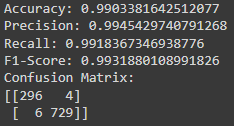
Do klasyfikowania wiadomości email, w modelu użyta została regresja logistyczna w połączeniu z wektoryzacją, która ułatwia modelowi analizę danych treningowych. Główną zaletą regresji logistycznej w tym problemie jest jego efektywność obliczeniowa. Inne modele mogłyby w celu wykonywania tego samego zadania opierać się na sieci neuronowej, bądź drzewie decyzyjnym, jednak uznałem, że tutaj najlepiej sprawdzi się prosty model regresji. Ma to też swoje odzwierciedlenie w wynikach na zbiorze treningowych. Aby ocenić model, korzystałem z 4 metryk podanych poniżej oraz z confusion matrix:



Jeżeli chodzi o formatowanie i podział danych, zdecydowałem się na usunięcie znaków interpunkcyjnych oraz liczb w preprocessingu. Uważam, że w kontekście wykrywania spamu, te elementy tekstu mają znikome znaczenie, a ich usunięcie wpływa pozytywnie na wymiarowość danych. 80% danych przypisałem do zbioru treningowego z celu zmniejszenia ryzyka na overfittingu. Miało to w tym przypadku pozytywny wpływ na wyniki modelu na zbiorze testowym. Ważnym aspektem podziału danych było zachowanie proporcji między ilością spamu, a hamu w zbiorze testowym i treningowym, jako, że w datasecie było znacząco mniej wiadomości typu spam.