Analiza wpływu zastosowania wybranych technik przygotowania danych do analizy, na jakość analizy danych

Dariusz Litwiński

21 maja 2023

1 Problem Analizy Danych

1.1 Analiza Danych

Analiza danych to proces, w którym surowe dane są przekształcane w wiedzę i spostrzeżenia, na podstawie których można podejmować lepsze decyzje. [] Wewnątrz tego procesu można wyróżnić następujące fazy:

1.1.1 Pozyskiwanie: gromadzenie danych

Zanim analiza danych będzie możliwa należy pozyskać dane. Dane mogą mieć różne źródła: czujniki, ankiety, Mogą być różnie przechowywane: pliki csv, bazy danych

1.1.2 Przygotowanie: przetwarzanie danych

Aby móc w pełni korzystać ze zgromadzonych danych, należy je przygotować: odpowiednio sformatować, wykorzystać techniki preprocessingu. To głównie tą częścią analizy danych będziemy się zajmować w dalszej części pracy.

1.1.3 Analiza: modelowanie danych

Najczęsciej kiedy ktoś mówi "Analiza danych"to ma na myśli własnie tą część całego procesu, jakim jest Analiza Danych. To właśnie wykorzystywane są techniki trenowania sztucznej inteligencji, tworzone są klasyfikatory, które później możemy wykorzystywać do różnych zadań

1.1.4 Działanie: podejmowanie decyzji

Kiedy mamy już gotowe wyniki analizy, wtedy możemy je wykorzystać aby podjąć konkretne decyzje w prawdziwym świecie: Wykorzystać stworzony klasyfikator w diagnostyce chorób,

1.2 Problemy z Danymi

Bardzo często zebrane dane nie nadają się bezpośrednio do pracy z nimi. Należy najpierw wykonać szereg operacji aby pozbyć się następujących problemów:

1.2.1 Brakujące wartości

W danych mogą występować brakujące wartości, na przykład czujniki mogą różnić się między sobą ilością pobieranych parametrów, ankietowani mogą pozostawić niektóre pytania bez odpowiedzi. Brakujące wartości stanowią poważny problem, ponieważ model nie potrafi ich jednoznacznie zinterpretować, dlatego w trakcie przygotowania danych musimy podjąć decyzję, czy usunąć rekordy z brakującymi wartościami, przez co możemy znacznie zmniejszyć liczebność zbioru danych. Alternatywnym podejściem jest wypełnienie brakujących wartości. W miejsce braku może być wstawiona średnia, minimum, maksimum, lub też inna arbitralnie wybrana wartość. Pozornie wartości wstawiane w puste miejsca są kompletnie arbitralne, jednak bardzo często takie podejście skutkuje najlepszymi rezultatami, pod warunkiem że dobierzemy odpowiednią wartość do wstawiania.

1.2.2 Wartości odstające

W niektórych przypadkach w danych mogą pojawić się takie wartości, które wyraźnie odstają od reszty i nie wnoszą sobą zbyt wiele informacji w kontekście analizy danych. Co więcej, mogą one zaciemniać pozostałą część danych, maskując trendy bądź prowadząc do błędnych wniosków. Dlatego najlepszym podejściem jest wykrywanie oraz usuwanie wartości, które możemy uznać za odstające. Istnieją algorytmy pozwalające nam na odrzucenie wartości odstających.

1.2.3 Kolumny kategoryczne

Wiele z modeli może pracować jedynie na wartościach liczbowych, podczas kiedy w zbiorach danych możemy znaleźć nie tylko takie wartości, ale również kategoryczne. Rezygnując z analizy tych danych tracilibyśmy wiedzę, jaką można z nich pozyskać. Nie jest to jednak konieczne, gdyż istnieją sposoby, aby zamienić te dane na postać liczbową za pomocą kodowania

2 Metody Przygotowania Danych

Istnieje kilka najczęściej używanych metod przygotowania danych, które dzielą się na następujące grupy:

2.1 Uzupełnianie brakujących wartości

Najczęściej brakujące wartości w zbiorach danych uzupełnia się średnią lub najczęściej występującą wartością, jednak może zdarzyć się tak, że najlepszym roz-

wiązaniem jest uzupełnienie braków minimum, maksimum, zerem bądź inną arbitralnie wybraną wartością

2.2 Wykrywanie wartości odstających

Do wykrywania wartości odstających możemy wykorzystać manualne metody, ale również i algorytmy, które wykryją te wartości za nas. Do manualnych metod możemy zaliczyć: Wykrywanie za pomocą rozkładu normalnego, Z-score, IQR, wykrywanie za pomocą percentyli. Natomiast spośród automatycznych metod mamy do dyspozycji między innymi Las Izolacji lub Local Outlier Factor

2.3 Kodowanie wartości kategorycznych

Jeżeli znamy zależności między klasami i możemy je uporządkować, wtedy jesteśmy w stanie dokonać kodowania ręcznie, na przykład najmniejszą wartość dla edukacji podstawowej, a najwyższą dla edukacji wyższej. Natomiast jeżeli nie znamy tych zależności, możemy wykorzystać LabelEncoder, jednak on ponumeruje klasy w kolejności alfabetycnzej, co nie zawsze jest pożądanym rezultatem. Innym podejściem jest One-Hot Encoding, który dla każdej z klas tworzy osobną kolumnę z wartością logiczną opisującą, czy dany rekord należy do tej klasy. Powoduje to wygenerowanie sporej ilości kolumn, jednak mamy wtedy pewność, że nie stworzymy nowych zależności między klasami

2.4 Standaryzacja

Standaryzacja jest procesem, po zakończeniu którego zmienna ma średnią wartość oczekiwaną zero oraz odchylenie standardowe równe jeden, dzięki czemu zyskujemy większą przejrzystość w jej analizie. Bardziej wyraźne są skupienia wokół konkretnych wartości, jednak należy zadbać o to, aby przed standaryzacją pozbyć się wartości odstających, gdyż będą one miały negatywny wpływ na zmienną po standaryzacji

3 Środowisko wykonawcze

3.1 Specyfikacja sprzetowa

3.2 System i środowisko

3.3 Zbiory danych

3.3.1 League of Legends stats

3.3.2 Australian Rain Forecast

3.3.3 Titanic Survival

4 Wykonane Eksperymenty

4.1 League of Legends stats

4.2 Australian Rain Forecast

4.3 Titanic Survival

5 Wyniki Eksperymentów

5.1 League of Legends stats

5.2 Australian Rain Forecast

5.3 Titanic Survival

6 Podsumowanie

6.1 Średnie wyniki

6.2 Wnioski