Imię i nazwisko: Przemysław Ziaja
Grupa: 5a

Podstawy Sztucznej Inteligencji
Uczenie Nadzorowane – Sprawozdanie

Zadania: https://colab.research.google.com/drive/1foGRmOJIeEU0F2bV871rciHwP_YB4hW9
Należy utworzyć kopię i pracować na swoim koncie.
Jeśli są jakieś uwagi (błąd, niejasne), co do kodu/zadania – należy umieścić je przy pomocy funkcji komentarza w pliku udostępnionym powyżej.

Część 4:

1. Jaki jest rozmiar zbioru danych? Podaj wszystkie wymiary!
   `rozmiar datasetu (6, 7)`
   `rozmiar na dysku: 272 bytes`
2. Ile atrybutów występuje w zbiorze danych?
   `liczba atrybutów: 7`
3. Ile jest instancji pozytywnych (enjoy==yes) a ile negatywnych?
   `liczba atrybutów pozytywnych: 4`
4. Który z atrybutów najlepiej rozdziela dane względem enjoy?
   `najlepiej rozdziela dane: sky, maksymalny rozmiar podzbioru 3`
5. Ile elementów ze zbioru danych ma atrybut wilgotność ustawiony jako wysoki (humidity==high)? Jakie mają numery w zbiorze danych (liczymy od 0)?
   `liczba elementow z high humidity 4 indexy to Int64Index([1, 2, 3, 5], dtype='int64')`

Część 5:

1. Czym te zbiory danych się różnią?
   Mają różną liczbę atrybutów, większość atrybutów mają wspólną. Zminiona jest kolejność rekordów lub atrybuty class nie odpowiadają sobie w obu zbiorach.
2. Jaki typ ma seria danych vendor?
   Kolumna vendor ma typ object
3. Co się stało po użyciu funkcji pd.get_dumies()?
   Dane z kolumny vendor zostały zakodowane przy pomocy rzadkiej macierzy, macierz jest tworzona w ten sposób, że wiersze zostają tak jak były ale dla każdego unikalnego obiektu w kolumnie wendors jest tworzona osobna kolumna w macierzy rzadkiej. W ten sposób można zakodować przynależność do danego obiektu z pola vendor za pomocą 0 i 1.
4. Zamieść kod dla regresji liniowej – od momentu wczytania danych.

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import pandas as pd
cpu = pd.read_csv('cpu.csv')
vendor = pd.read_csv('cpu-vendor.csv')
cpu_train, cpu_test, cpu_class_train, cpu_class_test = train_test_split(
   cpu.drop(columns=['class']).values,cpu['class'].values)
vendor_train, vendor_test, vendor_class_train, vendor_class_test = train_test_split(
   vendor.drop(columns=['class','vendor']).values,vendor['class'].values)
cpu_reg = LinearRegression()
vendor_reg = LinearRegression()
cpu_reg.fit(cpu_train,cpu_class_train)
vendor_reg.fit(vendor_train,vendor_class_train)
print(f'Dokładność przewidywań na cpu {cpu_reg.score(cpu_test,cpu_class_test)}')
```

print(f'Dokładność przewidywań na vendor {vendor_reg.score(vendor_test,vendor_class_test)}')

5. W jaki sposób można radzić sobie z overfittingiem? Podaj przynajmniej 1 przykład.
   Zwiększyć ilość danych lub zmniejszyć liczbę parametrów modelu.

Część 6:

1. Wklej cały kod dla przykładu ze swimming.

```
x = pd.get_dummies(data.drop('enjoy', axis=1))
y = data.enjoy
clf_entropy = tree.DecisionTreeClassifier(criterion='entropy')
clf_gini = tree.DecisionTreeClassifier()
clf_entropy.fit(x,y)
clf_gini.fit(x,y)
columns = X.columns
tree_ent = tree.export_graphviz(clf_entropy, out_file=None, rounded=True, filled=True, feature_names=columns)
tree_gini = tree.export_graphviz(clf_gini, out_file=None, rounded=True, filled=True, feature_names=columns)
graph1 = pydotplus.graph_from_dot_data(tree_ent)
graph2 = pydotplus.graph_from_dot_data(tree_gini)
```

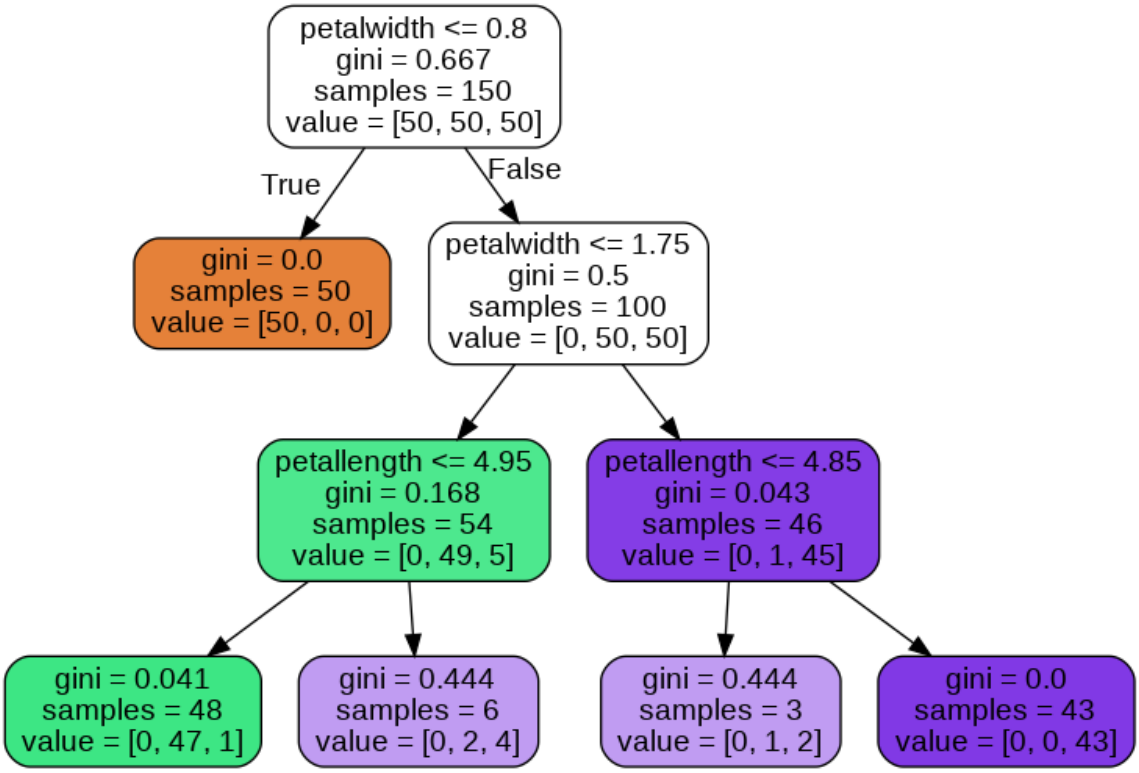
2. Czy w przypadku iris konieczne jest przekodowanie? Dlaczego?

W przypadku iris wymagane jest tylko przekodowanie klas(gatunków) na liczby. Algorytmy nie interpretują napisów jako klas. Natomiast pozostałe cechy są liczbami rzeczywistymi więc algorytm nie ma problemu z ich interpretacją.

3. Za co odpowiadają parametry `max_depth`, `min_samples_split`, `min_samples_leaf`?

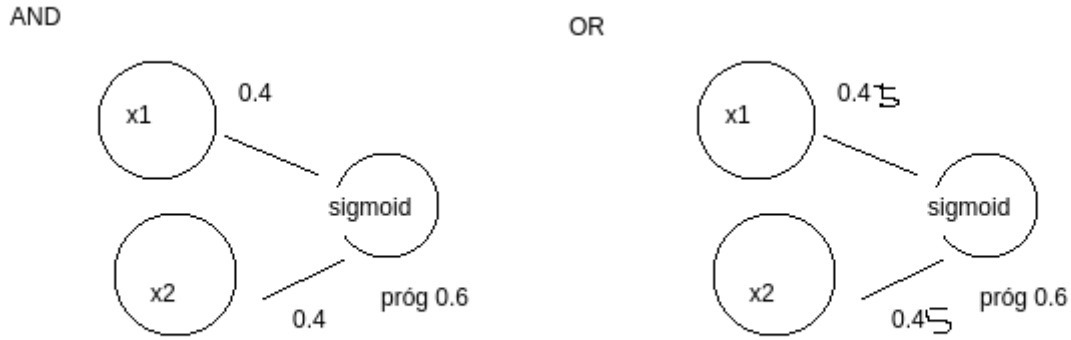
Max depth odpowiada za maksymalną głębokość drzewa, tj ile maksymalnie pytań zada przy ustalaniu klasy. min_samples_split określa liczbę lub procent wymagany do podziału węzła, tj jeżeli możemy podzielić node na 1 i 2, a min_samples_split jest ustawiony na 3 to węzeł nie zostanie podzielony. min_samples_leaf to parametr określający ile minimalnie trzeba przykładów aby węzeł mógł być liściem liściem

4. Zamieść graficzne reprezentacje drzew dla iris (przynajmniej 1 na parametr – podaj wartości).
   max_dept = 3; min_samples_split = 5, min_samples_leaf=2

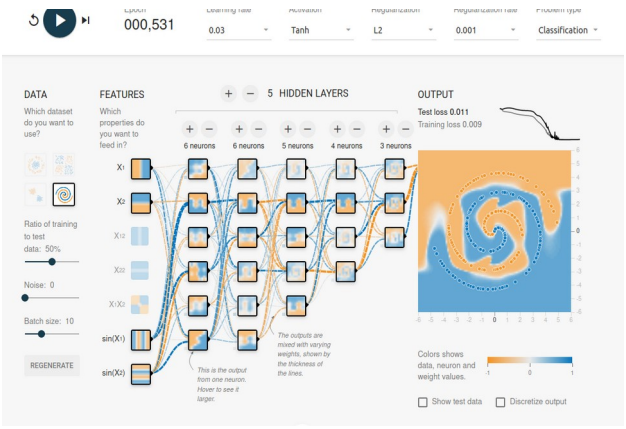


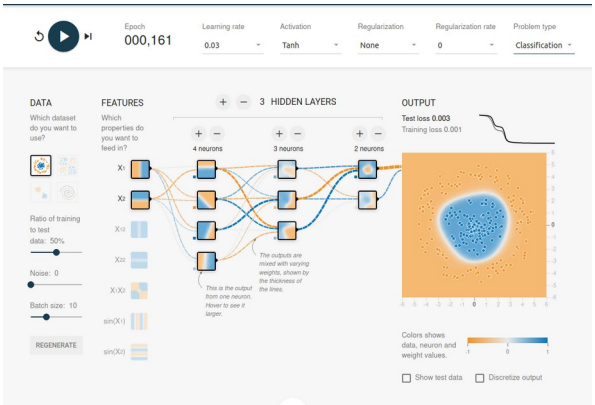
Część 7:

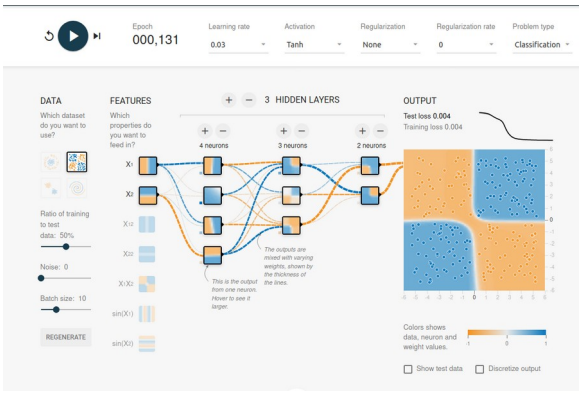
1. Wklej rysunki 2 neuronów dla AND i OR.

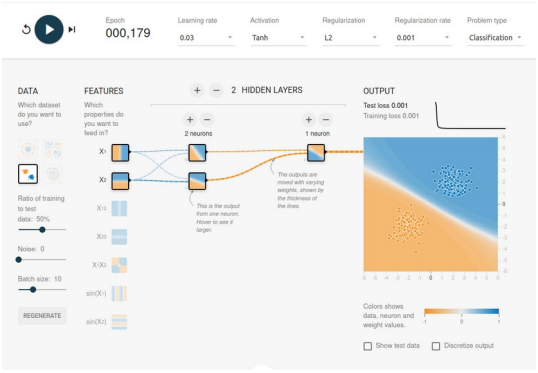


2. Wklej zrzuty ekranu wyuczonych sieci dla 4 datasetów ze wskazanej strony.





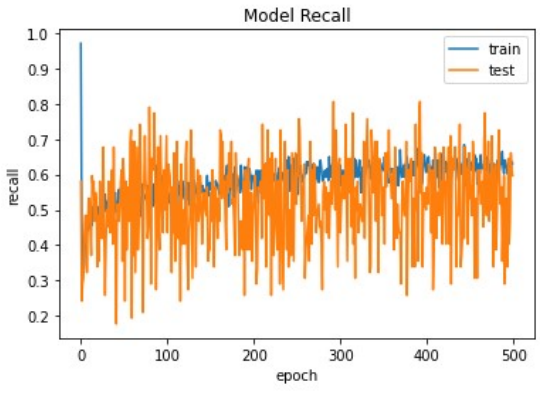


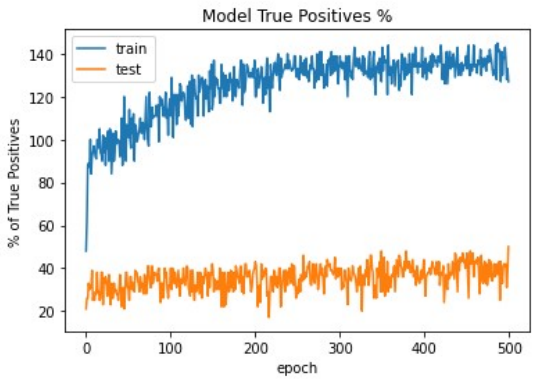


3. 3. Do czego służy funkcja `train_test_split()`? Po co ją wykorzystujemy?
   Służy do podziału danych na zbiór testowy i treningowy. Stosujemy po to, aby móc nauczyć i ocenić model.
4. Z ilu warstw składa się utworzona sieć? Która funkcja odpowiada za dodawanie warstw?
   2 warstwy ukryte wejście i wyjście.
5. Co oznacza batch_size? Czym jest liczba epok? Ile wynoszą te parametry w naszym przykładzie?
   Batch_size to liczba rekordów przetwarzanych jednocześnie, liczba epoch to liczba „okrążeń” zbioru treningowego podczas procesu uczenia, tj ile razy pokażemy cały zbiór treningowy modelowi.
6. Jakie funkcje aktywacji zastosowano? Czym jest funkcja aktywacji?
   W ukrytych Relu, w wyjściowej sigmoid. Funkcja aktywacji działa na wynik mnożenia wektor (poprzednia warstwa) razy macierz. Funkcja aktywacji przełamuje symetrię, co pozwala przy pomocy sieci neuronowej przybliżać każdą funkcję (tj nie tylko funkcje liniowe)

Część 8:

1. Uzasadnij poprzez zamieszczenie wykresu i odpowiedz na pytania:
   - Czy mamy do czynienia ze zjawiskiem overfittingu lub underfittingu?
     Tak, jest overfitting. Różnica pomiędzy zbiorem testowym i treningowym wynosi 5%.
   - Ile procent przypadków zostało poprawnie zaklasyfikowanych?
     Około 75%.
2. Czy to *dobry* wynik? Odnieś się do skuteczności modelu, gdyby ten *strzelał*.
   Zależy jak by strzelał. Jeżeli 1 lub 0 to tak. Jednakże nie jest to dobra miara, w społeczeństwie możemy mieć 10% ludzi którzy mają cukrzycę. Strzelając, że nikt nie ma cukrzycy mamy 90% dokładności, ale jest to fatalny wynik bo 10% nie jest leczonych. Trzeba wprowadzić inną metodę oceny modelu (F score)
3. Czym jest walidacja krzyżowa?
   Chodzi w niej o to, że mając zbiór danych dzielimy go i niektóre części wykorzystujemy do nauki/wnioskowania a inne do sprawdzenia przewidywań.
4. Zastosuj inne metryki (przynajmniej 2) i zaprezentuj rezultaty.





Informacja dodatkowa:
Inne elementy mogą, ale nie muszą, pojawić się w sprawozdaniu.