

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ
ΕΡΓΑΣΤΗΡΙΟ 6

WEKA: CLASS IMBALANCE

(HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA)

1. Αποθηκεύστε σε έναν φάκελο το αρχείο *bankruptcy.csv*, και ανοίξτε το με το Notepad, για να δείτε τι περιέχει.

Το αρχείο περιέχει δεδομένα 6819 εταιριών στην Taiwan, που πτώχευσαν ή όχι. Τα δεδομένα συλλέχθηκαν από το Taiwan Economic Journal για τα έτη 1999 έως 2009. Η πτώχευση μιας εταιρείας ορίστηκε με βάση τους επιχειρηματικούς κανονισμούς του Χρηματιστηρίου της Ταϊβάν. Κάθε εταιρία αναπαρίσταται μέσω 96 οικονομικών δεικτών.

Τα δεδομένα έχουν ληφθεί από το UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

Σχετικό Άρθρο: Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572.

Επεξήγηση των μεταβλητών:

Y - Bankrupt?: Class label
X1 - ROA(C) before interest and depreciation before interest: Return On Total Assets(C)
X2 - ROA(A) before interest and % after tax: Return On Total Assets(A)
X3 - ROA(B) before interest and depreciation after tax: Return On Total Assets(B)
X4 - Operating Gross Margin: Gross Profit/Net Sales
X5 - Realized Sales Gross Margin: Realized Gross Profit/Net Sales
X6 - Operating Profit Rate: Operating Income/Net Sales
X7 - Pre-tax net Interest Rate: Pre-Tax Income/Net Sales
X8 - After-tax net Interest Rate: Net Income/Net Sales
X9 - Non-industry income and expenditure/revenue: Net Non-operating Income Ratio
X10 - Continuous interest rate (after tax): Net Income-Exclude Disposal Gain or Loss/Net Sales
X11 - Operating Expense Rate: Operating Expenses/Net Sales
X12 - Research and development expense rate: (Research and Development Expenses)/Net Sales
X13 - Cash flow rate: Cash Flow from Operating/Current Liabilities
X14 - Interest-bearing debt interest rate: Interest-bearing Debt/Equity
X15 - Tax rate (A): Effective Tax Rate
X16 - Net Value Per Share (B): Book Value Per Share(B)
X17 - Net Value Per Share (A): Book Value Per Share(A)
X18 - Net Value Per Share (C): Book Value Per Share(C)
X19 - Persistent EPS in the Last Four Seasons: EPS-Net Income
X20 - Cash Flow Per Share
X21 - Revenue Per Share (Yuan ¥): Sales Per Share
X22 - Operating Profit Per Share (Yuan ¥): Operating Income Per Share

X23 - Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share
X24 - Realized Sales Gross Profit Growth Rate
X25 - Operating Profit Growth Rate: Operating Income Growth
X26 - After-tax Net Profit Growth Rate: Net Income Growth
X27 - Regular Net Profit Growth Rate: Continuing Operating Income after Tax Growth
X28 - Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth
X29 - Total Asset Growth Rate: Total Asset Growth
X30 - Net Value Growth Rate: Total Equity Growth
X31 - Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
X32 - Cash Reinvestment %: Cash Reinvestment Ratio
X33 - Current Ratio
X34 - Quick Ratio: Acid Test
X35 - Interest Expense Ratio: Interest Expenses/Total Revenue
X36 - Total debt/Total net worth: Total Liability/Equity Ratio
X37 - Debt ratio %: Liability/Total Assets
X38 - Net worth/Assets: Equity/Total Assets
X39 - Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
X40 - Borrowing dependency: Cost of Interest-bearing Debt
X41 - Contingent liabilities/Net worth: Contingent Liability/Equity
X42 - Operating profit/Paid-in capital: Operating Income/Capital
X43 - Net profit before tax/Paid-in capital: Pretax Income/Capital
X44 - Inventory and accounts receivable/Net value: (Inventory+Accounts Receivables)/Equity
X45 - Total Asset Turnover
X46 - Accounts Receivable Turnover
X47 - Average Collection Days: Days Receivable Outstanding
X48 - Inventory Turnover Rate (times)
X49 - Fixed Assets Turnover Frequency
X50 - Net Worth Turnover Rate (times): Equity Turnover
X51 - Revenue per person: Sales Per Employee
X52 - Operating profit per person: Operation Income Per Employee
X53 - Allocation rate per person: Fixed Assets Per Employee
X54 - Working Capital to Total Assets
X55 - Quick Assets/Total Assets
X56 - Current Assets/Total Assets
X57 - Cash/Total Assets
X58 - Quick Assets/Current Liability
X59 - Cash/Current Liability
X60 - Current Liability to Assets
X61 - Operating Funds to Liability
X62 - Inventory/Working Capital
X63 - Inventory/Current Liability
X64 - Current Liabilities/Liability
X65 - Working Capital/Equity
X66 - Current Liabilities/Equity
X67 - Long-term Liability to Current Assets
X68 - Retained Earnings to Total Assets
X69 - Total income/Total expense
X70 - Total expense/Assets
X71 - Current Asset Turnover Rate: Current Assets to Sales
X72 - Quick Asset Turnover Rate: Quick Assets to Sales

X73 - Working capitcal Turnover Rate: Working Capital to Sales
X74 - Cash Turnover Rate: Cash to Sales
X75 - Cash Flow to Sales
X76 - Fixed Assets to Assets
X77 - Current Liability to Liability
X78 - Current Liability to Equity
X79 - Equity to Long-term Liability
X80 - Cash Flow to Total Assets
X81 - Cash Flow to Liability
X82 - CFO to Assets
X83 - Cash Flow to Equity
X84 - Current Liability to Current Assets
X85 - Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise
X86 - Net Income to Total Assets
X87 - Total assets to GNP price
X88 - No-credit Interval
X89 - Gross Profit to Sales
X90 - Net Income to Stockholder's Equity
X91 - Liability to Equity
X92 - Degree of Financial Leverage (DFL)
X93 - Interest Coverage Ratio (Interest expense to EBIT)
X94 - Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise
X95 - Equity to Liability

2.Από το ΕΝΑΡΞΗ επιλέξτε Weka 3.6.6 -> Weka 3.6. Από το παράθυρο του Weka επιλέξτε *Explorer*.

3. Από το μενού επιλέξτε *Preprocess -> Open File*. Ανοίξτε το αρχείο *bankruptcy.csv*.

4. Επιλέξτε σαν έξοδο την μεταβλητή bankrupt.

5. Παρατηρήστε την ανισοκατανομή των τιμών της εξόδου στα δεδομένα. Πόσα είναι τα θετικά και πόσα τα αρνητικά παραδείγματα;_____
_____

6.Τρέξτε τον αλγόριθμο επαγωγής δέντρου απόφασης στα δεδομένα σας. Τι παρατηρείτε σχετικά με την απόδοση του αλγορίθμου στις δυο τιμές της εξόδου; Είναι παραπλήσια η απόδοση ανάμεσα στις δυο τιμές;_____
_____

Τυχαία Υπο-δειγματοληψία

7. Εφαρμόστε το φίλτρο Resample (supervised) στα δεδομένα σας. Επιλέξτε
   a. τιμή 1 στην παράμετρο biasToUniformClass (δηλαδή ισοκατανομή στις τιμές τις εξόδου).
   b. τιμή TRUE στην παράμετρο noReplacement ώστε να μην γίνεται επανατοποθέτηση.
   c. τιμή 10 στο sampleSizePercent ώστε το τελικό σετ εκπαίδευσης να έχει μετά την υποδειγματοληψία το 10% του αριθμού παραδειγμάτων που είχε πριν.

Πατήστε Apply.

Τρέξτε πάλι τον αλγόριθμο επαγωγής δέντρου απόφασης στα δεδομένα σας. Τι παρατηρείτε τώρα σχετικά με την απόδοση του αλγορίθμου στις δυο τιμές της εξόδου;

_____

_____

Κάντε από το μενού Preprocess Undo, ώστε να αναιρεθεί το προηγούμενο φίλτρο.

8. Εφαρμόστε το φίλτρο SpreadSubsample (supervised) στα δεδομένα σας. Επιλέξτε τιμή 10 στην παράμετρο distributionSpread (δηλαδή κατανομή 10:1 στις τιμές τις εξόδου).
Πατήστε Apply.
Τρέξτε πάλι τον αλγόριθμο επαγωγής δέντρου απόφασης στα δεδομένα σας. Τι παρατηρείτε τώρα σχετικά με την απόδοση του αλγορίθμου στις δυο τιμές της εξόδου;

_____

_____

Κάντε από το μενού Preprocess Undo, ώστε να αναιρεθεί το προηγούμενο φίλτρο.

Synthetic Minority Oversampling (SMOTE)

9. Εφαρμόστε το φίλτρο SMOTE (supervised) στα δεδομένα σας. Επιλέξτε σωστά το index του χαρακτηριστικού της εξόδου. Επιλέξτε την τιμή του ποσοστού των συνθετικών παραδειγμάτων που θα δημιουργηθούν.
Πατήστε Apply.
Τρέξτε πάλι τον αλγόριθμο επαγωγής δέντρου απόφασης στα δεδομένα σας. Τι παρατηρείτε τώρα σχετικά με την απόδοση του αλγορίθμου στις δυο τιμές της εξόδου;

_____

_____