Projekt końcowy

Państwa zadaniem jest zbudowanie, ocena i dostrojenie modeli klasyfikatorów (predykcyjnych), w oparciu o jeden ze zbiorów danych, wybrany z poniższej listy:

- HR Analytics: Job Change of Data Scientists (https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists)
- Airline Passenger Satisfaction (https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction)
- Rain in Australia (https://www.kaggle.com/jsphyg/weather-dataset-rattle-package) – proszę przeanalizować dane dla jednej, wybranej lokalizacji
- Churn Modelling (https://www.kaggle.com/shrutimechlearn/churn-modelling)
- Income Classification (https://www.kaggle.com/lodetomasi1995/income-classification)
- Banking Dataset - Marketing Targets (https://www.kaggle.com/prakharrathi25/banking-dataset-marketing-targets)

Proszę zwrócić uwagę, iż niektóre z powyższych przykładów zawierają dwa zbiory danych – treningowy oraz testowy, z czego ów drugi nie posiada zazwyczaj kolumny zmiennej celu, nie będzie więc mógł być użyty do uczenia modeli. Niemniej jednak, może on zostać wykorzystany np. do uzupełnienia brakujących wartości w zbiorze treningowym oraz do weryfikacji skuteczności zbudowanych modeli.

Modele te, w oparciu o zmienne, wybrane spośród predyktorów (atrybutów predykcyjnych) mają wyznaczyć wartość zmiennej celu. Należy także zbadać wypracowane modele pod kątem tzw. wyjaśnialności oraz analizy, jak poszczególne zmienne objaśniające wpływają na zmienną celu – te cztery etapy proszę przeprowadzić analogicznie jak w kursie „Machine Learning Basics" (https://academy.dataiku.com/machine-learning-basics) z Dataiku Academy oraz zaraportować zgodnie z szablonem (plik „Projekt – struktura raportu.pdf" na MS Teams).

By przyjrzeć się nieco bliżej danym, należy zacząć od tzw. eksploracyjnej analizy danych (EDA) – jej efektem powinny być m. in. podstawowe metryki dla każdej ze zmiennych, np. wartości minimalne, wartości maksymalne, średnie, mediany, analiza korelacji itd.

Państwa zadaniem jest także odnalezienie ewentualnych błędów w danych – wartości brakujących, błędnie zinterpretowanego rodzaju zmiennych, danych odstających (ang. *outliers*) itp. oraz ich wyczyszczenie. Następnie należy podjąć decyzję, które atrybuty wybrać do analizy (pod kątem ich istotności dla przewidywań modelu), czy i jak uzupełnić brakujące dane[1], przyciąć dane odstające, dokonać normalizacji/transformacji atrybutów lub wytworzyć nowe (ang. *feature engineering*) itd.

Projekt proszę realizować w grupach 2-osobowych (ew. 3-osobowych – wtedy należy wybrać zbiór danych z większą liczbą atrybutów). Termin nadsyłania raportów (e-mail: abobyk@pjwstk.edu.pl) to 28 stycznia 2024 r.

[1] Można posłużyć się tu jedną z metod tzw. imputacji, zawartą w pakiecie *scikit-learn* (https://scikit-learn.org/stable/auto_examples/impute/plot_missing_values.html) – w szczególności dobre efekty daje metoda MICE (ang. *Multivariate Imputation by Chained Equations*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/), implementowana poprzez klasę sklearn.impute.IterativeImputer (https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html).