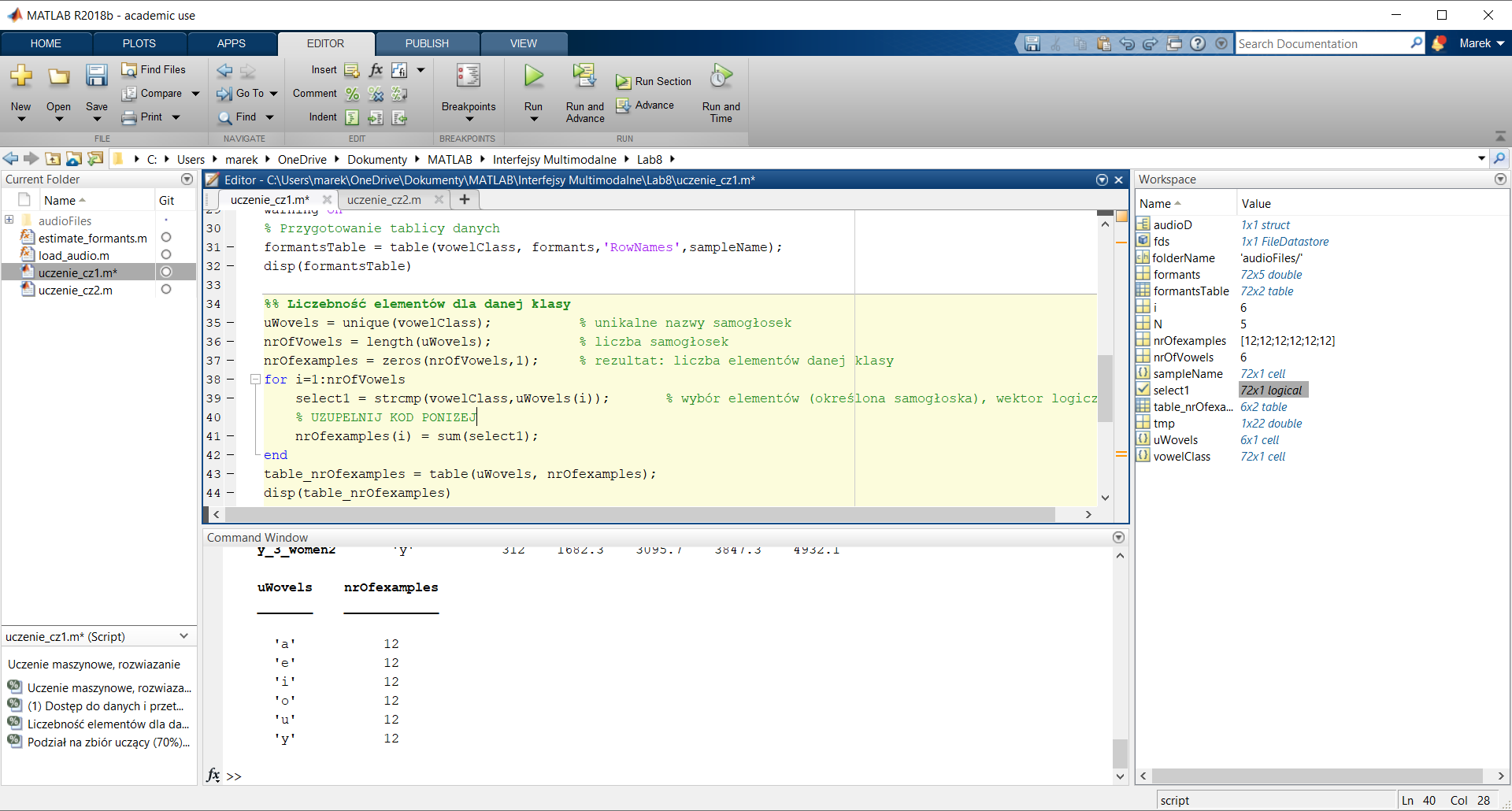
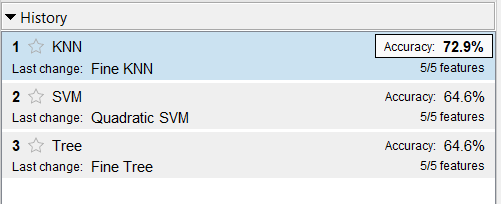
**Data:** 23.05.2020

**Imię i nazwisko:** Marek Matys

**Rezultaty**

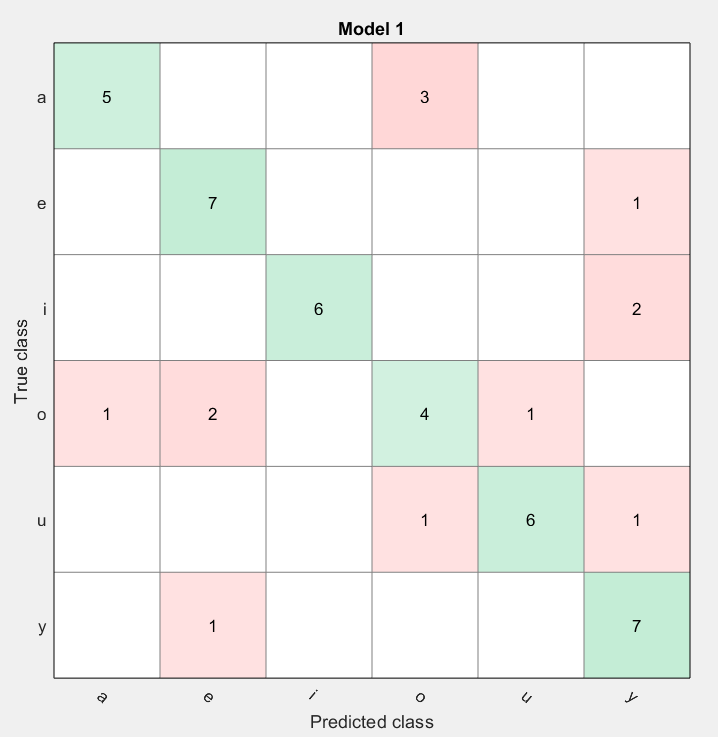
***Część I – Import i przygotowanie danych***

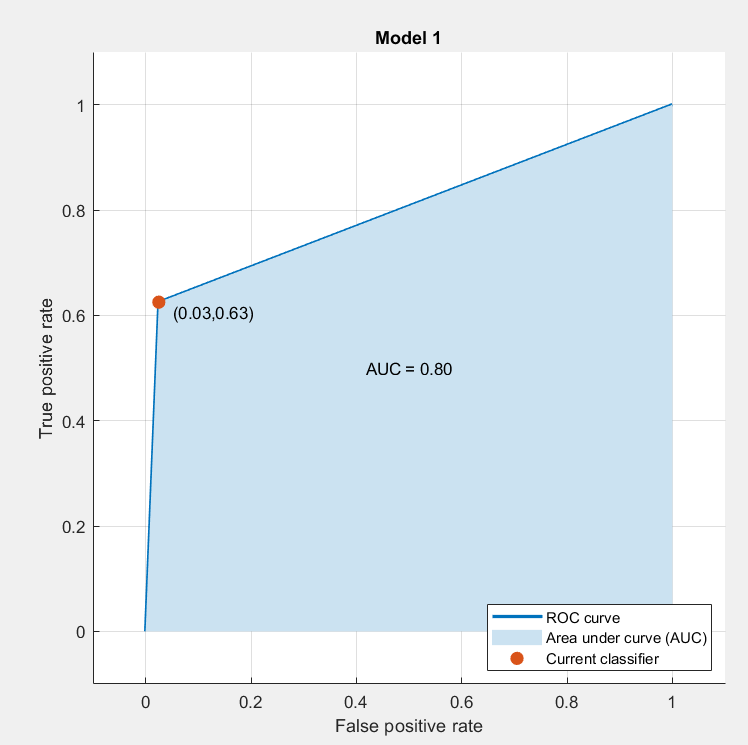
***Część II – budowa modelu (uczenie klasyfikatorów)***



Najlepszy wynik treningu uzyskał klasyfikator KNN

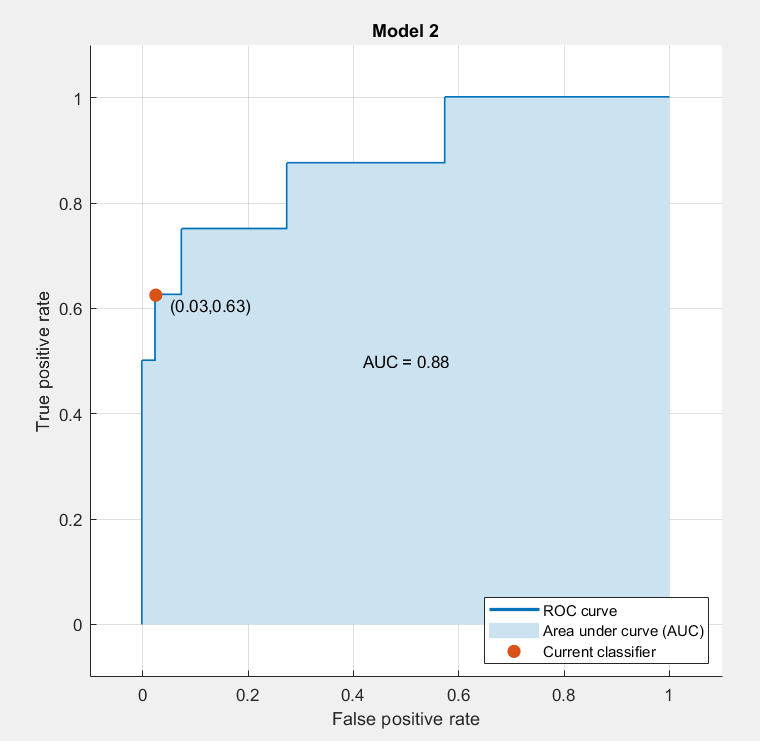
KNN



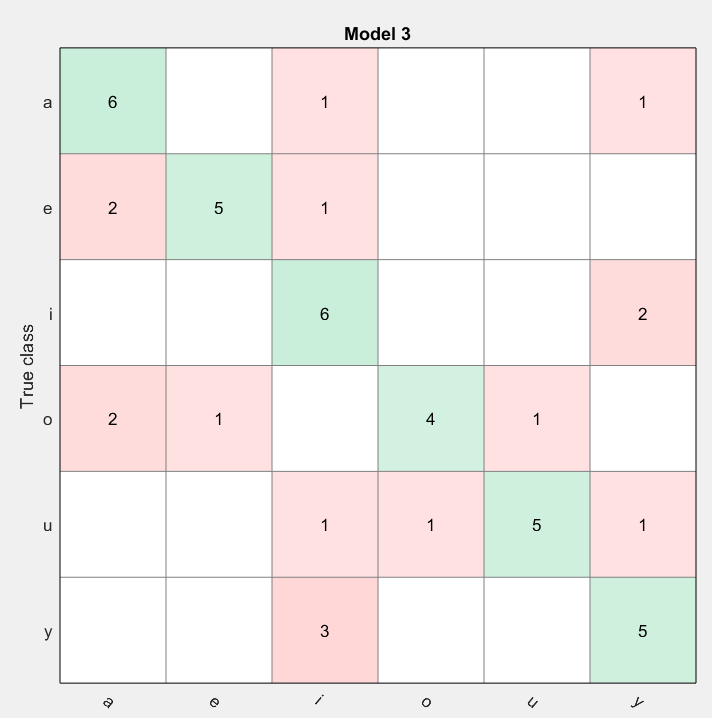


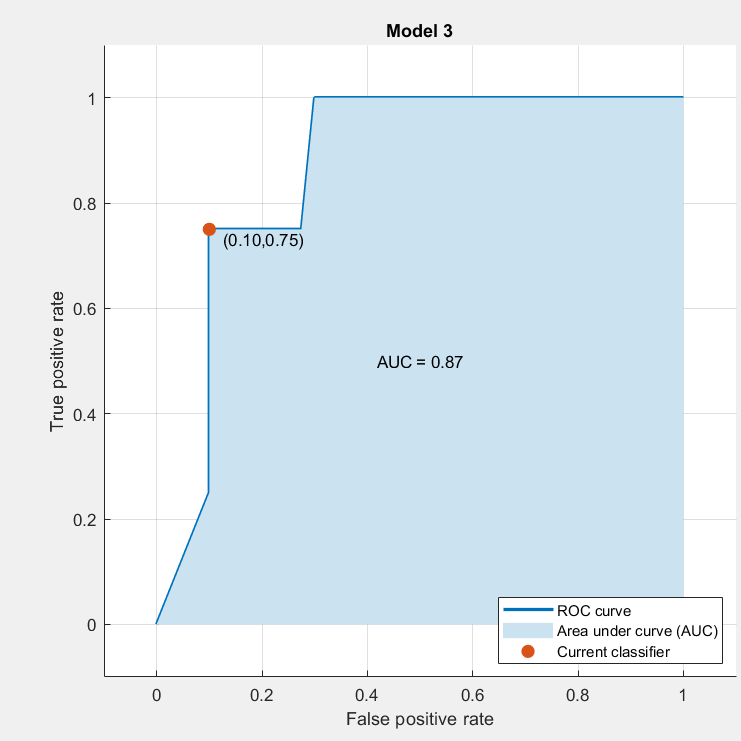
SVM



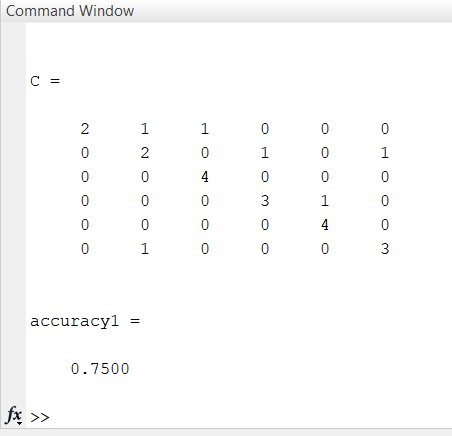


Tree





Wyniki na zbiorze testowym dla KNN



**Analiza i wnioski**

*Czy liczba wczytanych elementów klas jest podobna do siebie? Wyjaśnij dlaczego to ma znaczenie w kontekście uczenia maszynowego*

Liczba wczytanych elementów każdej z klas jest równa 12.

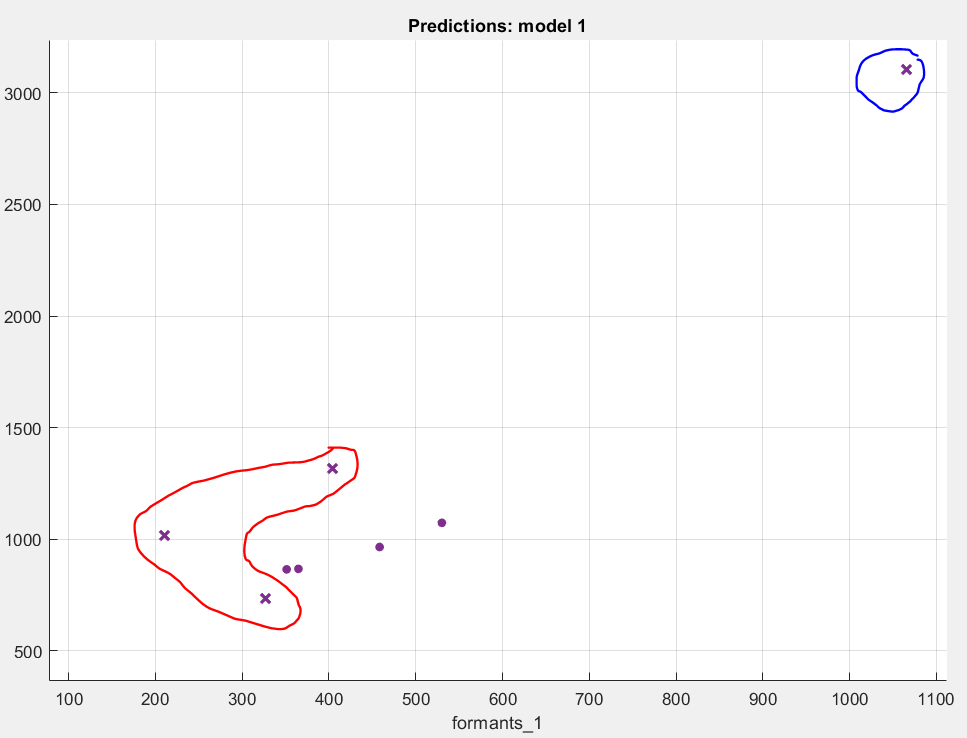
Tak, liczebność klas ma znaczenie. Gdyby jedna głoska np.: ‘a’ występowała więcej razy niż inne, to nasz model lepiej rozpoznawałby głoskę ‘a’ od innych głosek (mniej błędów w macierzy pomyłek). Jest to w pewnym sensie intuicyjne – im więcej model (mówiąc kolokwialnie) “coś przećwiczy” – tym będzie w tym lepszy.

Balansowanie danych (równanie liczebności klas) nie jest zawsze potrzebne. Zależy to w głównej mierze od problemu jaki chcemy rozwiązać i do czego dany model będzie używany.

*Wyjaśnij znaczenie elementów w macierzy pomyłek. Jakie klasy są ze sobą najczęściej mylone?*

W macierzy pomyłek oś pionowa odpowiada za faktyczną klasę, a pozioma za przewidzianą/odgadniętą przez model klasę.

Na przykładzie macierzy dla KNN (dla zbioru uczącego) – w kolumnie od głoski ‘o’ mamy: 3 w wierszu od ‘a’ (3 elementy zostały uznane za ‘o’ choć tak naprawdę to były ‘a’) oraz 1 w wierszu od ‘u’. Na wykresie widać te błędy:



Czerwoną linią zaznaczono te elementy które w faktycznie były głoskami ‘a’, a niebieską linią tę która była głoską ‘u’

Klasy najczęściej ze sobą mylone:

i&y, u&o, o&a (na podstawie macierzy pomyłek dla KNN, SVM i Tree).

**Pytania**

*Na czym polega walidacja krzyżowa?*

Walidacja krzyżowa, inaczej nazywana również kroswalidacją oraz sprawdzianem krzyżowym, polega na podziale próby statystycznej (zbioru wejściowego) na zbiór uczący i zbiór testowy. Prosta walidacja to najbardziej typowy rodzaj walidacji krzyżowej, w którym podział próby na wspomniane wyżej, rozłączne zbiory dokonuje się losowo. Zwykle zbiór testowy stanowi mniej niż 1/3 próby. Następnie model jest uczony na podstawie zbioru uczącego, a poprawność modelu jest sprawdzana na podstawie zbioru testowego.

Istnieje też tzw. k-krotna walidacja krzyżowa, gdzie dzieli się zbiór wejściowy na k podzbiorów, a następnie każdy z tych k zbiorów będzie służył za zbiór testowy, a pozostałe za zbiór uczący.

*Krótko wyjaśnij jak działa klasyfikator KNN.*

Jest to jedna z ważniejszych nieparametrycznych metod klasyfikacji. W tej metodzie klasyfikowany obiekt przydzielamy do tej klasy, do której należy większość z jego k najbliższych sąsiadów.

Przykładowo jeżeli mamy daną głoskę i zauważymy, że z 4 jej najbliższych sąsiadów 3 zostało zaklasyfikowanych jako ‘y’, a jeden jako ‘a’, to daną głoskę zaklasyfikujemy jako ‘y’.