Raportu z projektu - „Diabetes”

# Cel projektu

Projekt ma na celu opracowanie skutecznego narzędzia do wczesnego wykrywania cukrzycy na podstawie analizy danych pacjentów.

# Dane

Dane pochodzą ze strony internetowej Kaggle: <https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data> Z danych tabelarycznych wygenerowano opisy tekstowe pacjentów w stylu:

"Female patient aged 54, has no hypertension, no heart disease, HbA1c level is 6.6 and blood glucose level is 80."

Dane te zostały następnie podzielone na zbiór uczący (80%) i testowy (20%).

# Modele

W projekcie porównano skuteczność trzech modeli NLP opartych na reprezentacji TF-IDF: **TF-IDF + Regresja logistyczna TF-IDF + Naiwny klasyfikator Bayesa (MultinomialNB) TF-IDF + MLP (sztuczna sieć neuronowa)**

# Ewaluacja

Do ewaluacji wykorzystano metryki *accuracy, precision, recall* i *F1-score*. Wyniki ewaluacji przedstawia poniższa tabelka:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| TF-IDF + Logistic Regression | 0.9462 | 0.9476 | 0.3917 | 0.5543 |
| TF-IDF + Naive Bayes | 0.9473 | 1.0000 | 0.3823 | 0.5532 |
| TF-IDF + MLP Neural Network | 0.9468 | 0.9667 | 0.3911 | 0.5569 |

# Wnioski

Wszystkie modele osiągnęły **bardzo zbliżone wyniki accuracy** (~0.946–0.947), co pokazuje, że reprezentacja tekstowa oparta na TF-IDF dobrze uchwyciła informacje zawarte w danych. **MLP (sieć neuronowa)** osiągnęła najlepszy F1-score (0.5569), co oznacza dobre wyważenie między precyzją i czułością. **Naive Bayes** uzyskał najwyższą precyzję (1.0), ale najniższy recall, co może sugerować nadmierną ostrożność (mało fałszywie pozytywnych, ale więcej przypadków pominiętych). Regresja logistyczna i sieć neuronowa były bardziej zbalansowane i lepiej radziły sobie z wykrywaniem przypadków cukrzycy.