Dokumentacja projektu

2019/2020

Analiza zbioru danych "US Births(2018)"

Kierunek: Informatyka

Członkowie zespołu:
Aleksandra Mincberg
Maciej Knera
Adrian Kurkowski

Gliwice, 2019/2020

Spis treści

1 Wprowadzenie

1.1 Geneza projektu

Co roku amerykańskie Centrum Kontroli i Zapobiegania Chorobom (CDC) publikuje dane dotyczące urodzeń dzieci. Zazwyczaj jest to bardzo duży zbiór danych obejmujący dziesiątki badanych parametrów oraz miliony rekordów. Powstaje więc pytanie, czy na podstawie takich danych, za pomocą uczenia maszynowego można stworzyć modele predykcyjne, które w jakiś sposób wspierałyby pracę lekarzy i pielęgniarek związane z urodzeniami, a także potrafiłyby wskazać, które ciąże są szczególnie narażone.

1.2 Cel projektu

Celem projektu jest analiza zbioru danych oraz próba stworzenia modeli predykcyjnych dla następujących zagadnień:

- metoda narodzin dziecka
- wystąpienie ryzyka zagrożenia ciąży
- masa dziecka przy narodzinach

2 Założenia projektowe

2.1 Zespół projektowy

Aleksandra Mincberg - model przewidujący, czy ciąża będzie miała czynniki ryzyka
Maciej Knera - model przewidujący metodę narodzin dziecka
Adrian Kurkowski - model przewidujący wagę dziecka w czasie narodzin

2.2 Założenia techniczne

Do realizacji projektu wykorzystana zostanie platforma Jupyter Notebook oraz język Python3 wraz z narzędziami do analizy i wizualizacji danych oraz do tworzenia modeli predykcyjnych. Do kontroli wersji wykorzystany zostanie GIT.

2.3 Zbiór danych

Zbiór danych dotyczy urodzeń w Stanach Zjednoczonych w 2018 roku.
Składa się z 3,8 miliona rekordów i 55 kolumn. Jest dostępny do wglądu i pobrania na stronie https://www.kaggle.com/des137/us-births-2018.

2.4 Stos technologiczny

Python3
PyCharm Community Edition
Jupyter
pandas
numpy
sci-kit learn
GIT (kontrola wersji w projekcie)

2.5 Oczekiwane rezultaty projektu

Przeprowadzenie analizy danych w celu odkrycia ciekawych zależności powiązanych z narodzinami dziecka.
Stworzenie modeli predykcyjnych na podstawie dostępnych danych.

3 Realizacja projektu

3.1 Pobranie danych i wczytanie ich do DataFrame'a biblioteki pandas

Z początku pojawił się problem z wczytaniem danych do DataFrame'a ze względu na spory rozmiar zbioru. Okazało się, że korzystaliśmy z 32-bitowej wersji Pythona, która nie jest zoptymalizowana pod wczytywanie dużej ilości danych. Po zainstalowaniu wersji 64-bitowej wszystko działało bez problemu.

3.2 Obróbka danych

Ze względu na dużą ilość danych, zmieniamy typy danych, aby ograniczyć ilość pamięci potrzebnej do działania na zbiorze. Przykładowo, dane z zakresu 1 do 9 (jak kolumna ĄTTEND") wymagają jedynie uint8, ale kolumna zawierająca dane z zakresu 227 do 9999 (kolumna DBWT) wymaga już uint16. Następnie, wczytujemy dane do dataframe. Aby ułatwić pracę na kolumnie "płeć"(SEX), zmieniamy jej wartości z ('M', 'F') do (0, 1). Za pomocą regexa zamieniamy wartość wszystkich pustych pól na 99 oraz ustawiamy typ danych za pomocą "df.astype()". W dokumentacji zbioru możemy przeczytać, że nieznane wartości najczęściej wyrażają się za pomocą liczb 9, 99 lub 9999. Korzystając z tej informacji, usuwamy właśnie te rekordy, które zawierą wyżej wymienione liczby. Oczywiście sprawdzamy wcześniej czy można to zrobić bez utraty ważnych rekordów.

3.3 Eksploracja danych

Eksploracja danych składa się w głównej mierze z szeregu wykresów obrazujących dane. Szczegóły na ten temat zawarte są w samym projekcie.

3.4 Tworzenie modeli predykcji danych

Szczegóły na temat modeli, które stworzyliśmy, są zawarte w Notebooku wraz z komentarzami.

4 Wnioski

- *Spostrzeżenia*
  W trakcie realizacji projektu okazało się, że najtrudniejszy był wybór tematu projektu oraz same przygotowania. Analiza danych polega przede wszystkim na odpowiednim ich przygotowania, dobraniu metod analizy oraz prawidłowym zinterpretowaniu wyniku. Sama technologia, którą wykorzystaliśmy nie była skomplikowana i posłużyła jedynie jako narzędzie a nie kolejne wyzwanie.

- *Osiągnięcia*
  Każdy członek grupy sporo nauczył się o metodach uczenia maszynowego oraz interpretacji wyników. Mogliśmy sobie również przypomnieć jak pracować w Jupyter Notebook oraz podstawowym narzędziem systemu kontroli wersji jakim jest Github, jak również lepiej poznać narzędzia do analizy danych takie jak pandas czy numpy.

- *Potencjał rozwoju*
  Projekt był dość niewielki i w obecnej formie nie widzimy dla niego szerszego zastosowania. Nie oznacza to jednak, że nasza praca poszła na marne, ponieważ nauczyliśmy się wielu rzeczy, które można wykorzystać przy poważniejszych zastosowaniach. Wszystko zależy od doboru zbioru danych oraz celu jaki sobie wyznaczymy.