Raport z analizy danych i modelu klasyfikacji spamu

1. Podsumowanie wyników analizy danych

Zbiór danych zawiera dwie główne kolumny:

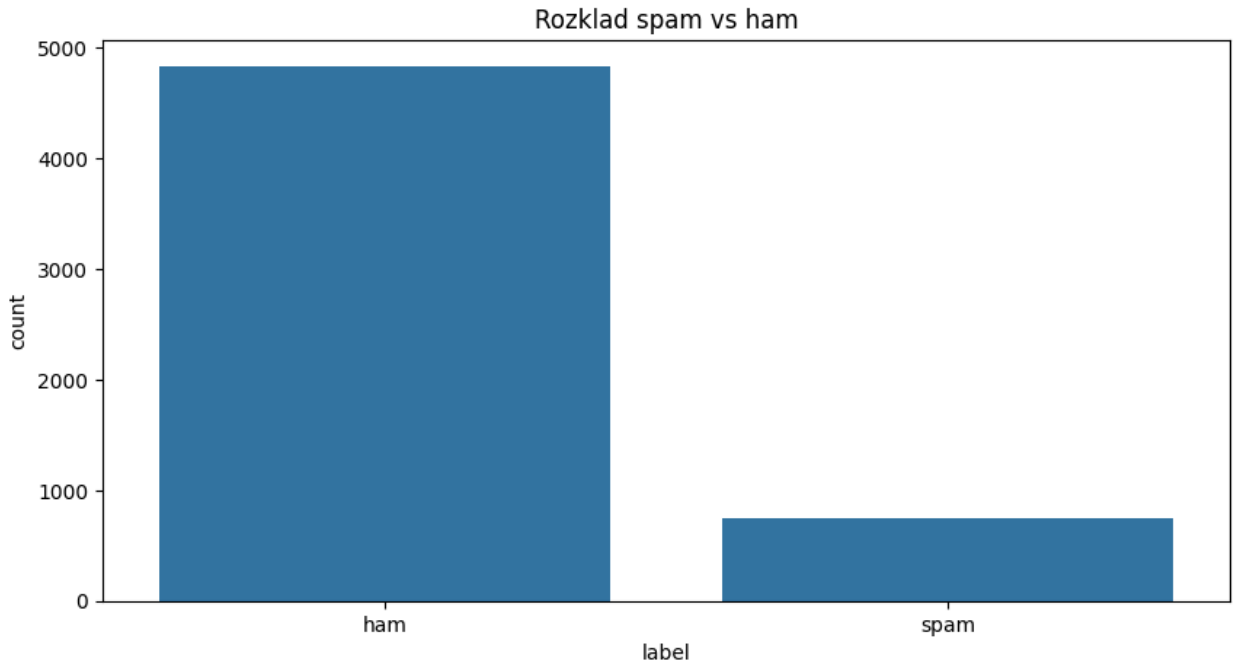
- **typ wiadomości** (ham/spam),
- **treść wiadomości**.

Dane zostały wstępnie przetworzone poprzez:

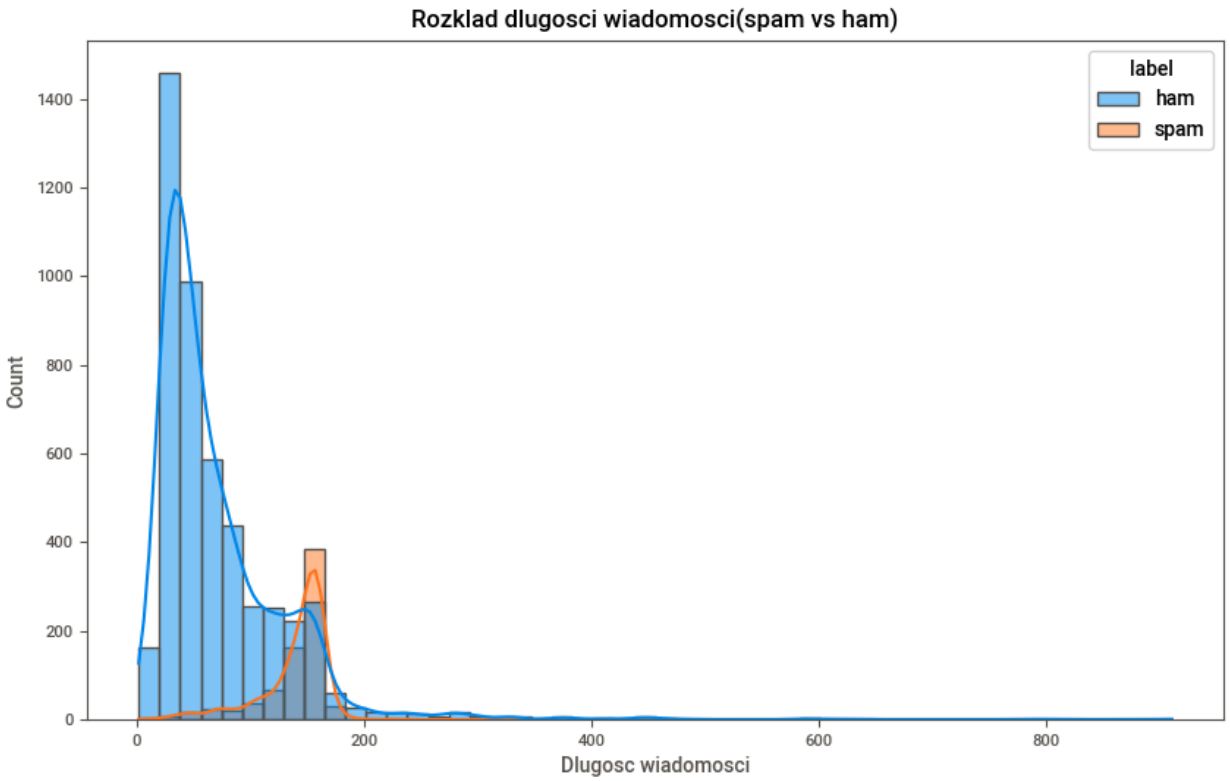
- Usunięcie kolumn o pustych nazwach i niepotrzebnych danych, które pierwotnie zaśmiecały dataset
- Przekształcenie zmiennej docelowej (label) na typ kategoryczny,
- Dalszą inżynierię cech, aby poprawić wydajność modelu. Nowe cechy obejmują między innymi: długości wiadomości, liczbę słów, liczbę znaków specjalnych i inne statystyki związane z tekstem wiadomości.

Po wstępnym przetwarzaniu, dane zostały podzielone na zbiór treningowy (70%) i testowy (30%)

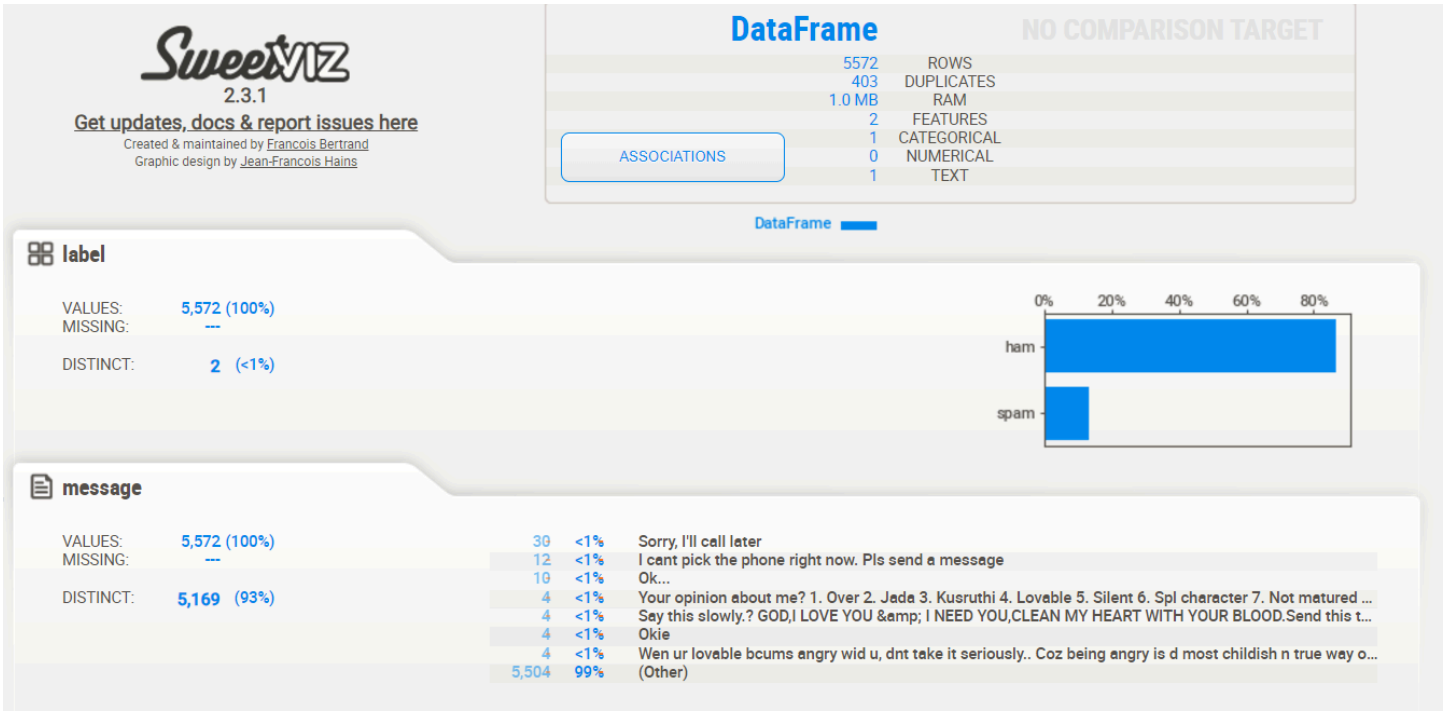
Wykres rozkładu spam vs ham (normalna wiadomość):



Histogram długości wiadomości:



Dane z analizy sweetwiz:



2. Wstępne wyniki modelu

Po przeprowadzeniu analizy z użyciem narzędzia H2O, najlepiej ocenionym modelem okazał się **GBM (Gradient Boosting Machine)**, który wykazał się dużą skutecznością w klasyfikacji spamu. Na podstawie wyników testów uzyskano następujące wyniki:

- **Dokładność (Accuracy): 0.9874**, co oznacza, że model poprawnie sklasyfikował około 98% wiadomości.
- **Średni błąd bezwzględny (MAE): 0.1310**, co wskazuje na stosunkowo mały błąd predykcji modelu.

Plan dalszego rozwoju modelu:

1. **Optymalizacja hiperparametrów**
2. **Wykorzystanie dodatkowych cech:** Można wzbogacić model o dodatkowe cechy, takie jak analiza częstotliwości słów, a także inne techniki przetwarzania tekstu, które mogą poprawić jakość klasyfikacji.