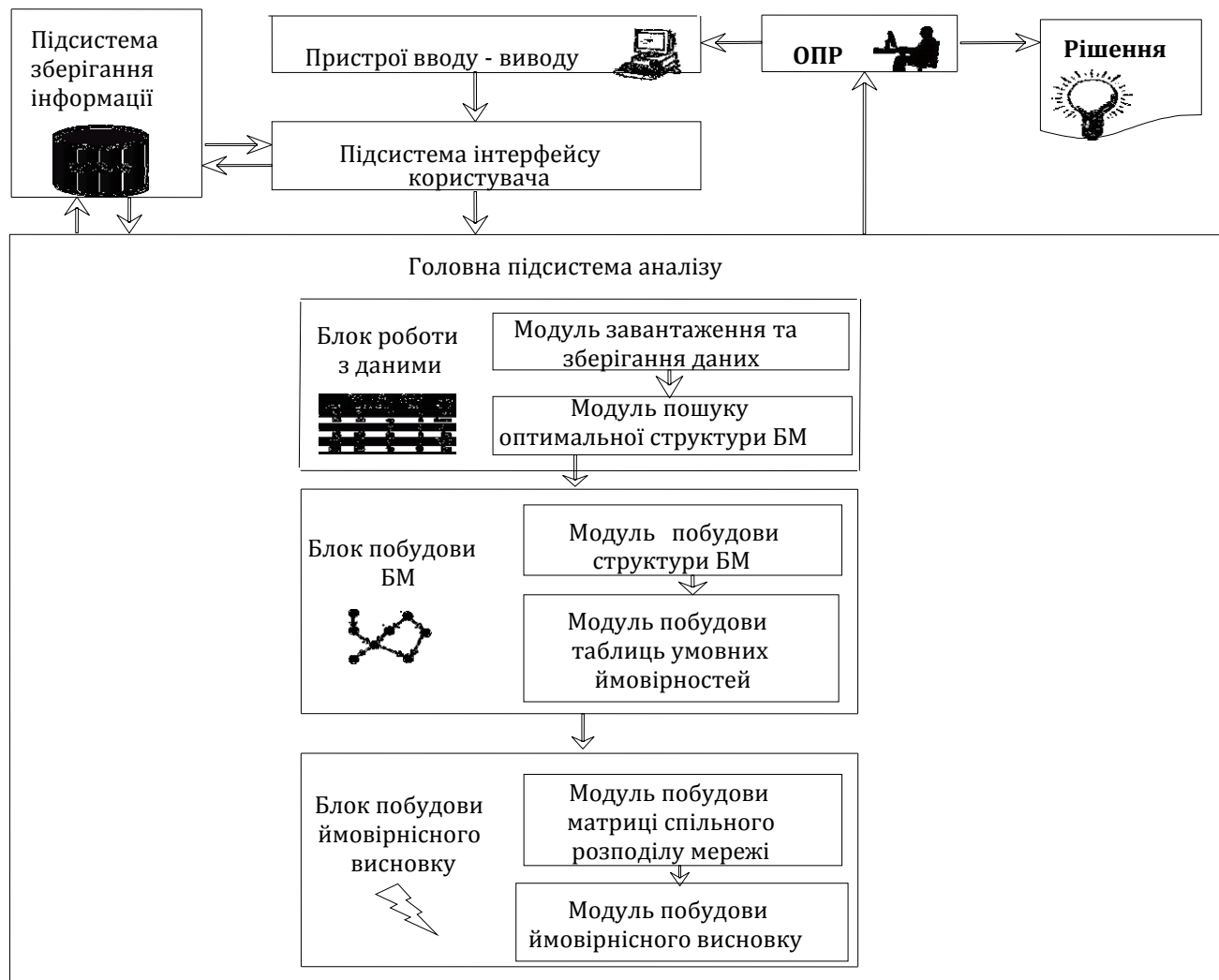
$$AUC = \int f(x) dx = \sum_i \left[\frac{X_{i+1} + X_i}{2} \right] \cdot (Y_{i+1} - Y_i)$$

Інтервал AUC	Індекс GINI	Якість моделі
0,9-1,0	0,8 – 1,0	Відмінна
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

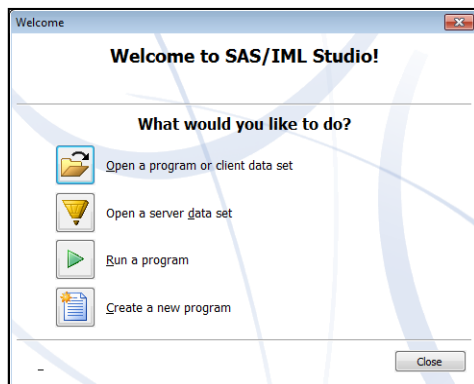


Компанія SAS

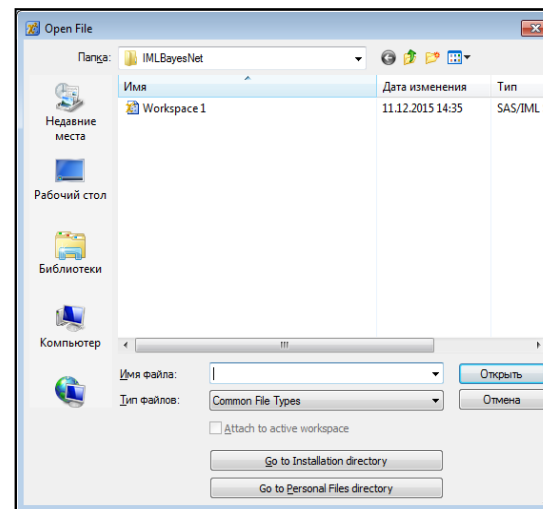
Архітектура комп'ютерної програми "IMLBayesNet"



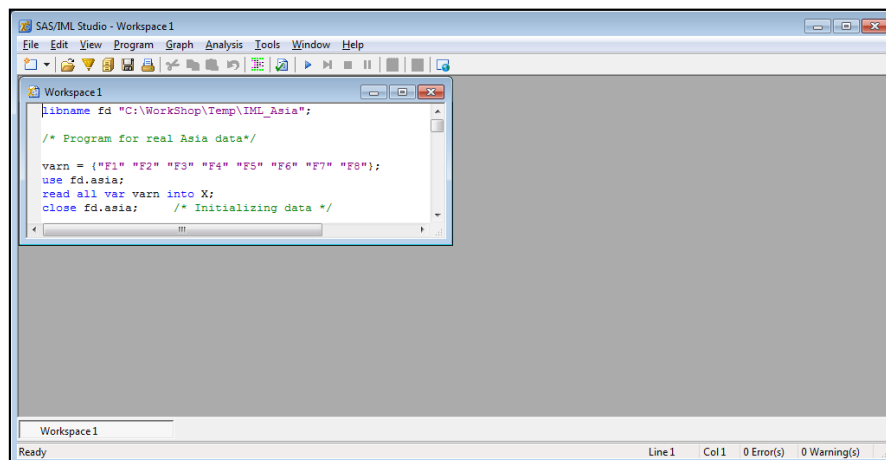
Загальний вигляд програми



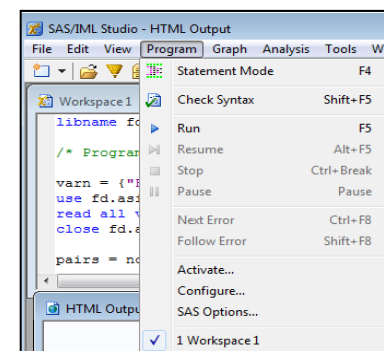
Вигляд вікна Welcome



Вигляд вікна Open File



Вигляд вікна SAS/IML Studio



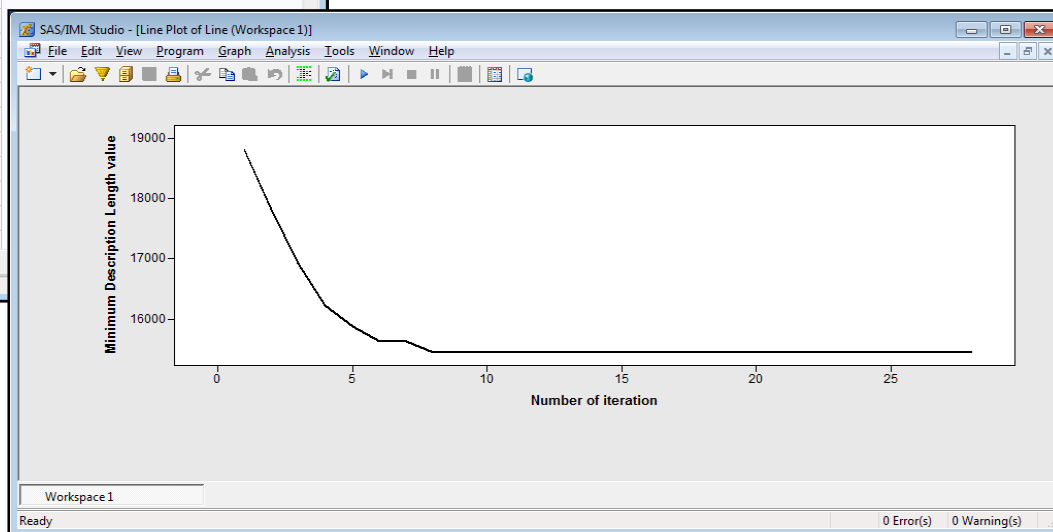
Запуск програми

Результати роботи

SAS/IML Studio - [HTML Output]

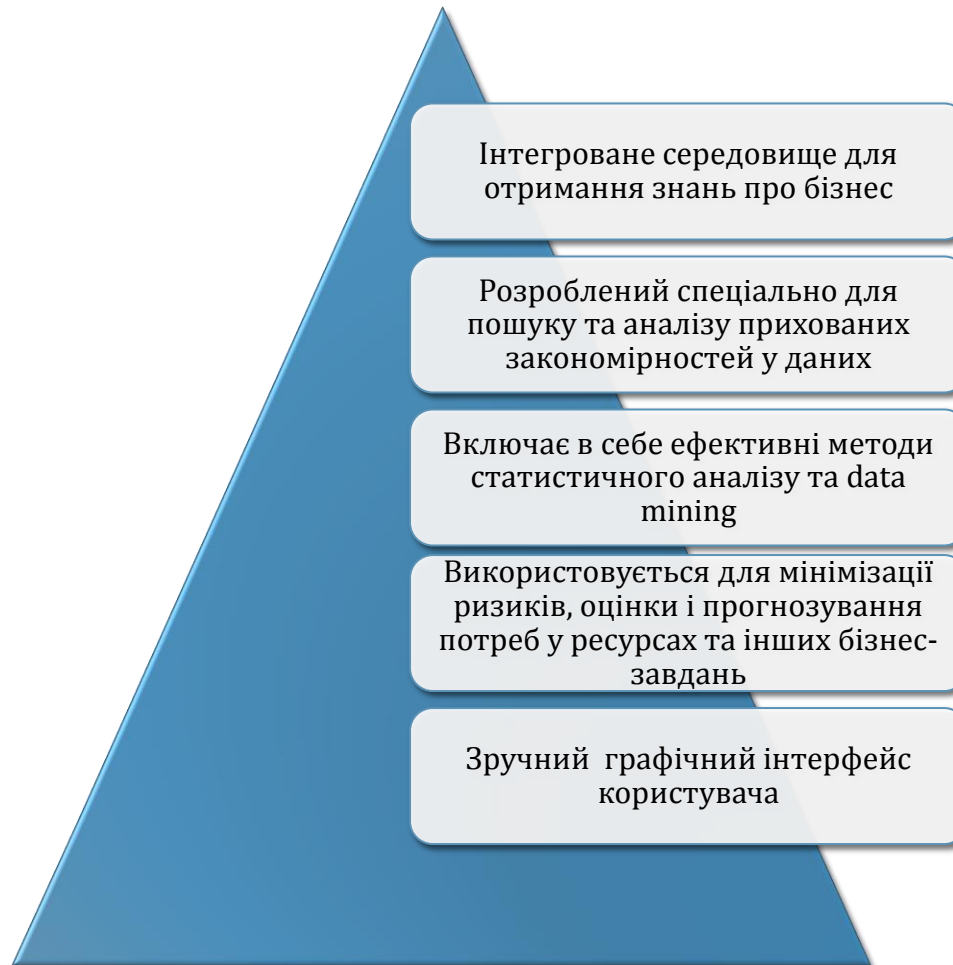
PXIXj					
Xi	Xj	Xi Xj=0 0	Xi Xj=0 1	Xi Xj=1 0	Xi Xj=1 1
1.0000	2.0000	0.5231	0.0132	0.4384	0.0253
1.0000	3.0000	0.5299	0.0064	0.4584	0.0053
1.0000	4.0000	0.5169	0.0194	0.4333	0.0304
1.0000	5.0000	0.5297	0.0067	0.4598	0.0039
1.0000	6.0000	0.0461	0.4902	0.0513	0.4124
1.0000	7.0000	0.3807	0.1556	0.1860	0.2777
1.0000	8.0000	0.3659	0.1704	0.2149	0.2488
2.0000	3.0000	0.9501	0.0114	0.0382	0.0003
2.0000	4.0000	0.9501	0.0114	0.0000	0.0385
2.0000	5.0000	0.9511	0.0104	0.0383	0.0001
2.0000	6.0000	0.0590	0.9025	0.0383	0.0001
2.0000	7.0000	0.5480	0.4135	0.0188	0.0197
2.0000	8.0000	0.5719	0.3896	0.0089	0.0296
3.0000	4.0000	0.9501	0.0382	0.0000	0.0117
3.0000	5.0000	0.9787	0.0096	0.0107	0.0010

Workspace 1
Ready



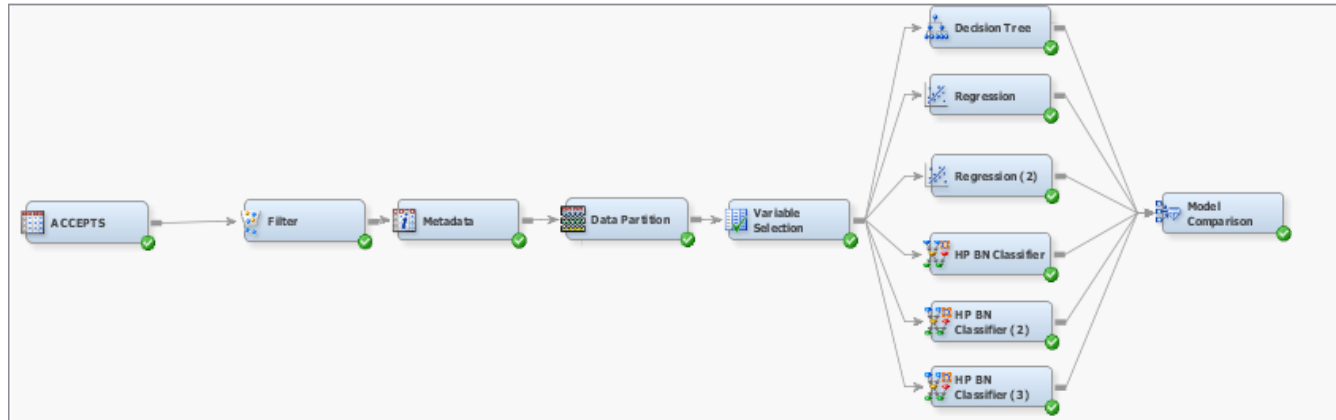
Побудова мережі для набору даних, що містить 5837 записів та 20 змінних, відбулася за 3 секунди

SAS Enterprise Miner



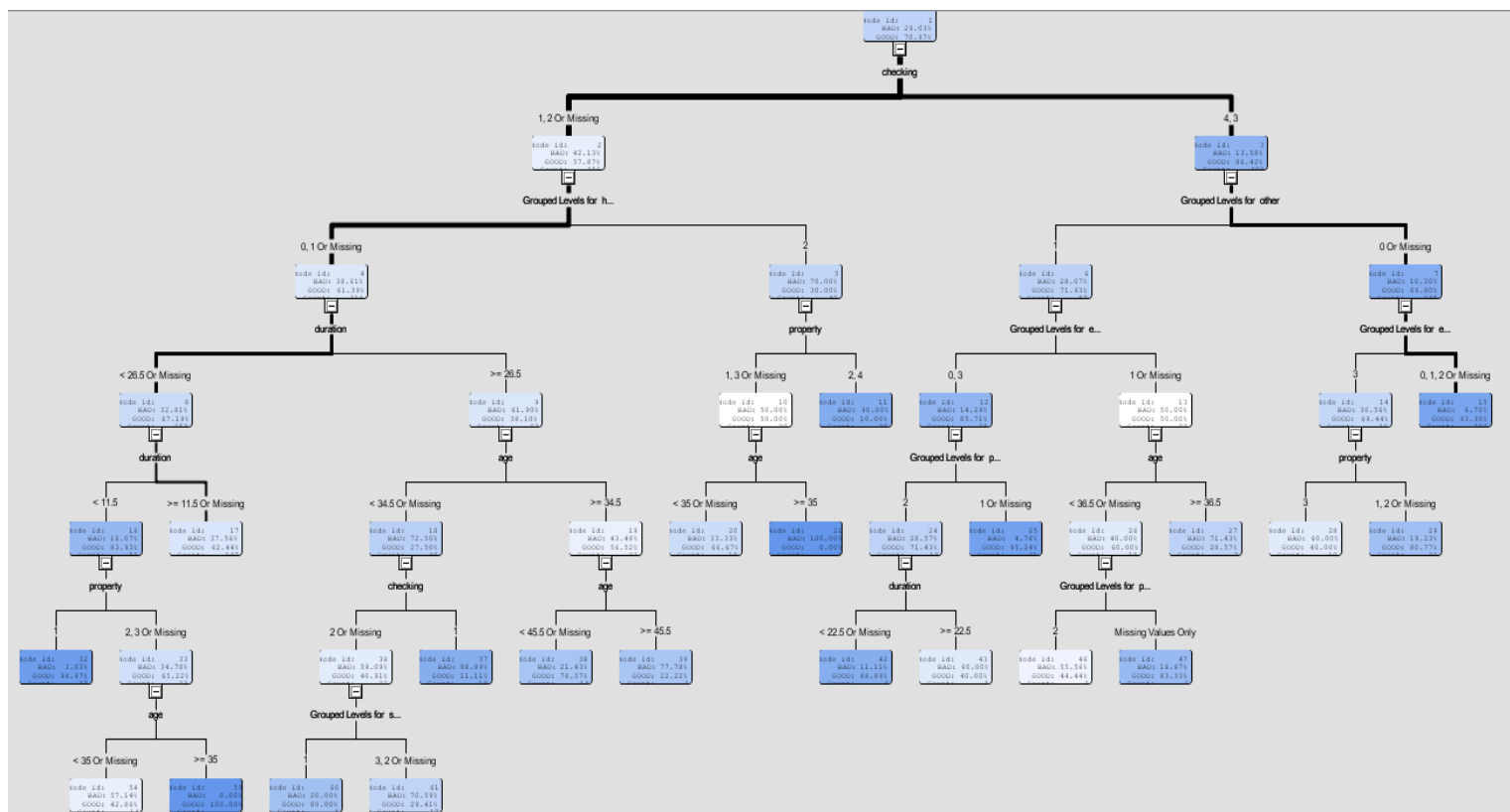
Моделювання кредитоспроможності

Використаний набір даних DMAGECR містить реальні статистичні дані за 2007 рік щодо кредитування 1000 фізичних осіб одного з найбільших банків Німеччини. Даний процес описується 21 змінною (з яких 20 вхідних і одна цільова змінна).

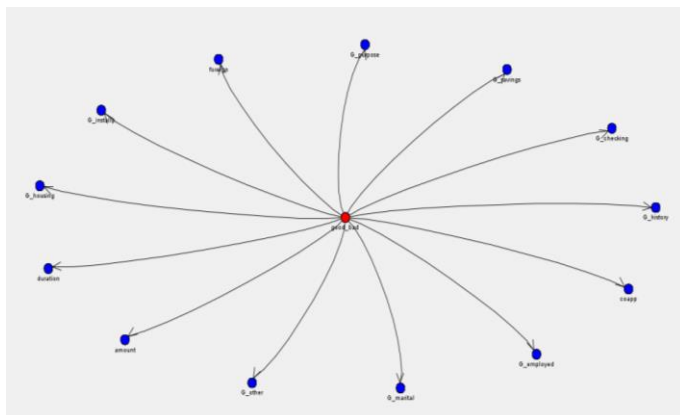


Діаграма технологічних процесів

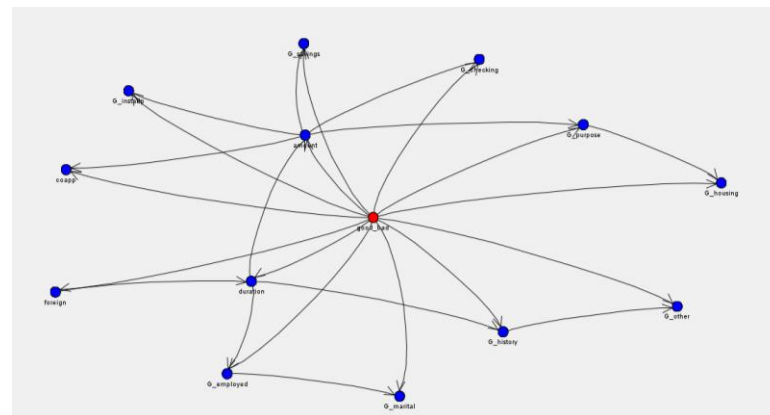
		25
Назва змінної	Опис	
Y	цільова змінна (приймає 0 якщо клієнт повернув кредит, 1-не повернув кредит)	
X ₁	стан чекового рахунку клієнта	
X ₂	тривалість співпраці з клієнтом, місяців	
X ₃	кредитна історія	
X ₄	мета кредиту	
X ₅	сума кредиту	
X ₆	ощадний рахунок	
X ₇	стаж роботи на цьому місці роботи	
X ₈	щомісячні платежі по кредиту по відношенню до місячного доходу клієнта	
X ₉	сімейний стан клієнта	
X ₁₀	наявність поручителів	
X ₁₁	дата, з якої резидент проживає за вказаною адресою	
X ₁₂	наявність власного майна	
X ₁₃	вік	
X ₁₄	наявність інших розстрочок	
X ₁₅	умови проживання	
X ₁₆	кількість непогашених кредитів в даному банку	
X ₁₇	професія	
X ₁₈	кількість утриманців	
X ₁₉	наявність телефону	
X ₂₀	іноземець	



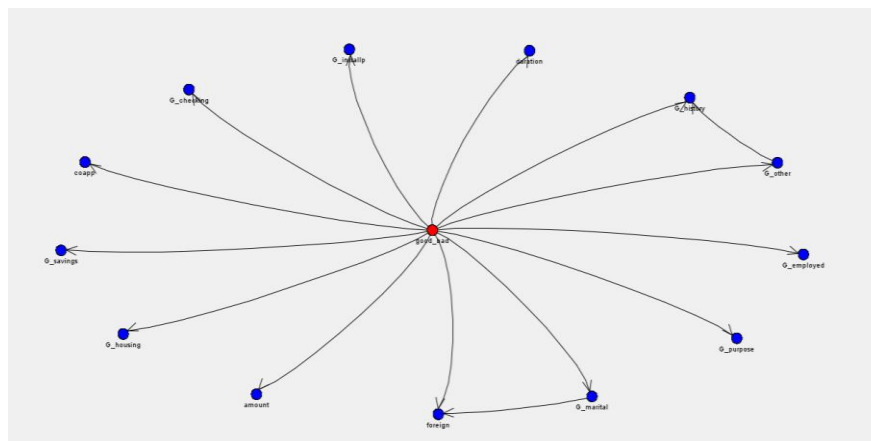
Побудовані мережі Байєса



- Наївна мережа

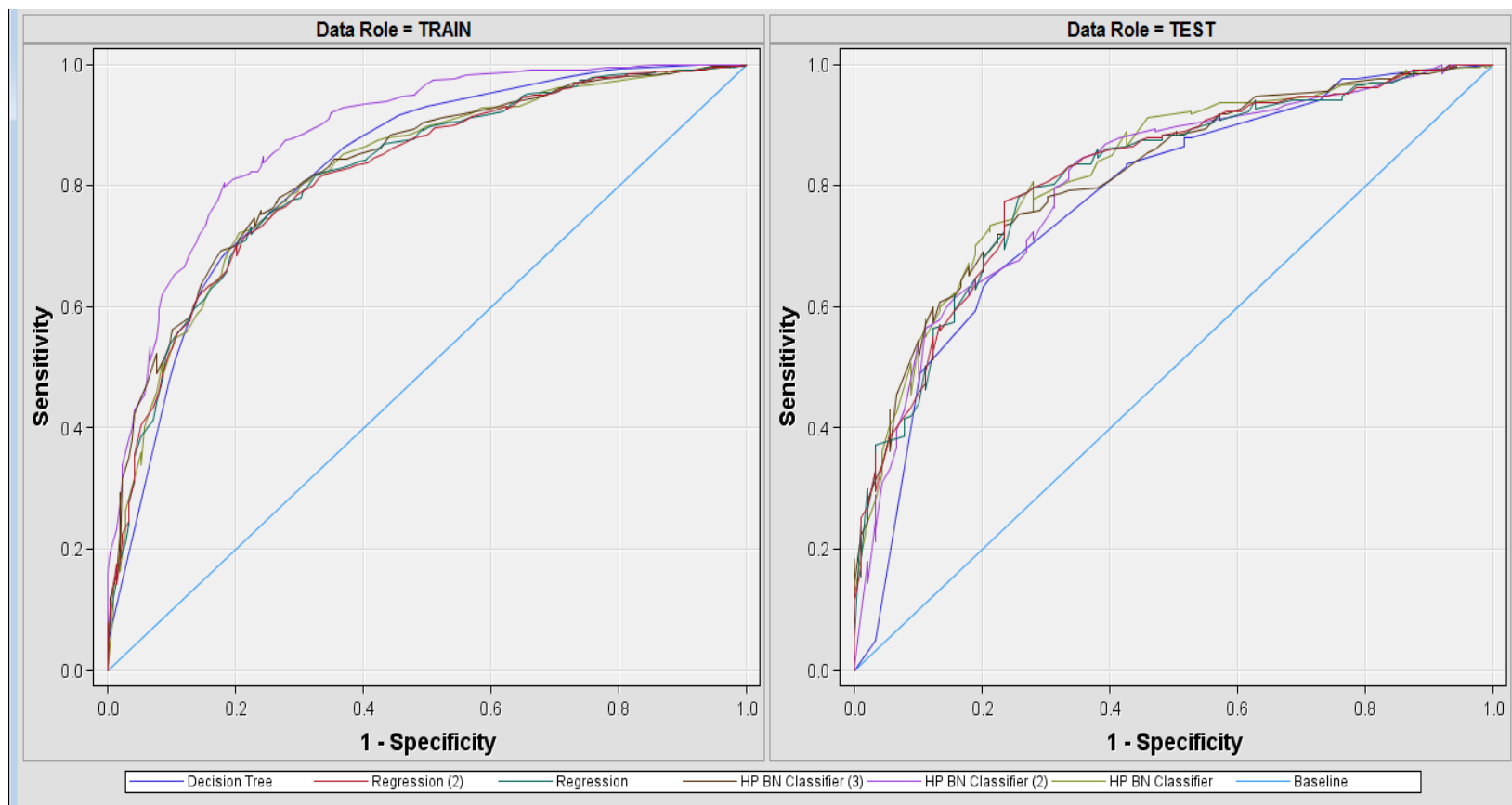


- Доповнена деревом мережа



- Доповнена наївна мережа

Рос криві для навчальної та тестової вибірки



Значення статистичних характеристик

Назва моделі	Загальна точність %	Помилка I роду %	Помилка II роду %	Помилка I+II роду %	ROC індекс	Коефіцієнт GINI
Наївна мережа	74.78	6.53	18.69	25.23	0,78	0,56
Доповнена деревом мережа	82.07	4.71	13.22	17.93	0,783	0,566
Доповнена наївна мережа	75.68	6.23	18.09	24.32	0,796	0,593
Дерево рішень	79.48	16.11	4.41	20.52	0,714	0,429
Логіт	77.51	14.74	7.75	22.49	0,79	0,511
Пробіт	76.9	15.05	8.05	23.1	0,792	0,515

ВИСНОВКИ

- Проаналізовано методи інтелектуального аналізу, що можуть бути використані для прогнозування кредитоспроможності фізичних осіб у сучасній банківській системі;
- Розроблена програма IMLBayesNet для побудови БМ за допомогою евристичного алгоритму на мові SAS/IML
- Порівняно ефективності описаних методів за допомогою аналітичної системи SAS Enterprise Miner
- Проаналізовано результати моделювання кредитоспроможності фізичних осіб з використанням двох наборів реальних статистичних даних

Перспективи щодо подальших досліджень

- Удосконалення вибраних типів класифікаційних моделей з метою підвищення якості класифікації клієнтів на дві групи
- Застосування інших типів моделей для прогнозування кредитоспроможності фізичних осіб. Наприклад, нейронні мережі та нечітку логіку
- Застосування текстової аналітики для розробки індикаторів, що включаються в математичну модель надання кредиту споживачу

Результати впровадження, публікації

- «Застосування наївного та дереводоповненого байєсівських класифікаторів для прогнозування кредитоспроможності фізичних осіб» - тези конференції SAIT-2016;
- «Моделювання фінансових ризиків з використанням наївного та доповненого деревом класифікаторів у формі байєсівських мереж» - стаття у виданні «Наукові вісті НТУУ «КПІ»», 2016р.;
- «Text Mining Analysis of Agriculture Internet Sources Using SAS Software» - тези конференції ISDMCI-2016;
- *Свідоцтво про реєстрацію авторського права на твір: Комп'ютерна програма «IMLBayesNet».*
- *Акт впровадження результатів магістерської дисертації у компанії SAS Institute ("SAS");*

Дякую за увагу!