**Sprawozdanie**

**Wprowadzenie do sieci neuronowych i uczenia maszynowego**

**Drzewa decyzyjne – część 2**

Julia Samp 151775 Wojciech Kot 151879

Po przeanalizowaniu tematu zadania i listy atrybutów, wybraliśmy 8 najistotniejszych z nich. Wybraliśmy takie, które naszym zdaniem najbardziej motywują uczniów do nauki, które wspierają ich w tej motywacji i pomagają w nauce oraz które najbardziej je utrudniają.

* reason
* failures
* higher
* Medu
* Fedu
* schoolsup
* internet
* absences

Przeanalizowaliśmy również wszystkie metryki, dostępne w programie Weka. Wybraliśmy 3 z nich, które uznaliśmy za najistotniejsze dla analizy otrzymanych danych.

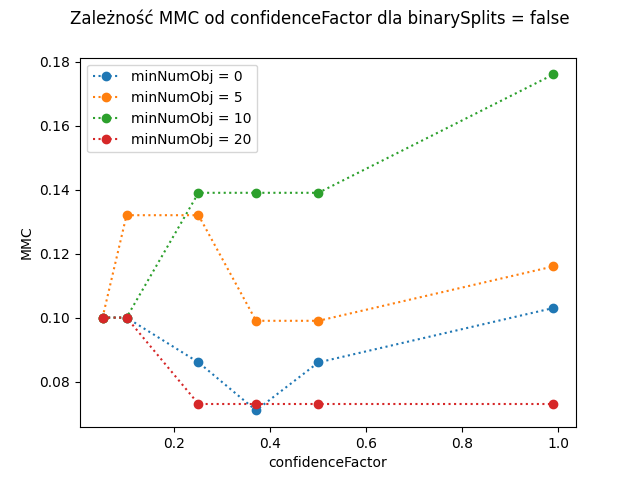
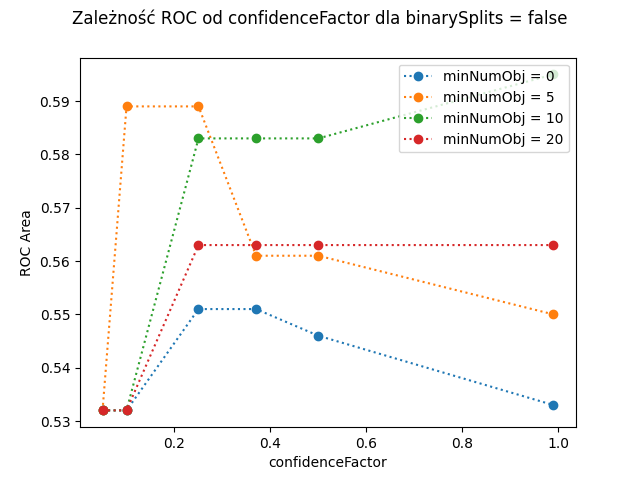
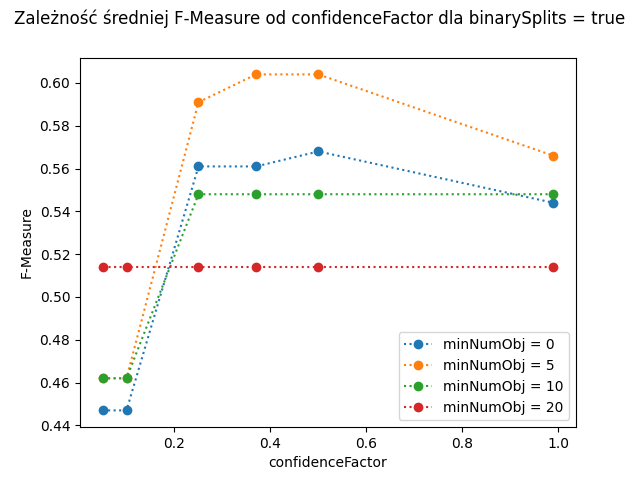
**F-Measure** – metryka, która ukazuje jakość klasyfikacji jednej klasy. Uwzględnia jednocześnie procent elementów zaklasyfikowanych pozytywnie, które są faktycznie pozytywne i procent elementów fałszywych, które są faktycznie fałszywe. F-Measure przyjmuje wartości między 0 a 1, gdzie 1 oznacza idealną równowagę między tymi dwoma miarami, a 0, że model nie jest skuteczny dla żadnej z tych miar.

**ROC-Area** – metryka, która pozwala określić skuteczność modelu w rozróżnianiu klas. Ukazuje relację między pozytywnie zaklasyfikowanymi elementami, które faktycznie są pozytywne, a pozytywnie zaklasyfikowanymi elementami, które faktycznie są pozytywne. ROC-Area przyjmuje wartości od 0 do 1, gdzie 1 oznacza, że model bezbłędnie rozróżnia klasy, 0,5 oznacza model losowy, natomiast wartości bliskie 0 wskazują, że model w większości dokonuje błędnych klasyfikacji.

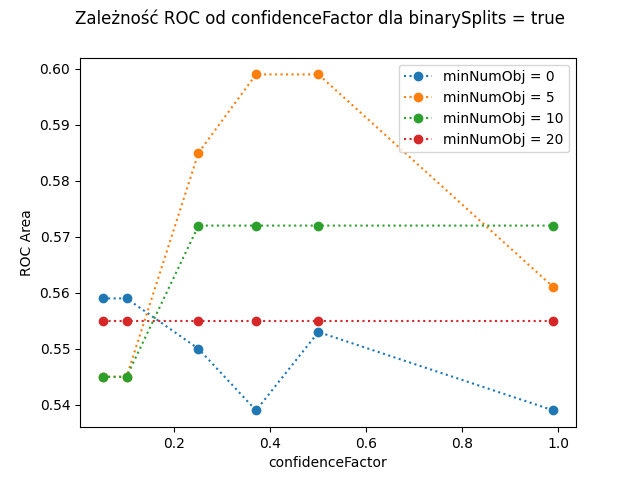
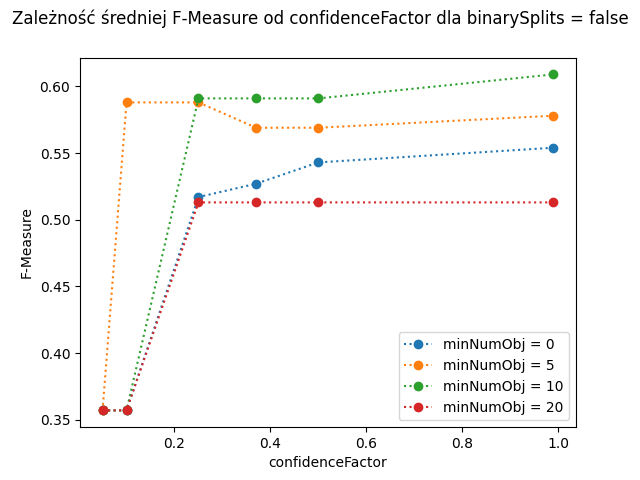
**MCC** – metryka, która uwzględnia wszystkie cztery kategorie klasyfikacji elementów.   
Przyjmuje wartości między -1 a 1, gdzie 1 oznacza doskonałą klasyfikację, 0 losową, a -1 błędną.

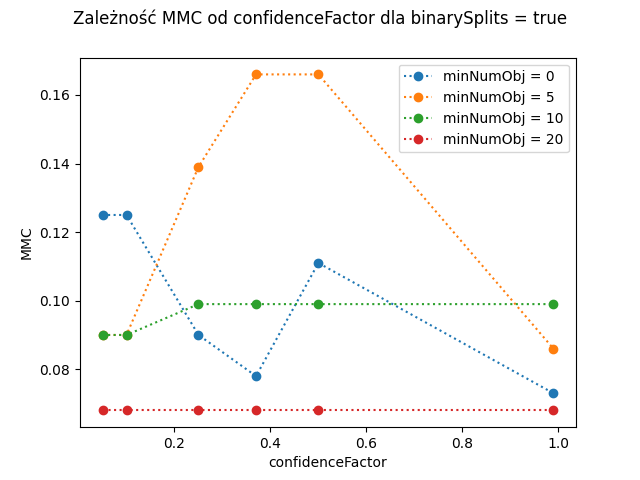
Nasze atrybuty

|  |  |
| --- | --- |
| cF = 0,05 , mNO = 0, bS = t  CCI 95 48.7179 %  ICI 100 51.2821 %  a b <-- classified as  29 89 | a = '(-inf-11.5]'  11 66 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,447 0,125 0,559 | cF = 0,05 , mNO = 0, bS = f  CCI 87 44.6154 %  ICI 108 55.3846 %  a b <-- classified as  15 103 | a = '(-inf-11.5]'  5 72 | b = '(11.5-inf)’  F-Measure MCC ROC Area  0,357 0,100 0,532 |
| cF = 0,10 , mNO = 5, bS = t  CCI 95 48.7179 %  ICI 100 51.2821 %  a b <-- classified as  34 84 | a = '(-inf-11.5]'  16 61 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,462 0,090 0,545 | cF = 0,10 , mNO = 5, bS = f  CCI 116 59.4872 %  ICI 79 40.5128 %  a b <-- classified as  84 34 | a = '(-inf-11.5]'  45 32 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,588 0,132 0,589 |
| cF = 0,25 , mNO = 10, bS = t  CCI 106 54.359 %  ICI 89 45.641 %  a b <-- classified as  61 57 | a = '(-inf-11.5]'  32 45 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,548 0,099 0,572 | cF = 0,25 , mNO = 10, bS = f  CCI 116 59.4872 %  ICI 79 40.5128 %  a b <-- classified as  82 36 | a = '(-inf-11.5]'  43 34 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,591 0,139 0,583 |
| cF = 0,37 , mNO = 20, bS = t  CCI 100 51.2821 %  ICI 95 48.7179 %  a b <-- classified as  51 67 | a = '(-inf-11.5]'  28 49 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,514 0,068 0,555 | cF = 0,37 , mNO = 20, bS = f  CCI 100 51.2821 %  ICI 95 48.7179 %  a b <-- classified as  50 68 | a = '(-inf-11.5]'  27 50 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,513 0,073 0,563 |
| cF = 0,5 , mNO = 0, bS = t  CCI 120 61.5385 %  ICI 75 38.4615 %  a b <-- classified as  89 29 | a = '(-inf-11.5]'  46 31 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,568 0,111 0,553 | cF = 0,5 , mNO = 0, bS = f  CCI 105 53.8462 %  ICI 90 46.1538 %  a b <-- classified as  61 57 | a = '(-inf-11.5]'  33 44 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,543 0,086 0,546 |
| cF = 0,99, mNO = 10, bS = t  CCI 106 54.359 %  ICI 89 45.641 %  a b <-- classified as  61 57 | a = '(-inf-11.5]'  32 45 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,548 0,099 0,572 | **cF = 0,99, mNO = 10, bS = f**  **CCI 120 61.5385 %**  **ICI 75 38.4615 %**  **a b <-- classified as**  **61 57 | a = '(-inf-11.5]'**  **33 44 | b = '(11.5-inf)'**  **F-Measure MCC ROC Area**  **0,609 0,176 0,595** |

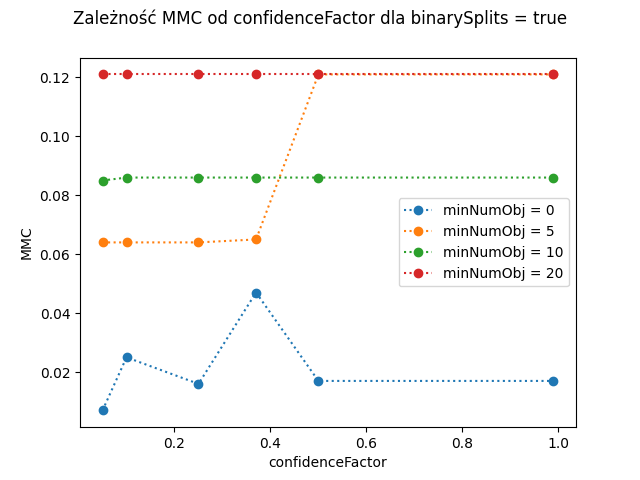
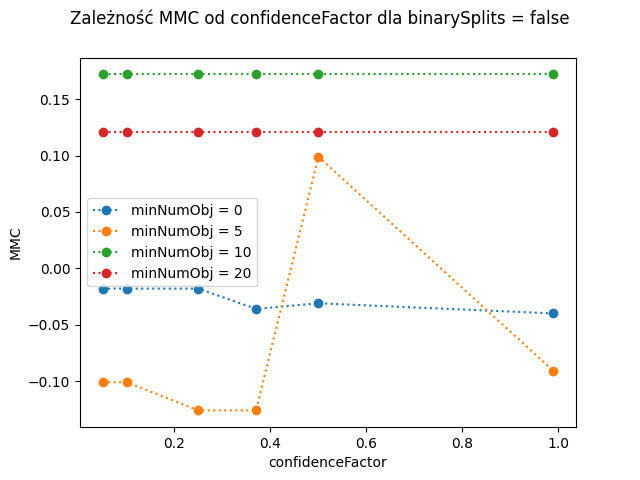
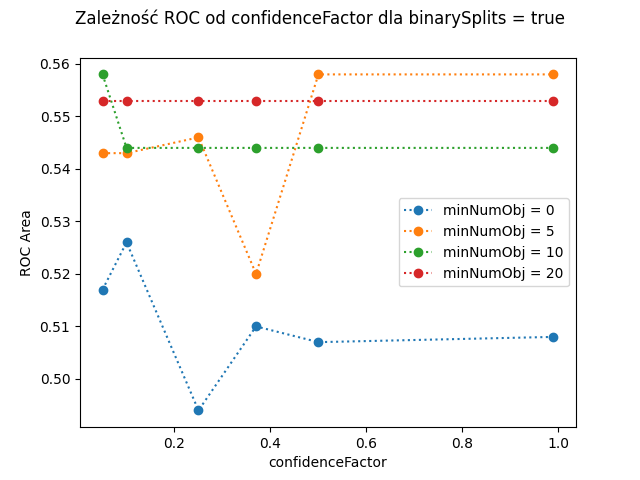
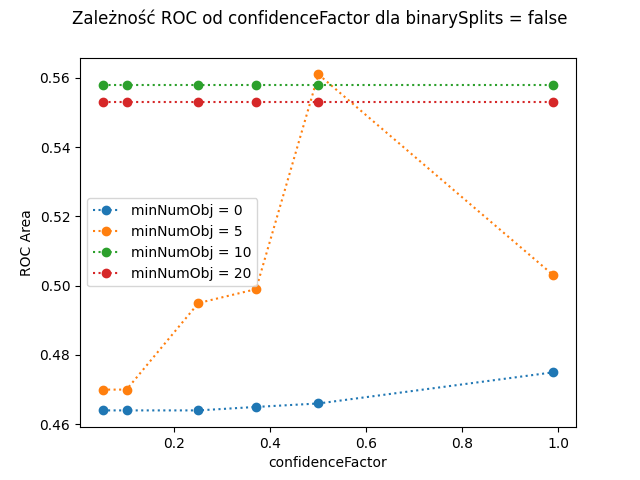
Nasze atrybuty

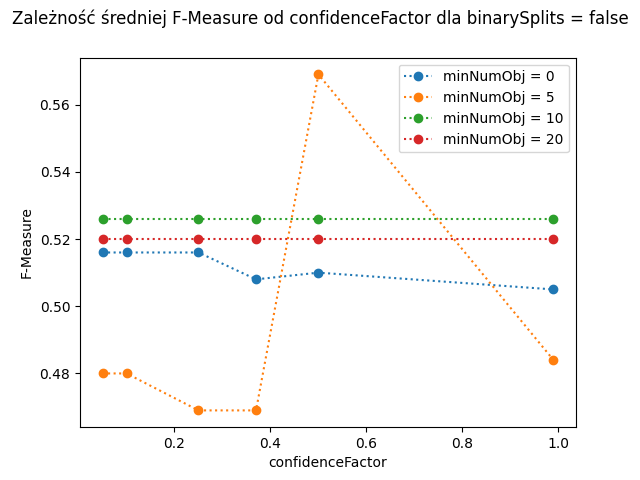
Najlepszy wynik otrzymujemy dla confidenceFactor równego 0,99 i minNumObj = 10.



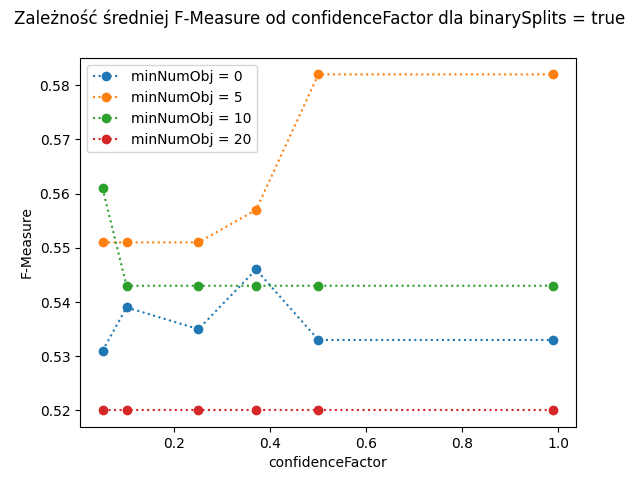
Wszystkie atrybuty

|  |  |
| --- | --- |
| cF = 0,05 , mNO = 20, bS = t  CCI 121 62.0513 %  ICI 74 37.9487 %  a b <-- classified as  114 4 | a = '(-inf-11.5]'  70 7 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,520 0,121 0,553 | cF = 0,05 , mNO = 20, bS = f  CCI 121 62.0513 %  ICI 74 37.9487 %  a b <-- classified as  114 4 | a = '(-inf-11.5]'  70 7 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,520 0,121 0,553 |
| cF = 0,10 , mNO = 10, bS = t  CCI 105 53.8462 %  ICI 90 46.1538 %  a b <-- classified as  61 57 | a = '(-inf-11.5]'  33 44 | b = '(11.5-inf)'F-Measure MCC ROC Area  0,543 0,086 0,544 | cF = 0,10 , mNO = 10, bS = f  CCI 123 63.0769 %  ICI 72 36.9231 %  a b <-- classified as  116 2 | a = '(-inf-11.5]'  70 7 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,526 0,172 0,558 |
| cF = 0,25 , mNO = 5, bS = t  CCI 107 54.8718 %  ICI 88 45.1282 %  a b <-- classified as  72 46 | a = '(-inf-11.5]'  42 35 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,551 0,064 0,546 | cF = 0,25 , mNO = 5, bS = f  CCI 99 50.7692 %  ICI 96 49.2308 %  a b <-- classified as  87 31 | a = '(-inf-11.5]'  65 12 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,469 -0,126 0,495 |
| cF = 0,37 , mNO = 0, bS = t  CCI 107 54.8718 %  ICI 88 45.1282 %  a b <-- classified as  76 42 | a = '(-inf-11.5]'  46 31 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,546 0,047 0,510 | cF = 0,37 , mNO = 0, bS = f  CCI 100 51.2821 %  ICI 95 48.7179 %  a b <-- classified as  74 44 | a = '(-inf-11.5]'  51 26 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,508 -0,036 0,465 |
| **cF = 0,5 , mNO = 5, bS = t**  **CCI 114 58.4615 %**  **ICI 81 41.5385 %**  **a b <-- classified as**  **80 38 | a = '(-inf-11.5]'**  **43 34 | b = '(11.5-inf)'**  **F-Measure MCC ROC Area**  **0,582 0,121 0,558** | cF = 0,5 , mNO = 5, bS = f  CCI 111 56.9231 %  ICI 84 43.0769 %  a b <-- classified as  76 42 | a = '(-inf-11.5]'  42 35 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,569 0,099 0,561 |
| cF = 0,99, mNO = 20, bS = t  CCI 121 62.0513 %  ICI 74 37.9487 %  a b <-- classified as  114 4 | a = '(-inf-11.5]'  70 7 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,520 0,121 0,553 | cF = 0,99, mNO = 20, bS = f  CCI 121 62.0513 %  ICI 74 37.9487 %  a b <-- classified as  114 4 | a = '(-inf-11.5]'  70 7 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,520 0,121 0,553 |

Wszystkie atrybuty



Najlepszy wynik otrzymujemy dla confidenceFactor równego 0,5 i minNumObj = 5.

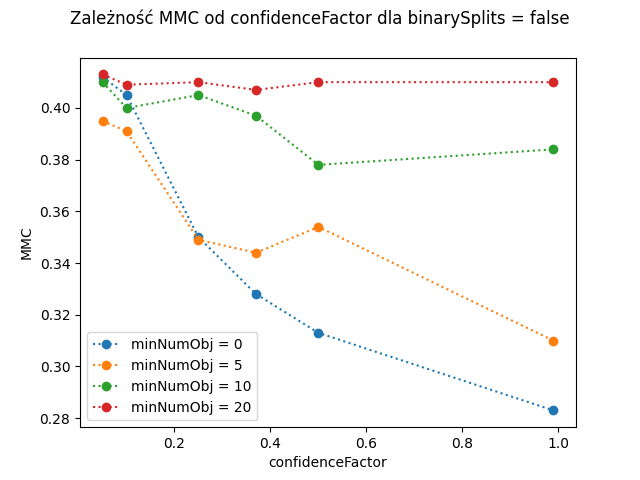


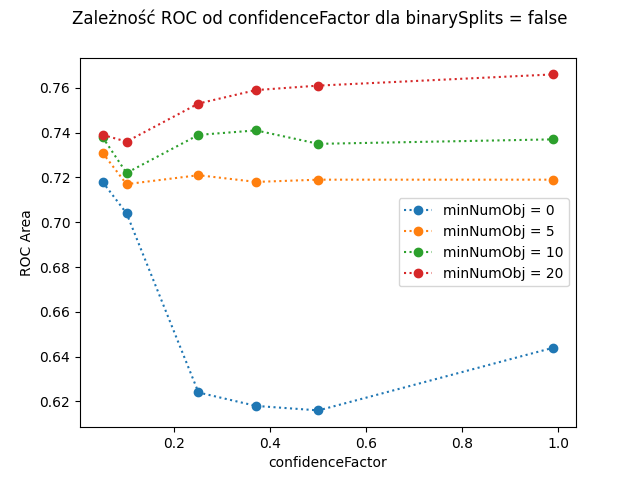
Portugalski (cross-validation, k = 10, bs = false)

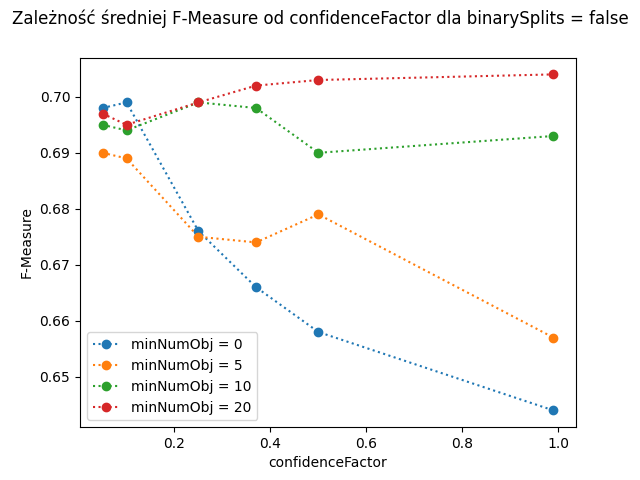
|  |
| --- |
| cF = 0,05 , mNO = 10, bS = f  CCI  ICI  a b <-- classified as  F-Measure MCC ROC Area |
| cF = 0,10 , mNO = 5, bS = f  CCI  ICI  a b <-- classified as  F-Measure MCC ROC Area |
| cF = 0,25 , mNO = 0, bS = f  CCI  ICI  a b <-- classified as  F-Measure MCC ROC Area |
| cF = 0,37 , mNO = 5, bS = f  CCI  ICI  a b <-- classified as  F-Measure MCC ROC Area |
| cF = 0,5 , mNO = 10, bS = f  CCI  ICI  a b <-- classified as  F-Measure MCC ROC Area |
| **cF = 0,99, mNO = 20, bS = f**  **CCI 459 70.7242 %**  **ICI 190 29.2758 %**  **a b <-- classified as**  **180 121 | a = '(-inf-11.5]'**  **69 279 | b = '(11.5-inf)'**  **F-Measure MCC ROC Area**  **0,704 0,410 0,766** |

Najlepszy wynik otrzymujemy dla confidenceFactor równego 0,99 i minNumObj = 20.

Portugalski (cross-validation, k = 10, bs = false)







Najlepsze drzewa:

|  |  |
| --- | --- |
| Matematyka  min 5 con 0,5 bs t  failures <= 1.0  | schoolsup = yes  | | studytime <= 1.0: '(11.5-inf)' (9.0/4.0)  | | studytime > 1.0: '(-inf-11.5]' (28.0/1.0)  | schoolsup != yes  | | Mjob = at\_home: '(-inf-11.5]' (15.0/4.0)  | | Mjob != at\_home  | | | Mjob = services  | | | | famrel <= 4.0  | | | | | health <= 2.0: '(11.5-inf)' (7.0)  | | | | | health > 2.0  | | | | | | absences <= 7.0: '(11.5-inf)' (12.0/3.0)  | | | | | | absences > 7.0: '(-inf-11.5]' (5.0)  | | | | famrel > 4.0: '(11.5-inf)' (9.0)  | | | Mjob != services  | | | | failures <= 0.0  | | | | | Mjob = health: '(11.5-inf)' (16.0/4.0)  | | | | | Mjob != health  | | | | | | paid = no  | | | | | | | freetime <= 3.0  | | | | | | | | age <= 15.0: '(11.5-inf)' (7.0/2.0)  | | | | | | | | age > 15.0  | | | | | | | | | studytime <= 1.0: '(11.5-inf)' (6.0/2.0)  | | | | | | | | | studytime > 1.0: '(-inf-11.5]' (6.0)  | | | | | | | freetime > 3.0: '(11.5-inf)' (20.0/3.0)  | | | | | | paid != no  | | | | | | | Mjob = other  | | | | | | | | sex = F: '(-inf-11.5]' (12.0/3.0)  | | | | | | | | sex != F: '(11.5-inf)' (7.0/1.0)  | | | | | | | Mjob != other: '(-inf-11.5]' (13.0/2.0)  | | | | failures > 0.0: '(-inf-11.5]' (5.0/1.0)  failures > 1.0: '(-inf-11.5]' (23.0) | Portugalski  min 20 con 0,99 bs f  failures <= 0  | higher = yes  | | school = GP  | | | schoolsup = yes  | | | | age <= 15: '(-inf-11.5]' (22.0/9.0)  | | | | age > 15: '(11.5-inf)' (26.0/10.0)  | | | schoolsup = no  | | | | Walc <= 3: '(11.5-inf)' (254.0/50.0)  | | | | Walc > 3  | | | | | activities = no: '(-inf-11.5]' (26.0/11.0)  | | | | | activities = yes: '(11.5-inf)' (32.0/11.0)  | | school = MS  | | | studytime <= 2  | | | | Walc <= 2  | | | | | Mjob = at\_home: '(-inf-11.5]' (21.0/7.0)  | | | | | Mjob = health: '(-inf-11.5]' (7.0/3.0)  | | | | | Mjob = other: '(11.5-inf)' (34.0/13.0)  | | | | | Mjob = services: '(11.5-inf)' (11.0/2.0)  | | | | | Mjob = teacher: '(-inf-11.5]' (6.0/3.0)  | | | | Walc > 2: '(-inf-11.5]' (50.0/13.0)  | | | studytime > 2: '(11.5-inf)' (24.0/5.0)  | higher = no: '(-inf-11.5]' (36.0/5.0)  failures > 0  | Fjob = teacher: '(11.5-inf)' (1.0)  | Fjob = other: '(-inf-11.5]' (60.0/3.0)  | Fjob = services: '(-inf-11.5]' (29.0)  | Fjob = health: '(11.5-inf)' (1.0)  | Fjob = at\_home: '(-inf-11.5]' (9.0/2.0) |

Drzewo dla matematyki jest większe niż dla języka portugalskiego. Dla obu drzew najistotniejszym atrybutem jest failures, które w tworzą korzeń. Dla języka portugalskiego, dwoma z trzech najistotniejszych atrybutów jest higher i Fjob, które w ogóle nie pojawia się w drzewie stworzonego dla matematyki. W obu drzewach pojawiają się Mjob, schoolsup i studytime, aczkolwiek dla drzewa matematyki są to atrybuty istotniejsze niż dla drzewa języka portugalskiego.

Zatem atrybuty, na których podstawie możemy określić pilność uczniów wobec obu przedmiotów to failures i schoolsup.

Porównanie z innymi algorytmami

|  |  |
| --- | --- |
| J48 | cF = 0,5 , mNO = 5, bS = t  CCI 114 58.4615 %  ICI 81 41.5385 %  a b <-- classified as  80 38 | a = '(-inf-11.5]'  43 34 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,582 0,121 0,558 |
| Naive Bayes | CCI 126 64.6154 %  ICI 69 35.3846 %  a b <-- classified as  101 17 | a = '(-inf-11.5]'  52 25 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,617 0,215 0,668 |
| Modlem | CCI 125 64.1026 %  ICI 70 35.8974 %  a b <-- classified as  106 12 | a = '(-inf-11.5]'  58 19 | b = '(11.5-inf)'  F-Measure MCC ROC Area  0,594 0,194 0,573 |

Na podstawie powyższego porównania widzimy, że najskuteczniejszym algorytmem jest Naive Bayes, które otrzymuje najlepsze wyniki dla trzech wybranych miar.

W poniższej tabeli zestawiono 8 wybranych na początku zadania atrybutów oraz 8 najistotniejszych atrybutów dla Naive Bayes i Modlem.

|  |  |  |
| --- | --- | --- |
| Nasze atrybuty   * **reason** * ***failures*** * ***higher*** * ***Medu*** * Fedu * ***schoolsup*** * internet * *absences* | Naïve Bayes   1. **failures**, 2. **schoolsup**, 3. **Medu**, 4. sex, 5. Mjob, 6. **higher**, 7. address 8. **reason** | Modlem   1. *failures*, 2. freetime, 3. *absences,* 4. *schoolsup*, 5. studytime, 6. *Medu*, 7. guardian 8. *higher* |

6 z wybranych na początku atrybutów znajduje się wśród 8 najistotniejszych atrybutów dla wskazanych algorytmów, przy czym 4 z nich znajdują się w obu zbiorach dla algorytmów. Atrybutami, które nie pojawiły się ani razu ani w 8 najistotniejszych atrybutach dla powyższych algorytmów, ani dla drzew najlepszych wyników, są internet i Fedu. To drugie jest tym ciekawsze, iż Medu pojawia się zarówno w Naive Bayes, jak i Modlem, aczkolwiek w żadnym z drzew wygenerowanych dla J48.