Realizacja projektu:
W zespołach 2-osobowych (można się mieszać pomiędzy grupami).

Cel projektu:
Sprawdzić, jak jakość klasyfikacji (mierzona częstością poprawnej decyzji) zależy od liczby stosowanych cech. Gdy liczba cech jest niewielka, to sprawdzić dla wszystkich przypadków. Gdy cech jest więcej, to można zacząć od 1 cechy, kolejno dokładać następne (według wyznaczonego rankingu) i robić to tak długo, dopóki jakość klasyfikacji będzie się poprawiała.
Badania przeprowadzić metodą 5 razy powtarzanej dwukrotnej walidacji krzyżowej (5x2CV). Zbiór danych dzielimy losowo na dwie równe części: jedna stanowi zbiór uczący (ZU), a druga zbiór testujący (ZT), uczymy klasyfikator na ZU i testujemy na ZT, potem zamieniamy miejscami zbiory: ZU jest teraz ZT, a ZT jest ZU i robimy to samo. Całą procedurę powtarzamy 5 razy (każdorazowo na nowo losujemy 2 części).Uzyskanych 10 wyników uśredniamy.

Kolejne etapy realizacji projektu:

1. Zapoznać się z algorytmami diagnostycznymi (algorytmami klasyfikacji), określonymi w temacie.
2. Zapoznać się z materiałem empirycznym – zdefiniować problem rozpoznawania (klasyfikacji) –określić liczbę i znaczenie klas, liczbę i znaczenie cech oraz charakter cech (ciągłe, wielowartościowe, binarne, itp.)
3. Wyznaczyć ranking cech pod względem ich przydatności do klasyfikacji) korzystając z dowolnej miary (kryterium) jakości cech stosowanych w selekcji cech z grupy metod zwanych *filtrami* (proponuję kryterium Kołmogorowa, gdyż jest bardzo proste rachunkowo).
4. Zaplanować badania eksperymentalne dla następujących założeń
   a) Trenowanie i testowanie zastosowanych klasyfikatorów z wykorzystaniem 5 razy powtarzanej metody 2-krotnej walidacji krzyżowej. Jakość klasyfikacji (poprawność diagnozy) należy mierzyć częstością poprawnych rozpoznań (diagnoz) na zbiorze testującym.
   b) Przeprowadzić badania dla różnej liczby cech (poczynając od jednej –najlepszej wg. wyznaczonego rankingu, a następnie dokładać kolejno po jednej (również według wyznaczonego rankingu) tak długo, aż zostanie znaleziona najlepsza liczba cech. Dodawanie cech powinno poprawiać jakość klasyfikacji, ale do pewnego momentu – dalsze dodawanie cech jakość pogorszy. Trzeba znaleźć optimum. Jak cech jest mało (<7), to przeprowadzić badania dla wszystkich cech.
   c) Przeprowadzić badania dla następujących algorytmów:
      - Dla sztucznych sieci neuronowych – sieć jednokierunkowa z 1 warstwą ukrytą dla 3 różnych liczb neuronów w warstwie ukrytej oraz dla uczenia metodą propagacji wstecznej z momentum i bez momentum.
      - Dla algorytmu k-NN – dla 3 różnych wartości k (1, 5, 10) oraz dla 2 różnych miar odległości (w tym Euklidesowa)
      - Dla naiwnego algorytmu Bayesa (zakładamy, że cechy są niezależne) –algorytm dla 0-1 funkcji strat. Prawdopodobieństwa *a priori* oraz prawdopodobieństwa cech w poszczególnych klasach (dla cech dyskretnych) szacujemy (estymujemy) ze zbioru uczącego metodą częstotliwościową i/lub warunkowe gęstości cech w klasach (dla cech ciągłych) szacujemy metodą histogramu, metodą empirycznej dystrybuanty lub metodami jądrowymi (*kernel methods*)
   d) Dla każdego pojedynczego eksperymentu (pojedynczy eksperyment to doświadczalne wyznaczenie jakości klasyfikacji dla danego algorytmu, danych wartości parametrów algorytmu i dla danej liczby cech) należy przedstawić wyniki (jakości klasyfikacji) w formie uśrednionej (względem 5 powtórzeń metody 2-krotnej walidacji krzyżowej). Dodatkowo, dla najlepszego przypadku należy przedstawić macierz konfuzji (pomyłek).
5. Zaimplementować algorytmy diagnostyczne (klasyfikacji) (środowisko implementacji dowolne), tak aby można było przeprowadzić badania eksperymentalne według przedstawionych założeń.
6. Zrealizować badania eksperymentalne według przedstawionych w punkcie 4 założeń.
7. Przeprowadzić dyskusję wyników i przedstawić wnioski.

Warunki zaliczenia:

1. Sporządzenie sprawozdania z wykonanego projektu zawierającego:
   - Opis problemu medycznego jako zadania klasyfikacji (liczba klas i ich medyczny sens, liczba cech i ich charakterystyka (znaczenie, czy ciągłe, czy dyskretne), liczba danych w dostępnym zbiorze);
   - Przedstawić zastosowany algorytm selekcji cech– forma opisu algorytmu: patrz punkt następny;
   - Przedstawienie stosowanego algorytmu: dla algorytmów minimalno-odległościowych i naiwnego algorytmu bayesowskiego w formie, w jakiej się algorytmy przedstawia (schemat blokowy, pseudokod, formuły matematyczne – wszystko powinno być precyzyjne, aby można było z opisu utworzyć kod), dla sztucznych sieci neuronowych opisać precyzyjnie strukturę stosowanej sieci (liczba warstw, liczba neuronów), funkcję aktywacji, parametry metody uczenia BP;
   - Opisać środowisko programistyczne – krótko;
   - Przedstawić plan eksperymentu oraz jego wyniki (w formie syntetycznej: tabela lub wykres + macierz konfuzji);
   - Przeprowadzić dyskusję otrzymanych wyników – czy są zauważalne jakieś prawidłowości, czy można sformułować jakieś wnioski, itp.;
   - Wykorzystana literatura.
2. Przesłanie sprawozdanie w formie elektronicznej do 2 grudnia godz. 23.59 (marek.kurzynski@pwr.edu.pl)
3. Możliwa odpowiedź (też elektronicznie wysłana):
   - Jest OK – wtedy należy dostarczyć wersję papierową sprawozdania do 12 grudnia (albo do pok. 110C-3, albo do mojej przegródki w sekretariacie pok. 16/17 C-3) oraz cieszyć się uzyskanym zaliczeniem;
   - Zaproszenie na rozmowę, gdy sprawozdanie jest niejasne – wtedy proszę przyjść w ostatni poniedziałek zajęć (10 grudnia) ze sprawozdaniem papierowym i –po udzielonych wyjaśnieniach, cieszyć się uzyskanym zaliczeniem.

DODATKOWE UWAGI:

1. W sprawozdaniu proszę się nie rozpisywać (krótko, ale treściwie i poprawnie pod względem formalizmów)
2. Bardzo proszę o przestrzeganie terminów – podane daty to są deadline'y. Można gotowe sprawozdanie przesłać wcześniej. Wtedy ja również wcześniej odpowiem i wcześniej będzie zaliczenie.