# Raport projektu

Julia Czerniecka Wiktoria Gałdusińska Jerzy Grunwald
Maciej Kosierb Krzysztof Mizgała

10 lutego 2024

## Spis treści

# 1 Wstęp

Jeżeli podczas pracy w TEX-u napotkamy problem, dobrym instynktem jest wyszukanie go w internecie. W większości przypadków jesteśmy w stanie znaleźć odpowiedź na nasze pytanie m.in. na forach, takich jak TEX Stack Exchange. Jeśli nie uda nam się znaleźć rozwiązania napotkanego problemu, możemy również zamieścić pytanie na forum, tym samym licząc na pomoc ponad 260 tysięcy użytkowników. Jeśli jednak osoby zadające pytania nie otrzymają oczekiwanych odpowiedzi, strony te nie będą użyteczne. Na Stack Exchange, jedynym znakiem pokazującym, że osoba zadająca pytanie uzyskała pożądaną odpowiedź, jest akceptacja odpowiedzi. W tym projekcie badamy pytania i zaakceptowane odpowiedzi z forum TEX Stack Exchange, w celu zrozumienia, jakie czynniki wpływają na akceptację odpowiedzi.

# 2 Opis danych

Skorzystaliśmy z dostępnych zrzutów danych Stack Exchange, dotyczących forum TEX-a, udostępnionych przez archive.org. Dane składają się z trzech istotnych dla nas tabel:

1. `posts`, zawierającej wszystkie nieusunięte posty; kolumny, z których korzystamy to:
    - PostTypeId – typ wpisu (1 – pytanie);
    - AcceptedAnswerId – ID zaakceptowanej odpowiedzi (NULL – brak zaakceptowanej odpowiedzi);
    - CreationDate – data zadania pytania;
    - Title – tytuł pytania;
    - Body – treść pytania;
    - OwnerUserId – ID użytkownika zadającego pytanie;
    - Tags – lista tagów pytania;
2. `users`, która posiada informacje na temat użytkowników forum; kolumny, z których korzystamy to:
    - Id – ID użytkownika;
    - CreationDate – data założenia konta na forum;
    - Reputation – reputacja użytkownika (podana w postaci liczby całkowitej);
    - Views – liczba wyświetleń profilu;
    - UpVotes – liczba pozytywnie ocenionych postów przez użytkownika;
    - DownVotes – liczba negatywnie ocenionych postów przez użytkownika;
3. `tags` z informacjami na temat tagów; kolumny, z których korzystamy to:
    - TagName – nazwa tagu;
    - Count – liczba pytań z danym tagiem.

# 3 Metodyka

Do efektywnego przetwarzania ogromnej ilości danych zastosowaliśmy popularne środowisko obsługujące Big Data, jakim jest Apache Spark. Umożliwił on równoległe i rozproszone przetwarzanie danych.

Przeprowadzono zaawansowane przetwarzanie cech (ang. feature engineering) na danych wejściowych, aby lepiej odzwierciedlić istotne aspekty pytania, takie jak:

- TitleLength – długość tytułu pytania (liczba słów);
- BodyLength – długość treści pytania (liczba słów);
- NumberOfTags – liczba tagów;
- TagsCountMax – popularność tagów (liczba pytań, w których wystąpił najpopularniejszy z tagów);
- informacje o pytającym:
  - OwnerReputation – reputacja;
  - OwnerViews – wyświetlenia profilu;
  - OwnerUpVotes, OwnerDownVotes – liczba pozytywnie i negatywnie ocenionych postów przez użytkownika;
  - OwnerExperience – doświadczenie w momencie zadawania pytania (liczba dni od daty założenia konta na forum do momentu zadania pytania);
- Accepted – stan akceptacji odpowiedzi (czy pytania posiada zaakceptowaną odpowiedź).

Rozważamy jedynie wyżej wymienione cechy, ponieważ są to wszystkie istotne informacje dostępne w momencie zadania pytania.

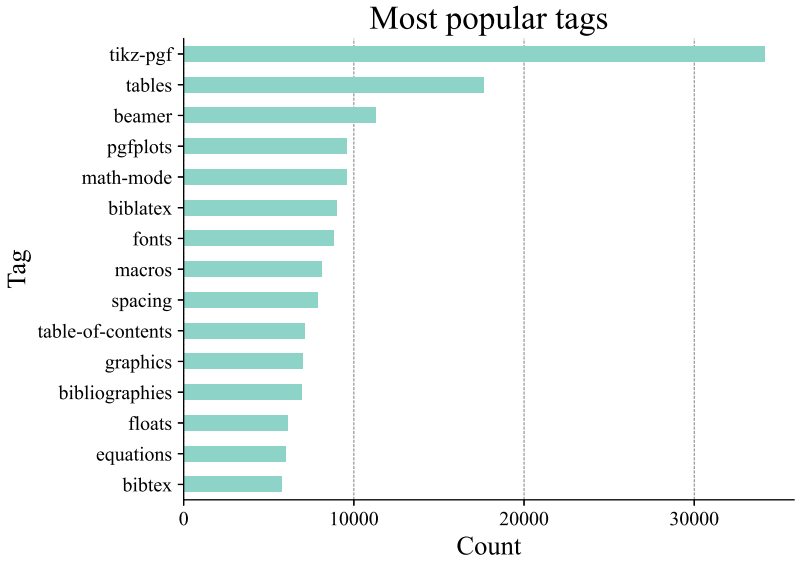
Tworzymy tabelę pomocniczą `questions`, która zawiera powyższe cechy dla każdego pytania w tabeli `posts`.

Implementujemy cztery modele predykcyjne, mające na celu przewidzenie, czy pytanie otrzyma zaakceptowaną odpowiedź, czy też nie. Do uczącej części danych dopasowano następujące modele:

- model regresji logistycznej;
- model lasów losowych;
- gradient boosting;
- model sieci neuronowych.

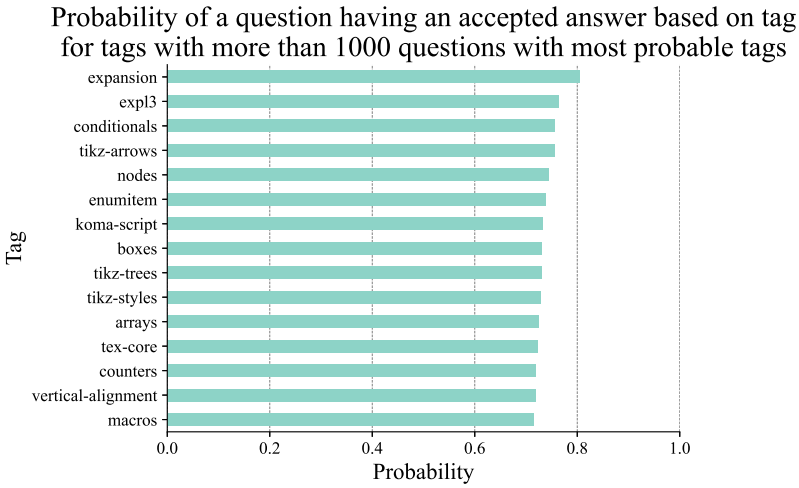
# 4 Analiza opisowa danych

Tabela `posts` zawiera 584 821 rekordów, z czego 255 804 stanowią pytania. Okazuje się, że 60,13% pytań ma zaakceptowaną odpowiedź oraz 47,16% spośród wszystkich odpowiedzi jest zaakceptowanych. Wysoki wskaźnik akceptacji może sugerować, że społeczność skutecznie zapewnia pomocne odpowiedzi.

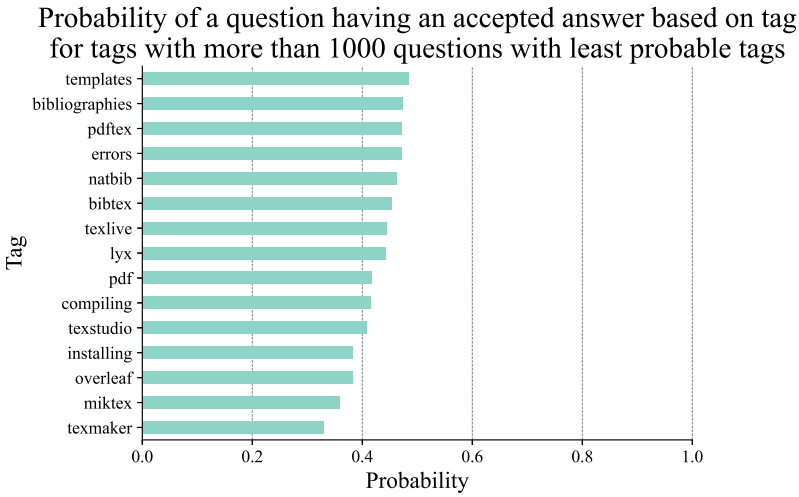



Rysunek 1: Najpopularniejsze tagi

Wykres ten przedstawia najczęściej wykorzystywane przez użytkowników forum tagi. Najbardziej popularnym tagiem jest „tikz-pgf", co wskazuje, że użytkownicy często zadają pytania związane z TikZ – pakietem LaTeX umożliwiającym programowe tworzenie grafiki.

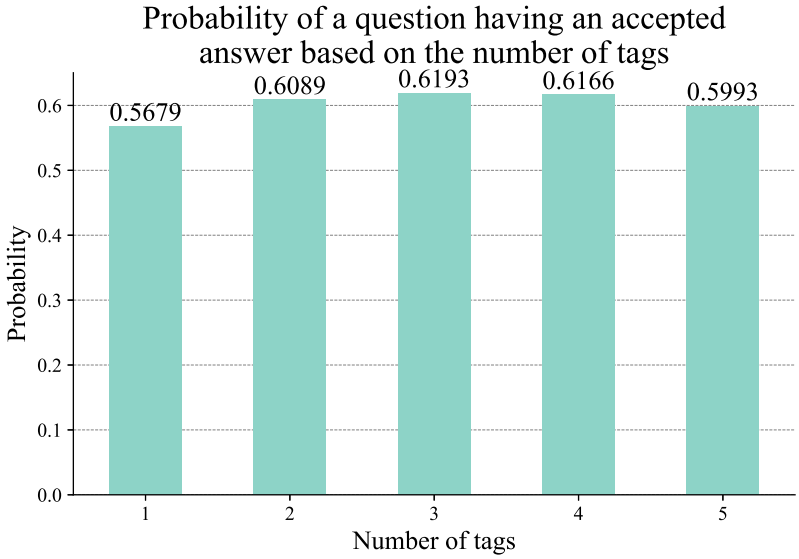



Rysunek 2: Najwyższe prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi dla tagów posiadających ponad 1 000 pytań




Rysunek 3: Najniższe prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi dla tagów posiadających ponad 1 000 pytań

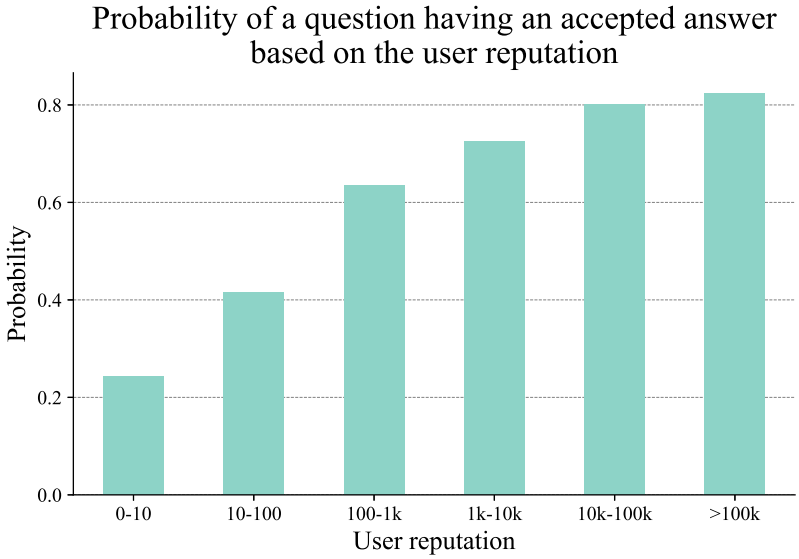
Wykresy te wskazują na to, że prawdopodobieństwo uzyskania zaakceptowanej odpowiedzi istotnie zależy od wykorzystanych tagów. Pytania z tagami „expansion" i „expl3" mają wyższe prawdopodobieństwo zaakceptowania odpowiedzi, co może oznaczać, że społeczność jest najbardziej aktywna wokół tematów z nimi powiązanych lub że te tematy są łatwiejsze do rozwiązania. Tagi „texmaker" i „miktex" posiadają najniższe prawdopodobieństwo uzyskania zaakceptowanej odpowiedzi, co może oznaczać, że problemy, z którymi zmagają się użytkownicy w tej kategorii są wymagające i społeczność nie potrafi ich rozwiązać.




Rysunek 4: Prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi w zależności od liczby tagów

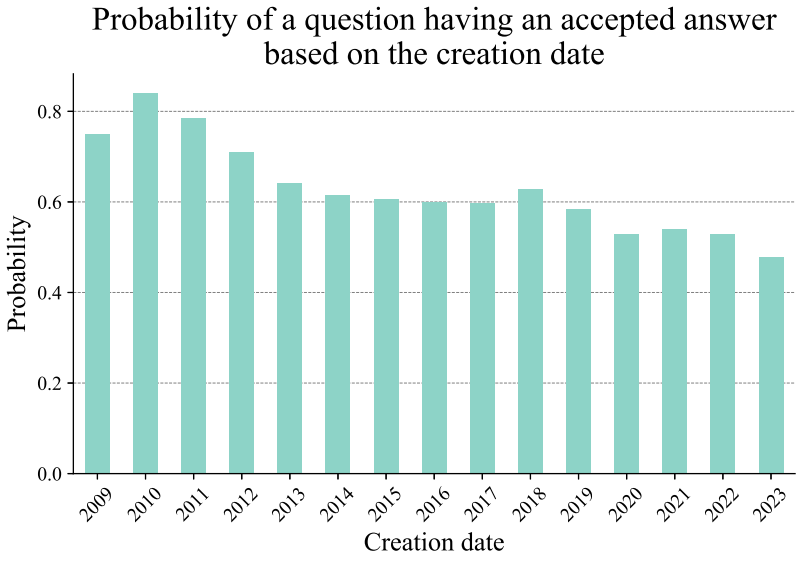
Powyższy wykres słupkowy pokazuje, że pytania zawierające od dwóch do pięciu tagów mają nieco większe szanse na otrzymanie zaakceptowanej odpowiedzi w porównaniu

do tych z jednym tagiem. Może to wskazywać, że liczba tagów nie wpływa w dużym stopniu na prawdopodobieństwo uzyskania zaakceptowanej odpowiedzi.



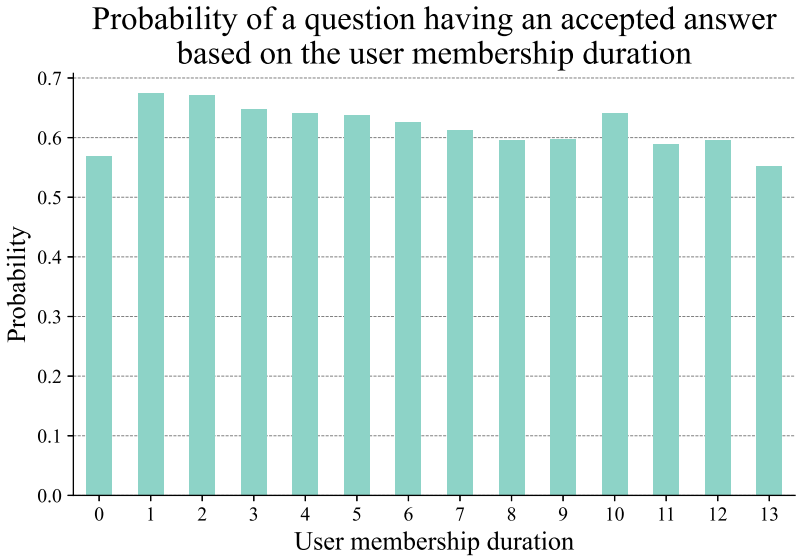

Rysunek 5: Prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi w zależności od reputacji zadającego pytanie

Wykres pokazuje pozytywną korelację między reputacją pytającego a prawdopodobieństwem zaakceptowania odpowiedzi na jego pytanie. W przypadku użytkowników z najwyższą reputacją, prawdopodobieństwo otrzymania odpowiedzi, która zostanie zaakceptowana, jest największe. Może to wynikać z większego doświadczenia w formułowaniu pytań na forum.



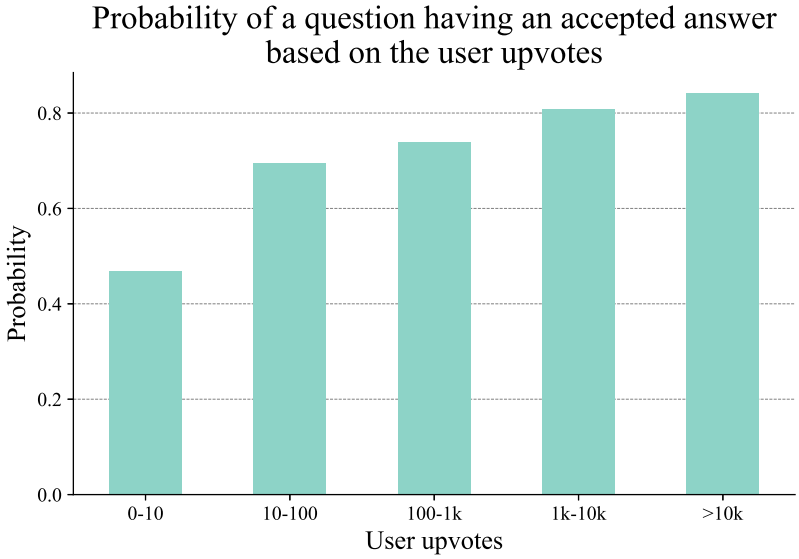

Rysunek 6: Prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi w zależności od daty publikacji pytania

Wykres wskazuje, że prawdopodobieństwo otrzymania zaakceptowanej odpowiedzi na pytanie wzrasta wraz z czasem. Im starsze pytanie, tym większa szansa, że posiada ono zaakceptowaną odpowiedź.



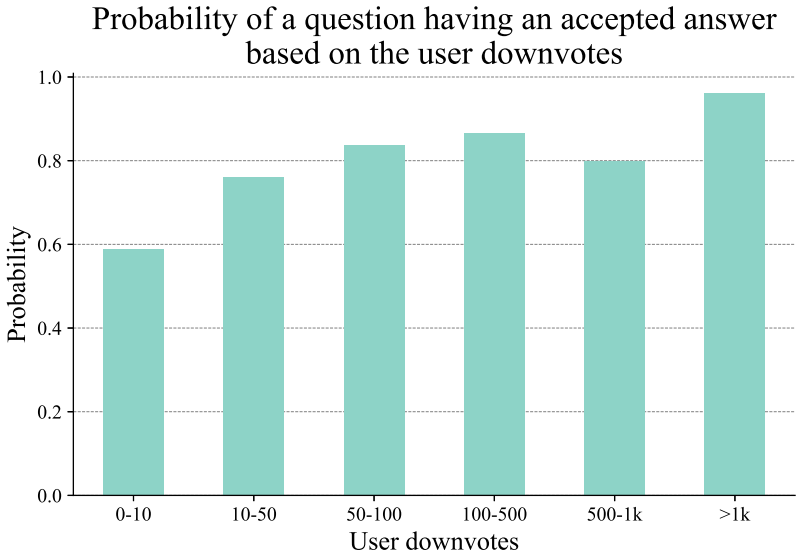

Rysunek 7: Prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi ze względu na długość członkostwa pytającego

Powyższy wykres pozwala wysnuć wniosek, że prawdopodobieństwo pozostaje stosunkowo stabilne, niezależnie od tego, jak długo użytkownik jest członkiem forum, co sugeruje, że długość członkostwa nie wpływa znacząco na szanse otrzymania zaakceptowanej odpowiedzi.




Rysunek 8: Prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi ze względu na liczbę pozytywnie ocenionych postów przez pytającego

Na podstawie wykresu można stwierdzić, że istnieje znacząca zależność prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi od liczby pozytywnie ocenionych postów przez pytającego.




Rysunek 9: Prawdopodobieństwa uzyskania zaakceptowanej odpowiedzi ze względu na liczbę negatywnie ocenionych postów przez pytającego

Warto wspomnieć, że użytkownicy mają możliwość ocenienia postów w negatywny sposób tylko wtedy, gdy przekroczą pewien próg reputacji. Ponadto, wystawiając negatywną opinię, tracą punkty reputacji. Z tego względu, użytkownicy rzadko decydują się na przyznawanie negatywnych ocen. Można spekulować, że jedynie doświadczeni i najbardziej aktywni członkowie forum wystawiają negatywne oceny, a zatem cecha OwnerDownVotes może być skorelowana z cechą OwnerReputation.

# 5 Analiza istotności cech

Tworzymy klasyfikator drzew losowych oparty o aspekty pytania opisane w sekcji 3. Dla wybranego ziarna generatora otrzymujemy następujące współczynniki istotności cech (w przybliżeniu do dwóch miejsc po przecinku):

- OwnerUpVotes: 45,80%;
- OwnerReputation: 33,41%;
- OwnerViews: 11,07%;
- OwnerExperience: 4,57%;
- TagsCountMax: 2,34%;
- OwnerDownVotes: 2,04%;

- BodyLength: 0,53%;
- NumberOfTags: 0,13%.
- TitleLength: 0,12%;

Można zaobserwować, że wnioski uzyskane w sekcji 4 są istotnie powiązane z powyższymi wynikami.

Aby uniknąć przeparametryzowania modeli, w dalszych analizach będziemy rozważać jedynie te cechy, których istotność wynosi przynajmniej 1%.

# 6 Wybór najlepszego modelu

Posiadając informacje na temat istotności cech możemy przystąpić do konstrukcji modeli. Dzielimy dane na część uczącą i testową w stosunku 80/20. Dla danych uczących dopasowujemy cztery modele. Następnie wyznaczamy predykcje uzyskania zaakceptowanej odpowiedzi dla danych ze zbioru testowego i badamy jakość modeli w oparciu o:

- dokładność dopasowania (Accuracy);
- wskaźnika Bookmaker Informedness (BM);
- współczynnik korelacji Matthewsa (MCC).

## 6.1 Wykorzystywane miary dokładności

### 6.1.1 Dokładność (Accuracy)

Jest to miara statystyczna wykorzystywana, w kontekście klasyfikacji, w uczeniu maszynowym i statystyce do mierzenia stopnia zgodności między przewidywaniami modelu a rzeczywistymi wartościami. Określa ona stosunek liczby poprawnych przewidywań do całkowitej liczby przypadków w zbiorze danych i przedstawia się wzorem

$$\text{Dokładność} = \frac{\text{Dobrze dopasowane przypadki}}{\text{Liczność populacji}}.$$

Dużą zaletą tej miary jest jej prostota i łatwość w interpretacji. Pomimo tego, ze względu na brak uwzględnienia nierównomiernego rozkładu klas, wysoka dokładność może być myląca, gdy jedna klasa jest znacznie liczniejsza niż pozostałe. Model może osiągnąć wysoką dokładność, zawsze przewidując dominującą klasę, co nie świadczy o jego rzeczywistej skuteczności. Z tego względu, do oceny modeli wykorzystamy również inne miary.

### 6.1.2 Bookmaker Informedness (BM)

Jest to miara oceniająca skuteczność klasyfikatora binarnego, która mierzy zdolność klasyfikatora do unikania fałszywych klasyfikacji, łącząc czułość i specyficzność w jedną metrykę. Wartość BM wynosi od 0 (brak wartości informacyjnej) do 1 (doskonała informacyjność), gdzie wyższe wartości wskazują na lepszą wydajność klasyfikatora. Miara ta przedstawia się wzorem

$$BM = \text{Czułość} + \text{Swoistość} - 1.$$

Zaletami tej miary są prostota interpretacji oraz zdolność do oceny ogólnej skuteczności klasyfikatora w unikaniu fałszywych klasyfikacji. Pomimo tego jest mniej przydatna w niezbalansowanych zbiorach danych, gdzie liczba przypadków w klasach jest istotnie różna.

### 6.1.3 Współczynnik korelacji Matthewsa (MCC)

Jest to miara służąca do oceny jakości klasyfikacji binarnych, uwzględniająca wszystkie cztery części macierzy pomyłek: prawdziwie pozytywne, fałszywie pozytywne, prawdziwie negatywne i fałszywie negatywne przypadki. Oferuje zbilansowaną ocenę klasyfikatorów, szczególnie w niezbalansowanych zbiorach danych. MCC przyjmuje wartości od $-1$ do 1, gdzie 1 oznacza dobre dopasowanie, 0 na poziomie klasyfikatora losowego, a $-1$ oznacza całkowite niedopasowanie. Miara ta przedstawia się wzorem

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

gdzie

- $TP$ – liczba prawdziwie pozytywnych wyników,
- $TN$ – liczba prawdziwie negatywnych wyników,
- $FP$ – liczba fałszywie pozytywnych wyników,
- $FN$ – liczba fałszywie negatywnych wyników.

## 6.2 Model regresji logistycznej

Model regresji logistycznej to matematyczny model służący do przewidywania prawdopodobieństwa przynależności obserwacji do jednej z dwóch klas. Regresja logistyczna używa funkcji logistycznej do przekształcania wyników regresji liniowej na przedział (0,1), co interpretowane jest jako prawdopodobieństwo.

Dopasowujemy model regresji logistycznej przyjmując maksymalną liczbę iteracji równą 100.

## 6.3 Model lasów losowych

Model lasów losowych to model zespołowy, który łączy wiele drzew decyzyjnych, a następnie agreguje wyniki. Każde drzewo w lesie losowym jest trenowane na innym podzbiorze danych, a także losowo wybiera podzbiór cech do rozważania podczas węzłów.

Dopasowujemy model lasów losowych przyjmując liczbę drzew równą 100.

## 6.4 Gradient boosting

Ogólna idea boostingu polega na trenowaniu modeli iteracyjnie, a każdy nowy model jest dostosowywany do błędów poprzedniego.

Dopasowujemy model gradient boosting przyjmując maksymalną liczbę iteracji równą 100.

## 6.5 Model sieci neuronowych

Model sieci neuronowych to system obliczeniowy składający się z warstw neuronów, z których każdy jest połączony z wieloma innymi neuronami, reprezentowanymi przez funkcje matematyczne. Model uczy się rozpoznawania wzorców, dokonywania klasyfikacji i przewidywania wyników.

Warstwa wejściowa przyjmuje dane wejściowe, które mają 6 cech. Mamy dwie warstwy ukryte oraz jedną warstwę wyjściową. Pierwsza warstwa ukryta ma 10 neuronów, a druga 8. Warstwa wyjściowa ma 2 neurony, ponieważ cecha Accepted jest typu binarnego. Parametry te zostały dobrane na bazie wielokrotnej walidacji.

## 6.6 Zestawienie dokładności dopasowania

| Model | Accuracy | BM | MCC |
|---|---|---|---|
| Regresja logistyczna | 60,06% | — | 0 |
| Lasy losowe | 70,27% | 0,4232 | 0,3603 |
| Gradient boosting | 70,92% | 0,4243 | 0,3747 |
| Sieci neuronowe | 66,38% | 0,2980 | 0,2706 |

Tabela 1: Tabela dokładności dopasowania modeli dla trzech wybranych miar w zaokrągleniu

Miara BM przyjmuje dla regresji liniowej wartość nieokreśloną, ponieważ przy wyznaczaniu czułości dzielimy przez zero. Podobnie w przypadku obliczania MCC dochodzi do dzielenia przez zero, jednakże w tej sytuacji przypisanie wartości zero jest uzasadnione. Pełny dowód można znaleźć w artykule.

# 7 Podsumowanie

Interpretując wyniki w tabeli 1 zauważamy, że gradient boosting wykazuje najlepszą ogólną wydajność wśród analizowanych modeli, zarówno pod względem dokładności,

jak i obu zastosowanych miar korelacji. Lasy losowe również prezentują się dobrze, będąc blisko wartości gradient boosting. Sieci neuronowe, choć lepsze od regresji logistycznej pod względem dokładności, nie osiągają tak dobrych wyników jak dwa wcześniej wspomniane modele w kategoriach BM i MCC. Regresja logistyczna ma najniższą dokładność i MCC równy 0, co sugeruje jej ograniczoną użyteczność w porównaniu z innymi modelami w tym zestawieniu.

# 8 Lista plików

Spis plików wchodzących w skład projektu:

- folder tex.stackexchange.com – tabele zawierające analizowane dane;
- statistics.ipynb – analiza opisowa danych;
- features.ipynb – analiza istotności cech;
- analysis.ipynb – ocena dokładności modeli predykcyjnych;
- requirements.txt – spis wykorzystywanych bibliotek;
- README.md – informacje potrzebne do inicjalizacji analiz.