Katedra Systemów Informacyjnych

Grzegorz Cichosz

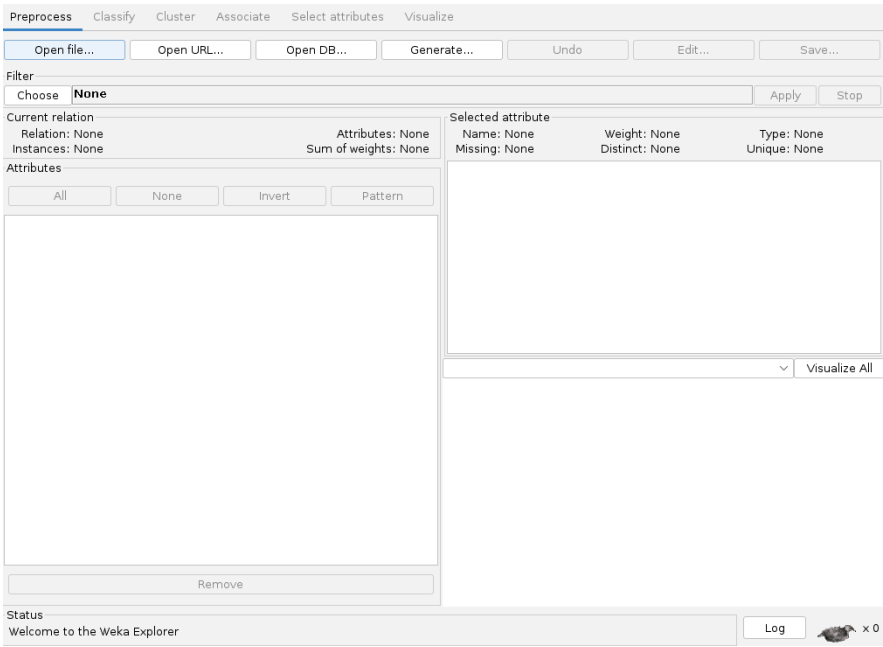
Algorytmy analizy danych
Kierunek Informatyka, WE, sem. VII

Laboratorium nr 7

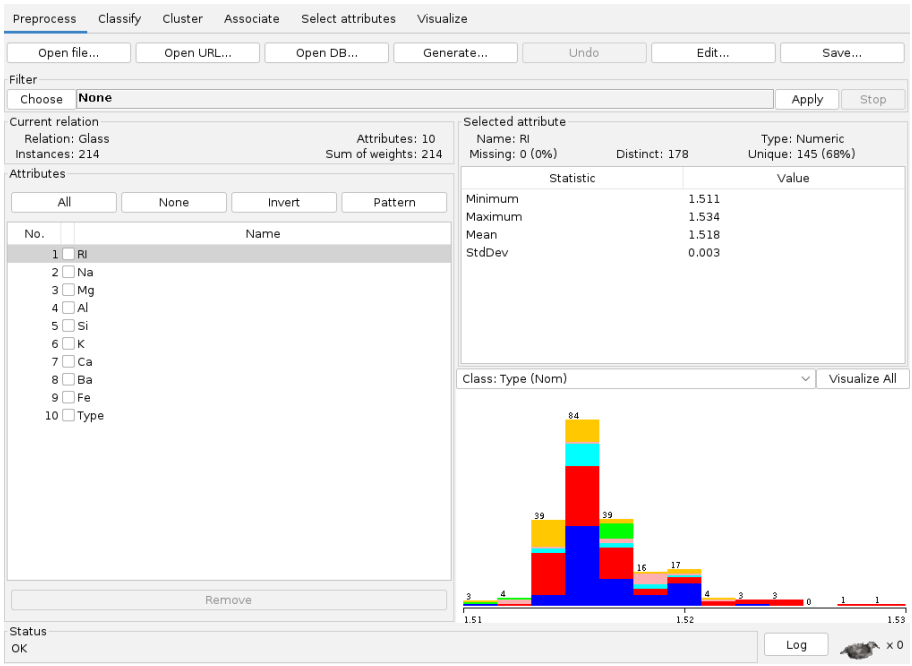
Celem zajęć poznanie działania wybranych algorytmów klasyfikacji danych oraz procesu uczenia klasyfikatorów. Celem zajęć jest także poznanie narzędzia WEKA oraz jego interfejsu dedykowanego obliczeniom uczenia maszynowego.

Przygotuj sprawozdanie odpadając na poniższe polecenia. Sprawozdanie prześlij w iliasie.

1. Cześć 2 - polecenia
    a. Uruchom program WEKA a następnie przejdź do programu Explorer



b. W zakładce Preprocess wczytaj plik z danymi glass.arff
Opis danych glass znajdziesz pod adresem:
https://archive.ics.uci.edu/ml/datasets/glass+identification
Zapoznaj się z charakterem problemu opisanego tym zestawem danych?

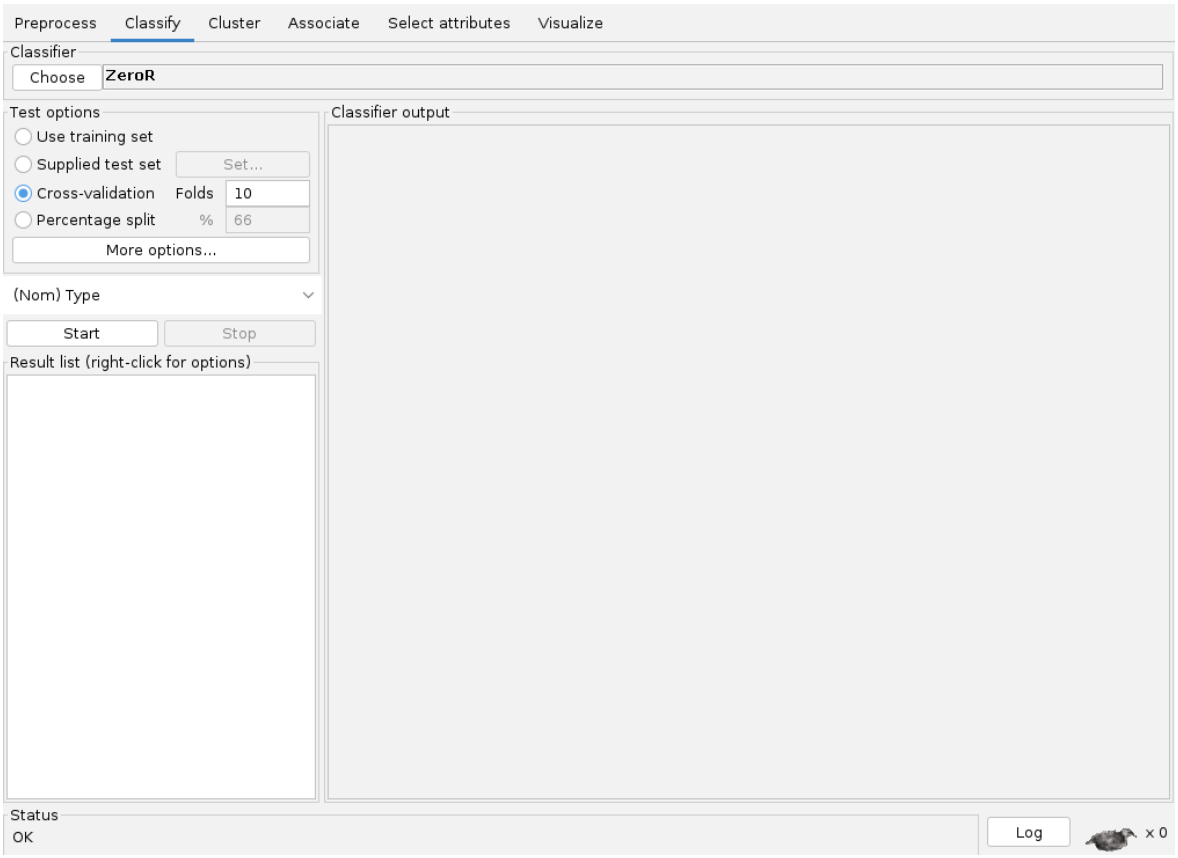


Zbiór danych Glass opisuje szkło za pomocą cech numerycznych, takich jak skład chemiczny i współczynnik załamania światła. Zmienną decyzyjną jest atrybut Type, określający przeznaczenie szkła. Dane nie zawierają braków, a cechy różnią się między klasami.

c. Przejrzyj wszystkie informacje dostępne na zakładce Preprocess, które pojawiły się po wczytaniu danych.

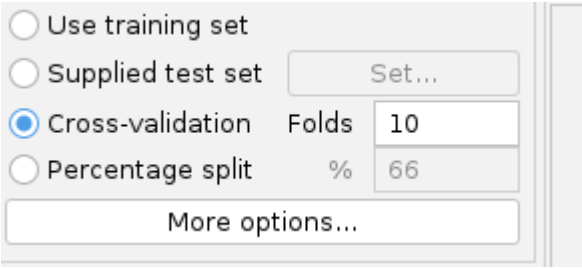
d. Przejdź do zakładki Classify
Teraz będziemy budować różne model klasyfikatorów.
Mając na uwadze to zadanie oraz charakter wczytanych danych, czego tak naprawdę będziemy szukać?
Odpowiedź umieść w sprawozdaniu.

Celem zadania jest sprawdzenie, czy na podstawie składu chemicznego oraz współczynnika załamania światła można poprawnie określić przeznaczenie szkła. Będziemy szukać zależności między tymi cechami a klasą, aby móc automatycznie klasyfikować próbki szkła.

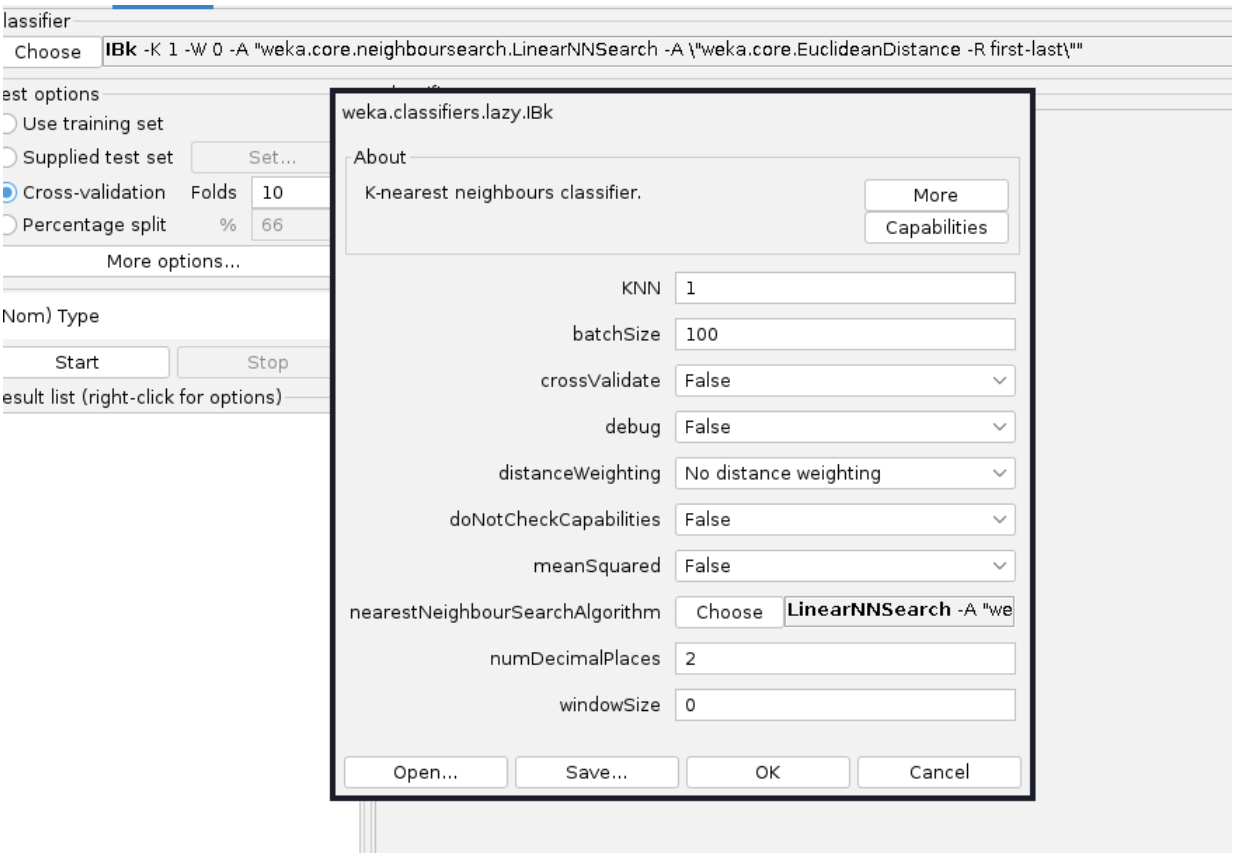


e. W Test option ustaw Cross-validation oraz Fold na 10.
Czego dotyczy to ustawinie? Możesz o niej poczytać np. pod adresem:
https://machinelearningmastery.com/k-fold-cross-validation/
W sprawozdaniu opisz swoimi słowami jakie na znaczenie opcja Cross-validation.

Cross-validation służy do oceny i porównywania klasyfikatorów.



f. W opcji Classifer (przycisk Choose) wybierz ***lazy*** a następnie IBk – wybrałeś algorytm najbliższego sąsiada. Klikając w opis tego algorytmu (obok przycisku Choose) zobacz jakie parametry dla tego algorytmu są możliwe do ustawiania. Zwróć uwagę na KNN. W szczególności czego dotyczy to ustawinie? Opisz w sprawozdaniu.



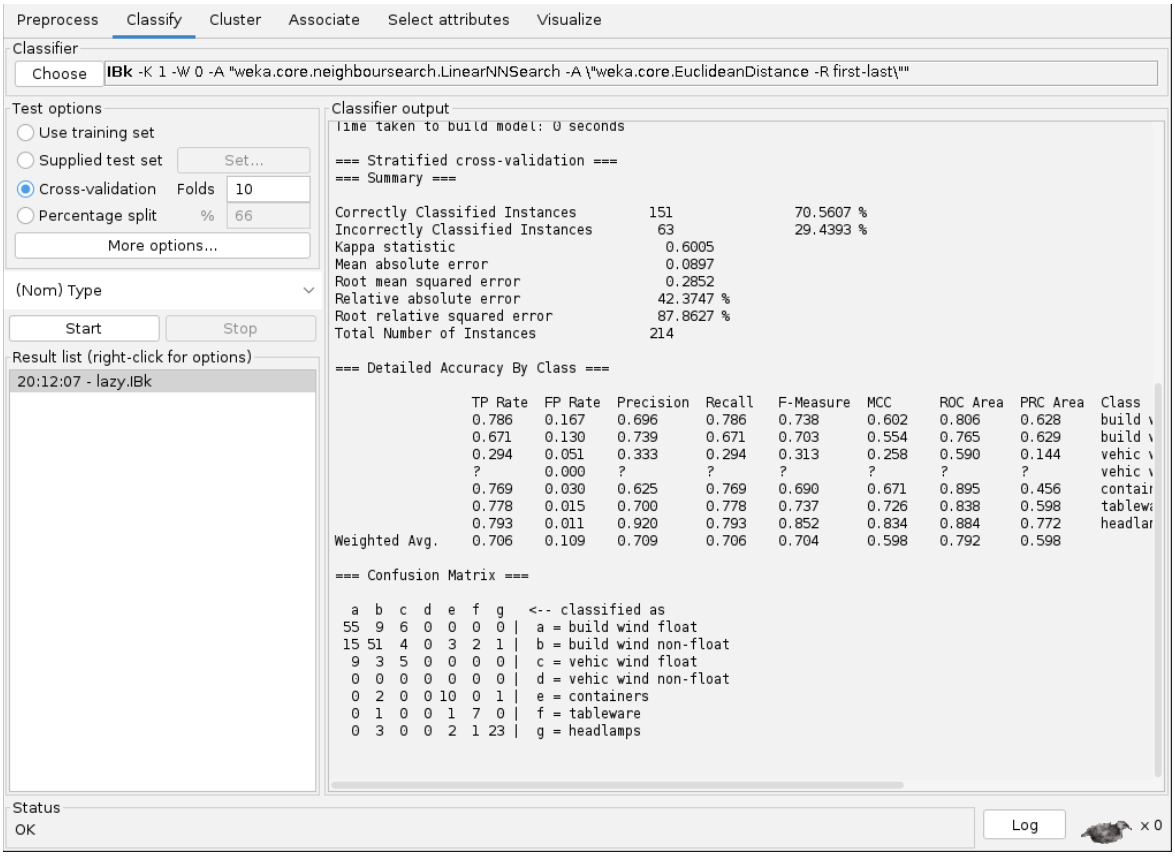
KNN to liczba grup

g. Czy IBk można zastosować dla danych glass.arff?
Odpowiedź z uzasadnieniem umieść w sprawozdaniu.

Tak, ponieważ dane są poprawnie sformatowane

h. Uruchom uczenie klasyfikatora (przycisk Start) a następnie w sprawozdaniu opisz poszczególne składowe raportu (Classifier output). Szczególnie przeanalizuj
- Summary
- Detailed Accuracy by Class
- Confusion Matrix

Opisz w sprawozdaniu poszczególne parametry związane z otrzymanym modelem klasyfikatora.
Która z klas decyzyjnych jest lepiej identyfikowana przez otrzymany model? Odpowiedź uzasadnij wskazując argumenty za.



Sekcja Summary
Przedstawia ogólna skuteczność klasyfikatora. Model poprawnie klasyfikuje około 70% obiektów, a wartość statystyki Kappa wskazuje na umiarkowanie dobrą jakość klasyfikacji.

Detailed Accuracy by Class
Ta część pokazuje skuteczność klasyfikatora dla poszczególnych klas. Najlepsze wyniki (wysokie Precision, Recall i F-Measure) osiąga klasa headlamps, natomiast klasy szkła samochodowego są rozpoznawane słabiej.

Confusion Matrix
Macierz pomyłek pokazuje liczbę poprawnych i błędnych klasyfikacji. Wynika z niej, że niektóre klasy są często mylone, natomiast klasa headlamps jest rzadko błędnie klasyfikowana.

Najlepiej rozpoznawaną klasą jest headlamps, co potwierdzają wysokie wartości miar jakości oraz duża liczba poprawnych klasyfikacji w macierzy pomyłek.

i. Przeprowadź eksperyment, którego celem będzie wskazanie najlepszego możliwego modelu klasyfikatora opartego na koncepcji kNN.
Eksperyment przeprowadź zmieniając k od 1 do 10, wpierw badając odległość w oparciu o miarę Euclidean, potem zmieniając ją na Manhattan i inne miary (jeśli jednak użyjesz innych miar, upewnij się czy są odpowiednie dla danych, które przetwarzasz, opisz też te miary w sprawozdaniu).
Sporządź tabele z wynikami (dwie tabela lub więcej, każda dla innej zastosowanej miary), gdzie dla różnych k podaj:
- Correctly Classified Instances
- Precision
- Recall
- F-Measure
- ROC Area

Tabele umieść w sprawozdaniu.
Wskaż najlepszy uzyskany model, oraz uzasadnij jego wybór.
Czy generalnie zmiany parametrów (k oraz miary odległości) miały wpływ na jakość modelu?

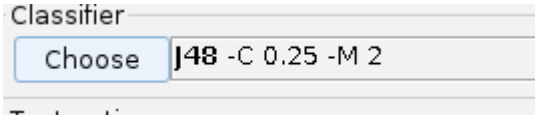
Euclidean

k	Correctly Classified Instances	Precision	Recall	F-Measure	ROC Area
1	151	0.696	0.786	0.738	0.806
2	145	0.612	0.857	0.714	0.835
3	154	0.656	0.843	0.738	0.865
4	147	0.632	0.857	0.727	0.863
5	145	0.641	0.843	0.728	0.867
6	143	0.621	0.843	0.715	0.869
7	137	0.598	0.829	0.695	0.876
8	137	0.61	0.871	0.718	0.881
9	135	0.592	0.829	0.69	0.881
10	142	0.602	0.886	0.717	0.881

Najlepsze wyniki uzyskano dla klasyfikatora kNN z k = 3 (oraz k = 1).
Dla k = 3 model osiągnął najwyższą liczbę poprawnych klasyfikacji (154) oraz najwyższą wartość F-Measure (0,738) przy jednocześnie wysokim ROC Area (0,865).
Model z k = 1 osiąga porównywalny F-Measure, jednak charakteryzuje się niższym ROC Area.
Zmiana parametru k miała wpływ na jakość modelu.
Dla małych wartości k (1–3) model osiągał najlepsze wyniki, natomiast wraz ze wzrostem k obserwowany jest spadek liczby poprawnych klasyfikacji oraz wartości F-Measure. Zbyt duże k prowadzi do pogorszenia skuteczności klasyfikatora.

j. Zmień algorytm na J48 (J48 to niekomercyjna wersja C4.5).
Jaka jest różnica pomiędzy J48 a kNN? Czym różnią się oba algorytmy? Odpowiedzi zawrzyj

w sprawozdaniu.

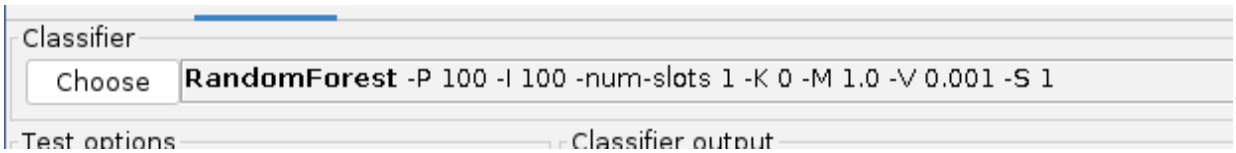


J48 tworzy drzewo decyzyjne, na podstawie którego podejmuje decyzje, natomiast kNN porównuje nowy obiekt z innymi przykładami i przypisuje klasę na podstawie najbliższych sąsiadów.

k. Uruchom algorytm J48.
Uruchom J48 dla dwóch ustawień: unpruned = False (standardowego) oraz unpruned = True.
Czego dotyczy to ustawianie opisz w sprawozdaniu.
Porównaj wyniki J48 z IBk.
W sprawozdaniu opisz który z algorytmów okazał się lepszym w teście 10CV.

Parametr unpruned określa, czy drzewo decyzyjne ma być przycinane.
lepszym algorytmem dla badanego zbioru danych okazał się IBk, który zapewnił wyższą jakość klasyfikacji w teście 10CV.

l. Zmień na algorytm RandomForest.
Co to za algorytm? Jego cechy opisz w sprawozdaniu.



RandomForest to algorytm wykorzystujący wiele drzew decyzyjnych, których wyniki są łączone przez głosowanie. Dzięki temu jest bardziej stabilny i mniej podatny na przeuczenie niż pojedyncze drzewo.

m. Uruchom algorytm dla standardowych ustawień dla bagSizePercent oraz batchSize, a następnie dla ustawień 100.
Czy zmiana tych ustawień miała wpływ na wyniki?

Zmiana ustawień miała wpływ na wyniki działania algorytmu. Większe wartości parametrów poprawiały dokładność klasyfikatora, jednak powodowały także wolniejsze działanie modelu

n. Porównaj RandomForest z J48 oraz IBk – oczywiście w sensie jakości modeli.

Pod względem jakości klasyfikacji najlepsze wyniki uzyskał RandomForest, następnie IBk, a najsłabsze rezultaty osiągnął J48

~~o. Zmień algorytm na MultilayerPerceptron.~~
~~Co to za algorytm? Odpowiedź umieść w sprawozdaniu.~~

~~Możesz pominąć punkty o-r (pkt. te mogą być kosztowne obliczeniowo, ale możliwe do wykonania!) i przejść do pkt. s, który jest niezbędny aby porównać uzyskane na tym etapie wyniki.~~

~~p. Zmieniając parametr hiddenlayers kolejno na a, i, o, t dokonaj uczenia modelu. Czego dotyczy ten parametr? Czy zmieniając go uzyskałeś lepszy model klasyfikatora?~~

q. ~~Dokonaj teraz serii eksperymentów, za każdym razem dla innego parametru hiddenlayers, ale zmieniaj wartość learning rate z 0.1 do 1 z krokiem 0.1, za każdym razem zmieniając momentum z 0.1 do 1 też krokiem 0.1. Wyniki Correctly Classified Instances, Precision, Recall, F-Measure, ROC Area umieść w stosownych tabelach oraz w sprawozdaniu. Wskaż też najlepszy z uzyskanych modeli (przy jakich parametrach go uzyskałeś?).~~

r. ~~W opcjach MultilayerPerceptron ustaw GUI na True, zmieniając hiddenlayers na a, i, o, t i wybierając Start (uczenie modelu klasyfikatora) zobacz jak WEKA w interfejsie GUI zobrazowała strukturę tego algorytmu. Następnie dane uczenie uruchom na przyciski Start okna GUI.~~
~~Do czego służy parametr Num Of Epoch. Jak zachowuje się proces uczenia zmieniając ten parametr? Odpwiedź umieść w sprawozdaniu,~~

s. Porównaj wyniki dla najlepszych IBk, J48, RandomForest oraz MultilayerPerceptron. Który z modeli klasyfikatora okazał się najlepszy dla danych glass biorąc pod uwagę wskaźniki jakościowe? Odpowiedź z uzasadnieniem umieść w sprawozdaniu.

Klasyfikator	Correctly Classified Instances	Precision	Recall	F-Measure	ROC Area
IBk (k=3 Euclidean)	154	0.656	0.843	0.738	0.865
J48 pruned	143	0.667	0.714	0.69	0.806
J48 unpruned	144	0.667	0.714	0.69	0.809
RandomForest (bagSize 100%)	171	0.782	0.871	0.824	0.934
Multilayer Perceptron	145	0.667	0.743	0.703	0.862

Na podstawie porównania wskaźników jakościowych najlepszym klasyfikatorem dla zbioru danych glass okazał się RandomForest z parametrami bagSizePercent = 100 oraz batchSize = 100. Model ten osiągnął najwyższą skuteczność klasyfikacji oraz najwyższe wartości Precision, F-Measure i ROC Area, przy jednocześnie bardzo wysokim Recall, co potwierdza jego przewagę nad pozostałymi algorytmami.

t. Uwaga. Każdy z modeli możesz zapisać (prawy przycisk na Result list). Zapisz więc każdy z najlepszych modeli dla użytych algorytmów uczenia maszynowego.

Sprawozdanie prześlij w iliasie !