

# Projekt zaliczeniowy

## Bootcamp Data Science ING

# sages

Warunkiem uzyskania certyfikatu pozytywnego ukończenia kursu jest zrealizowanie projektu końcowego. Niniejszy dokument określa zasady zaliczania projektu.

Warunki ogólne:

1. Tematem projektu musi być jeden wybrany problem z listy tematów zamieszczonej w dalszej części tego dokumentu.
2. Projekt realizowany jest samodzielnie.
3. Forma przekazania projektu do oceny: **projekt musi być umieszczony w repozytorium (publicznym) na githubie**. Projekt musi zawierać plik w formacie .ipynb (notebook Jupyter'owy), w którym znajdować się będzie cała realizacja projektu lub główna jego część przedstawiająca kluczowe elementy (dopuszczalne są ewentualne pliki pomocnicze pomagające zachować przejrzystość pliku głównego). Dane mogą, ale nie muszą znajdować się w repozytorium. Inne formy realizacji projektu nie są dopuszczalne - w szczególności **przesłanie plików z rozwiązaniem nie jest traktowane jako przekazanie rozwiązania**. Link do repozytorium z projektem należy przesłać na adres: **n.ryciak@sages.com.pl**
4. Termin oddania projektów: **14 dni od ostatniego dnia zajęć** grupy, w której uczestniczył uczestnik. Projekt oczywiście można oddać wcześniej.
5. Aby projekt został zaliczony:
  - a. musi być poprawny pod względem merytorycznym (oznacza to np.: poprawną metodologicznie ocenę modeli czy rozsądne przygotowanie danych bez ewidentnych "błędów w sztuce"),
  - b. musi spełniać wymogi określone dla danego tematu,
  - c. musi mieć czytelną, przejrzystą formę umożliwiającą osobie sprawdzającej zrozumienie jego przebiegu. W szczególności **niezbędne są komentarze i objaśnienia**, które wyjaśniają poszczególne kroki. **Projekty nieestetyczne/trudne w odbiorze/bez objaśnień słownych będą zwracane do poprawki.**
6. W przypadku niezaliczenia projektu za pierwszym podejściem, projekt zwracany jest do poprawki i można projekt przesłać ponownie w ciągu 14 dni od otrzymania informacji o niezaliczeniu pierwszego podejścia. Dalszej możliwości poprawy nie przewiduje się.
7. Wszelkie pytania dotyczące projektu należy kierować na adres **n.ryciak@sages.com.pl**

---

# Lista tematów

## Temat 1

Przewidywanie ceny domów. Celem projektu jest zastosowanie modeli regresji do przewidzenia ceny domu o podanych cechach.

Link do danych: [web.stanford.edu/class/stats191/data/ames2000\\_NAfix.csv](http://web.stanford.edu/class/stats191/data/ames2000_NAfix.csv)

Dokładny opis danych: <http://web.stanford.edu/class/stats191/data/amesdoc.txt>

Zmienna celu(cena domu): SalePrice

Wymogi:

- Musi zostać wykorzystany model regresji liniowej.
- Muszą zostać wykorzystane przynajmniej dwie inne metody regresji: regresja Ridge, regresja Lasso, drzewo regresyjne, las losowy regresyjny, XGBoost lub inne).
- Trzeba uwzględnić optymalizację modeli/pipelinów.
- Nie usuwamy żadnych obserwacji - braki danych uzupełniamy w jakiś sposób.
- Wszelkie nieoczywiste przekształcenia danych należy opatrzyć uzasadnieniem dlaczego dokonujemy tego przekształcenia (np. robimy wykres i uzasadniamy wzięci logarytmu ze zmiennej tym, że rozkład jest skośny).
- Musi pojawić się ocena graficzna predykcji modeli (np. wykresy wartości przewidywanych od prawdziwych czy wykresy reziduiów).
- Projekt musi być zakończony przejrzystym porównaniem przetestowanych rozwiązań (modeli/pipelinów) **w postaci tabeli**, gdzie jeden wiersz opisuje jedno rozwiązanie i jego wynik.

---

## Temat 2

Przewidywanie wzięcia pożyczki. Celem projektu jest zastosowanie modeli klasyfikacji (binarnej) do przewidzenia czy pożyczka zostanie udzielona danemu klientowi z danymi parametrami wniosku.

Link do danych: <https://raw.githubusercontent.com/saimadhu-polamuri/DataHackthon3X/master/dataSet/Train.csv>

Opis danych (najdokładniejszy jaki istnieje): <https://discuss.analyticsvidhya.com/t/hackathon-3-x-predict-customer-worth-for-happy-customer-bank/3802>

Zmienna celu (wyplacenie pożyczki): Disbursed

Wymogi:

- Zmienna LoggedIn nie może być uwzględniona w modelowaniu - należy ją od razu wyrzucić ze zbioru.
- Nie usuwamy żadnych obserwacji - braki danych uzupełniamy w jakiś sposób.
- Wszelkie nieoczywiste przekształcenia danych należy opatrzyć uzasadnieniem dlaczego dokonujemy tego przekształcenia (np. robimy wykres i uzasadniamy wzięci logarytmu ze zmiennej tym, że rozkład jest skośny).
- Należy wykorzystać przynajmniej 3 metody klasyfikacji.
- Należy uwzględnić optymalizację modeli/pipelinów.
- Projekt musi być zakończony przejrzystym porównaniem przetestowanych rozwiązań (modeli/pipelinów) **w postaci tabeli**, gdzie jeden wiersz opisuje jedno rozwiązanie i jego wynik.

---

## Temat 3

Klasyfikacja wydźwięku twittów. Celem projektu jest zastosowanie modeli klasyfikacji (binarnej) do rozpoznawania wydźwięku (pozytywny lub negatywny) twittów.

Link do danych: <https://archive.ics.uci.edu/ml/machine-learning-databases/00331/sentiment%20labelled%20sentences.zip>

Opis danych: <https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences>

Dane są podzielona na trzy pliki amazon, yelp oraz imdb - należy je połączyć i traktować jako jeden zbiór.

Dopuszczalne jest również użycie dowolnego innego zbioru tekstów z określonym wydźwiękiem. (np. twittów z serwisu kaggle).

Wymogi:

- W rozwiązaniu musi pojawić się "czyszczenie" tekstu.
- Należy wykorzystać metody reprezentacji tekstu: macierz licznosci słów, macierz TfIdf, redukcja wymiaru przy użyciu SVD (ewentualnie również inne)
- Należy wykorzystać przynajmniej 3 metody klasyfikacji.
- Należy uwzględnić optymalizację modeli/pipelinów.
- Projekt musi być zakończony przejrzystym porównaniem przetestowanych rozwiązań (modeli/pipelinów) **w postaci tabeli**, gdzie jeden wiersz opisuje jedno rozwiązanie i jego wynik.

---

## Temat 4

Klasyfikacja obrazów. Celem projektu jest zastosowanie algorytmów do klasyfikacji do problemu rozpoznawania rasy psów przedstawionych na zdjęciu.

Link do danych: [vision.stanford.edu/aditya86/ImageNetDogs/images.tar](http://vision.stanford.edu/aditya86/ImageNetDogs/images.tar)

Opis danych: <http://vision.stanford.edu/aditya86/ImageNetDogs/>

Dane zawierają zdjęcia psów 120 różnych ras. Jeżeli to będzie powodować problemy z czasem obliczeń można sobie ten zbiór zmniejszyć - np. wziąć zdjęcia tylko 10-ciu ras i na takim zbiorze pracować.

Wymogi:

- Jeżeli projekt będzie wykorzystywał algorytmy klasycznego uczenia maszynowego, to
  - należy porównać działanie modeli na danych surowych oraz danych w jakiś sposób przekształconych/cechach wygenerowanych przez nas
  - należy wykorzystać przynajmniej 3 metody klasyfikacji.
  - należy wówczas również uwzględnić optymalizację modeli/pipelinów.
- Jeżeli projekt będzie wykorzystywał konwolucyjne sieci neuronowe, to trzeba rozpatrzyć kilka różnych struktur/wariantów sieci.
- Projekt musi być zakończony przejrzystym porównaniem przetestowanych rozwiązań (modeli/pipelinów) **w postaci tabeli**, gdzie jeden wiersz opisuje jedno rozwiązanie i jego wynik.

