Laboratorium 7 - Modele Bayesowskie

1. Cel laboratorium

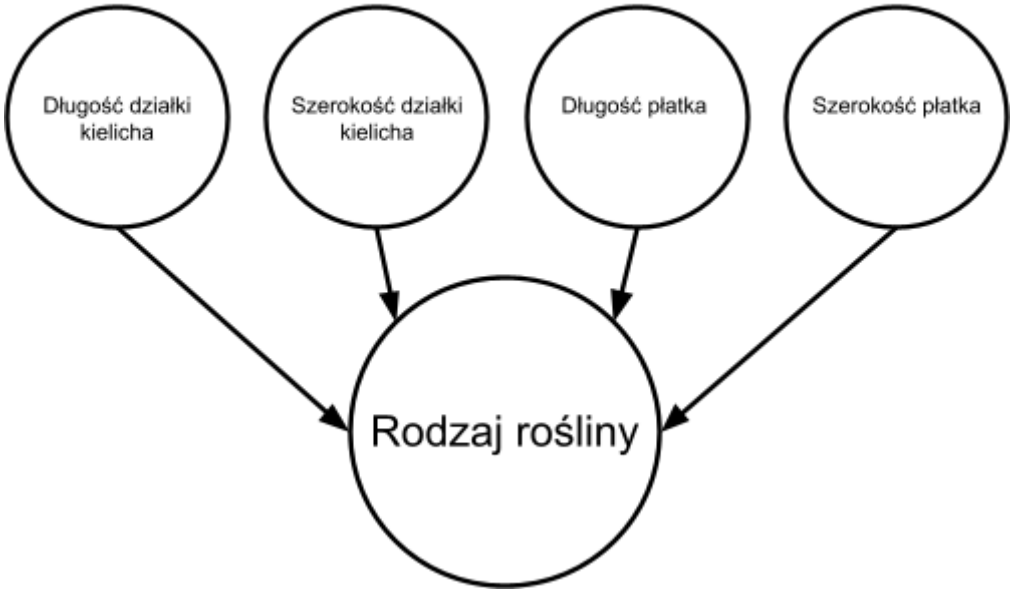
Celem laboratorium jest zaimplementowanie naiwnego klasyfikatora Bayesa oraz przetestowanie jego działania na zbiorze danych Iris

2. Zbiór danych Iris Data Set

Wymieniony wyżej zbiór danych przechowuje informację na temat określonych atrybutów rośliny:

- długość działki kielicha
- szerokość działki kielicha
- długość płatka
- szerokość płatka
- rodzaj rośliny (Setosa, Versicolor, Virginica)

Graficzny model zdarzenia wygląda następująco:



Mamy więc do czynienia ze wspólnym efektem, który ma cztery przyczyny. Oznacza to, że zdarzenia “przyczynowe” są niezależne do momentu zaobserwowania efektu.

3. Implementacja

Za działanie programu odpowiedzialne są trzy klasy:

3.1. Dataset

Klasa Dataset jako atrybut przyjmuje nazwę pliku csv, w którym zgromadzone są dane. Wszystkie dane odczytywane są przy pobieraniu jako wartości typu string. Metoda change_values zmienia parametry na wartości typu float, a seperate_data_by_title zmienia nazwy roślin na wartości typu int, co umożliwia prostsze kategoryzowanie danych.

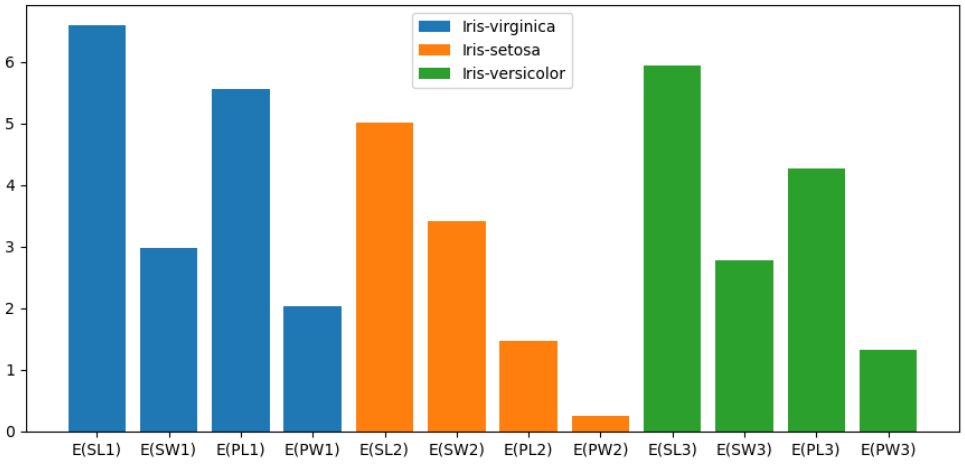
3.2. Splitter

Klasa Splitter jako atrybut przyjmuje obiekt klasy Dataset, który kategoryzuje ze względu na rodzaj rośliny i oblicza wartości:

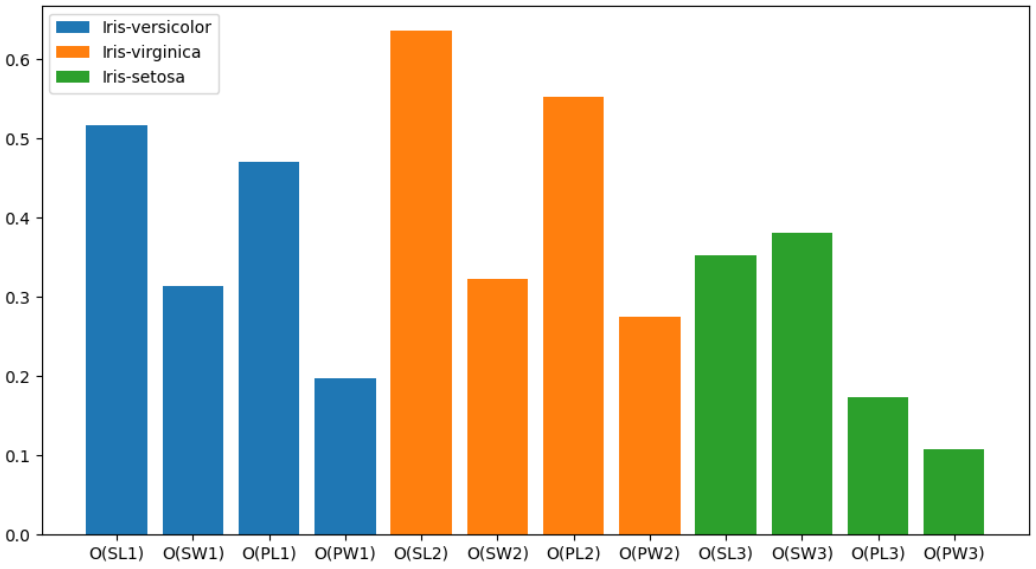
- wartość średnią z próby, ze wzoru: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- odchylenie standardowe, ze wzoru: $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2}$

dla każdego jej atrybutu

Poniżej przedstawione są wartości średnie dla poszczególnych parametrów każdego rodzaju rośliny pobrane z klasy Splitter metodą get_measures:



Oraz ich odchylenie standardowe:



3.3. NaiveBayes

Klasa NaiveBayes to naiwny klasyfikator Bayesowski. Metoda train_data_cretor dzieli zbiór danych na n_parts - równych podzbiorów, które potem wykorzystywane są jako część zbioru trenującego lub zbiór uczący.

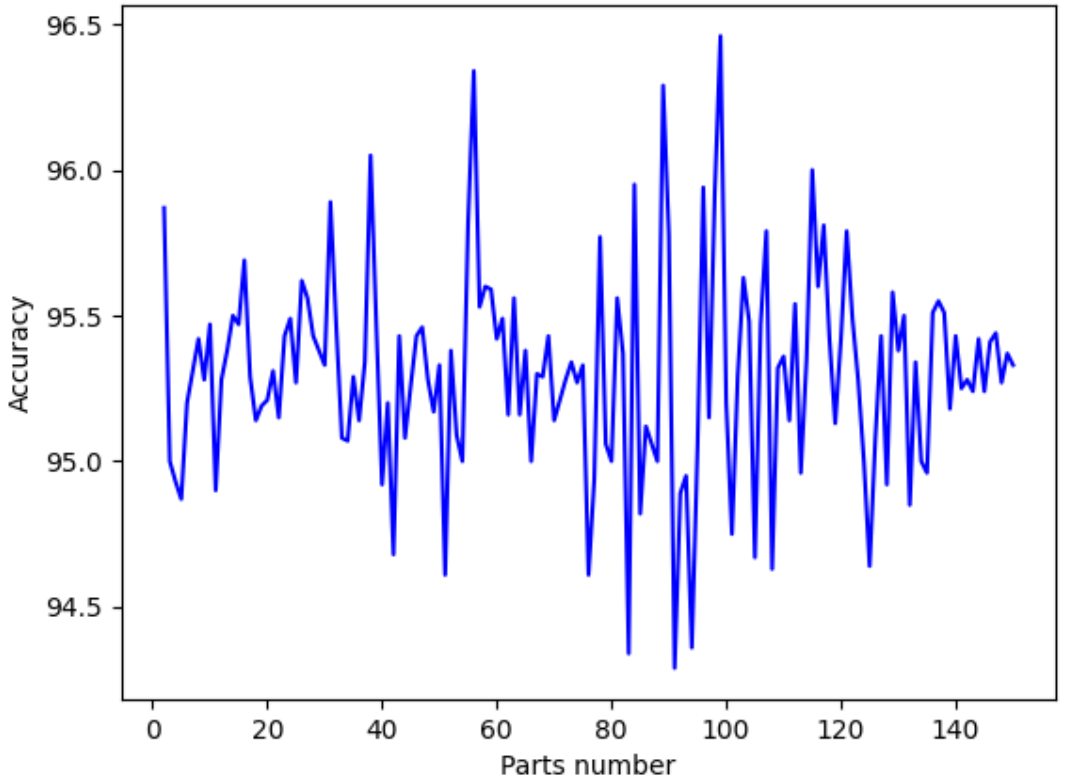
Początkowo prawdopodobieństwo jest równe liczbie obiektów wybranego rodzaju w zbiorze test podzielonego przez liczbę wszystkich obiektów. Następnie mnożymy je przez dystrybuantę rozkładu normalnego dla każdej wartości atrybutu (możemy tak zrobić, ponieważ są one niezależne), korzystając ze wzoru:

$$F(x) = \frac{e^{-\frac{(x-EX)^2}{2\sqrt{\sigma}}}}{\sqrt{2\pi\sigma}}$$

Algorytm wybiera rodzaj rośliny, która posiada największe prawdopodobieństwo

4. Testowanie

Wykres poniżej przedstawia średnią dokładność działania algorytmu w zależności od liczby n_parts:



Jak można zauważyć, dokładność jego działania waha się w przedziale od ~94,5 do ~96,5%, co wskazuje na wysoką trafność jego klasyfikacji

Należy dodać, że podany wykres może ulegać zmianie praktycznie z każdą próbą jego wykonania. Dokładność szacowania klasyfikatora zależy w dużej mierze od doboru danych przez metodę train_data_creator z klasy NaiveBayes. Dokładność nie zmieni się tylko w jednym przypadku, gdy n_parts = 150 (dokładność równa 95,33%), ponieważ w tym przypadku dobór zbiorów będzie zawsze taki sam.