Analiza czynników wpływających na szybkość adopcji zwierząt przy użyciu uczenia maszynowego.

1. Wstęp

1.1. **Cel projektu** Celem niniejszej pracy jest analiza danych pochodzących ze schronisk dla zwierząt oraz budowa modelu uczenia maszynowego, który pozwoli przewidzieć, jak szybko dane zwierzę znajdzie nowy dom. Problem bezdomności zwierząt jest istotnym wyzwaniem społecznym. Zrozumienie czynników wpływających na decyzje adopcyjne może pomóc schroniskom w lepszym promowaniu zwierząt „trudniej adopcyjnych" i optymalizacji procesów opiekuńczych.

1.2. **Przedmiot badań** Analiza została przeprowadzona na zbiorze danych udostępnionym w ramach konkursu "PetFinder.my Adoption Prediction". Zbiór zawiera informacje o tysiącach psów i kotów przebywających w schroniskach. Każdy rekord opisany jest szeregiem cech oraz etykietą określającą szybkość adopcji.

1.3. **Hipotezy badawcze.** W ramach projektu sformułowano szereg hipotez dotyczących czynników wpływających na proces adopcyjny, które poddano weryfikacji w fazie eksploracyjnej analizy danych (EDA):

- Różnice gatunkowe: Szybkość adopcji różni się w zależności od gatunku zwierzęcia – zakłada się, że istnieją wyraźne różnice w czasie oczekiwania na dom między psami a kotami.
- Wpływ wieku: Wiek zwierzęcia jest kluczowym wyznacznikiem szybkości adopcji (młodsze zwierzęta znajdują dom szybciej niż starsze).
- Preferencje rasowe: Zwierzęta rasowe oraz w typie rasy są adoptowane chętniej i szybciej niż mieszańce.
- Weryfikacja „Syndromu Czarnego Psa": Istnieje teoria o zjawisku polegającym na dłuższym czasie oczekiwania na adopcję zwierząt o czarnym umaszczeniu w porównaniu do zwierząt o innej maści.

2. Charakterystyka zbioru danych i narzędzia

2.1. Źródło danych. Dane wykorzystane w projekcie pochodzą z platformy Kaggle, z konkursu "PetFinder.my Adoption Prediction". Jest to rzeczywisty zbiór danych udostępniony przez organizację PetFinder.my. Główny plik z danymi (`train.csv`) zawiera rekordy dotyczące psów i kotów, opisane szeregiem cech.

2.2. Zmienna celu. Kluczowym elementem analizy jest zmienna AdoptionSpeed (Szybkość adopcji). Jest to zmienna, którą model ma za zadanie przewidzieć. Przyjmuje ona wartości w skali od 0 do 4:

- 0: Zwierzę adoptowane tego samego dnia (natychmiast).
- 1: Adopcja pomiędzy 1 a 7 dniem (1. tydzień).
- 2: Adopcja pomiędzy 8 a 30 dniem (1. miesiąc).
- 3: Adopcja pomiędzy 31 a 90 dniem (do 3 miesięcy).
- 4: Brak adopcji po 100 dniach.

2.3. Opis zmiennych objaśniających (Cechy) Każde zwierzę w zbiorze opisane jest za pomocą następujących atrybutów, które posłużyły do budowy modelu:

- Type: Gatunek zwierzęcia (1 = Pies, 2 = Kot).
- Age: Wiek zwierzęcia podany w miesiącach.
- Breed1 / Breed2: Rasa główna i dodatkowa (dla mieszańców).
- Gender: Płeć (1 = Samiec, 2 = Samica, 3 = Grupa mieszana).
- Color1 / Color2 / Color3: Dominujące kolory sierści.
- MaturitySize: Wielkość zwierzęcia po osiągnięciu dojrzałości (Mały, Średni, Duży, Bardzo duży).
- FurLength: Długość sierści (Krótka, Średnia, Długa).
- Zdrowie (Vaccinated, Dewormed, Sterilized, Health): Informacje o statusie medycznym – czy zwierzę jest szczepione, odrobaczone, wysterylizowane oraz ogólna ocena stanu zdrowia.
- Quantity: Liczba zwierząt w jednym ogłoszeniu (np. rodzeństwo).
- Fee: Wysokość opłaty adopcyjnej (0 oznacza adopcję darmową).
- VideoAmt / PhotoAmt: Liczba materiałów wideo i zdjęć dołączonych do ogłoszenia internetowego.

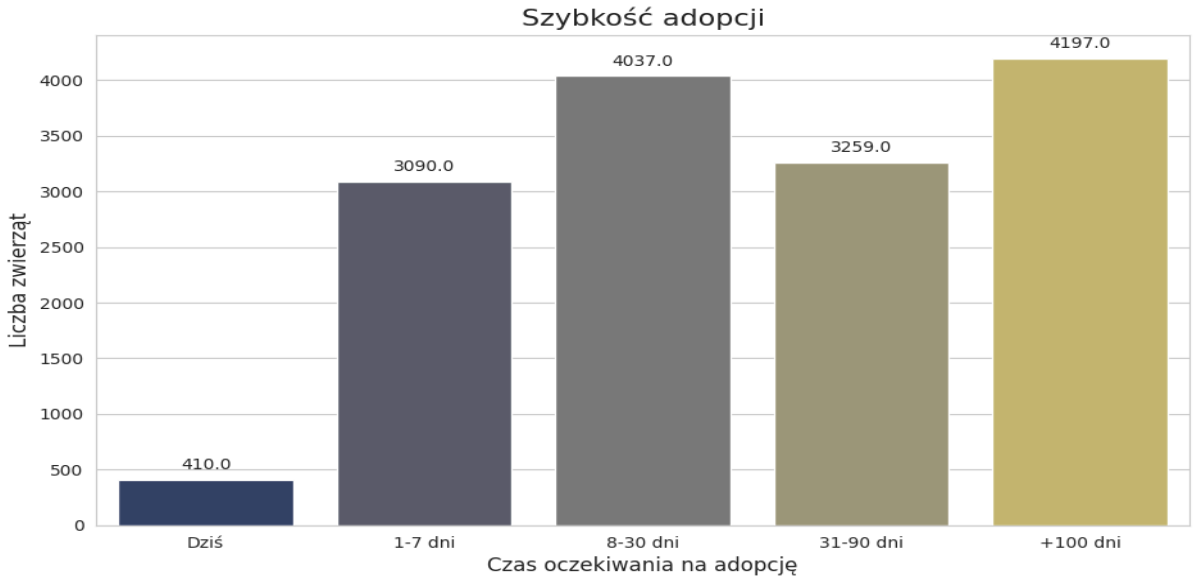
2.4. Środowisko pracy i narzędzia Projekt został zrealizowany w języku Python, z wykorzystaniem środowiska chmurowego Google Colab. Do analizy i modelowania wykorzystano następujące biblioteki:

- Pandas: Do wczytywania, czyszczenia i manipulacji danymi tabelarycznymi.
- Seaborn / Matplotlib: Do tworzenia wizualizacji i wykresów (EDA).
- Scikit-Learn: Do budowy modelu uczenia maszynowego (RandomForestClassifier), podziału zbioru na treningowy i testowy oraz oceny wyników.

3. Eksploracyjna Analiza Danych (EDA)

W tym rozdziale przedstawiono wizualizację kluczowych zależności w zbiorze danych. Celem tej analizy była weryfikacja postawionych wcześniej hipotez badawczych oraz zrozumienie struktury danych.

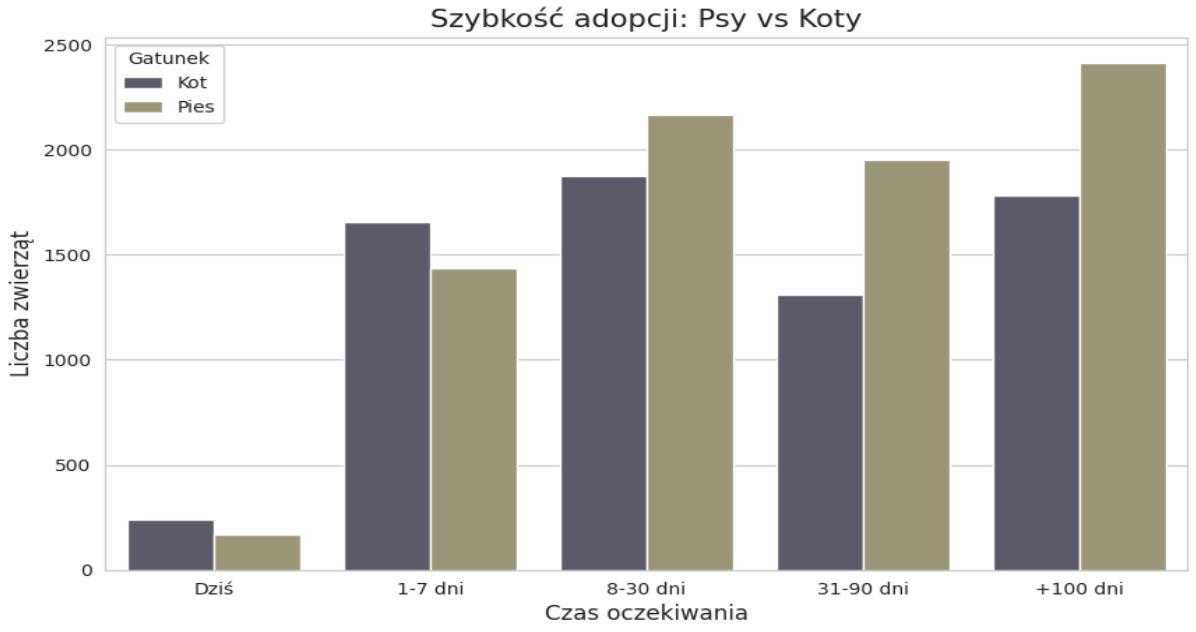
3.1. Rozkład zmiennej Pierwszym krokiem była analiza rozkładu zmiennej `AdoptionSpeed`. Pozwoliło to ocenić, czy klasy są zbalansowane, czy też występuje przewaga którejś z grup (np. bardzo szybkich lub bardzo wolnych adopcji).



Niestety, najliczniejszą grupę stanowią zwierzęta z kategorii 4, czyli te, które nie znalazły domu przez ponad 100 dni od momentu przyjęcia do schroniska. Stanowią one największe wyzwanie dla placówki.

Natomiast w grupie zwierząt, którym udało się znaleźć dom (kategorie 0-3), najwięcej adopcji ma miejsce w przedziale od 1 tygodnia do 3 miesięcy (kategorie 2 i 3). Adopcje błyskawiczne (tego samego dnia) są zjawiskiem marginalnym.

3.2. Wpływ gatunku i wieku na adopcję W pierwszej kolejności sprawdzono, czy gatunek zwierzęcia determinuje czas oczekiwania na adopcję. Poniższy wykres przedstawia liczbę adoptowanych psów i kotów w poszczególnych przedziałach czasowych.

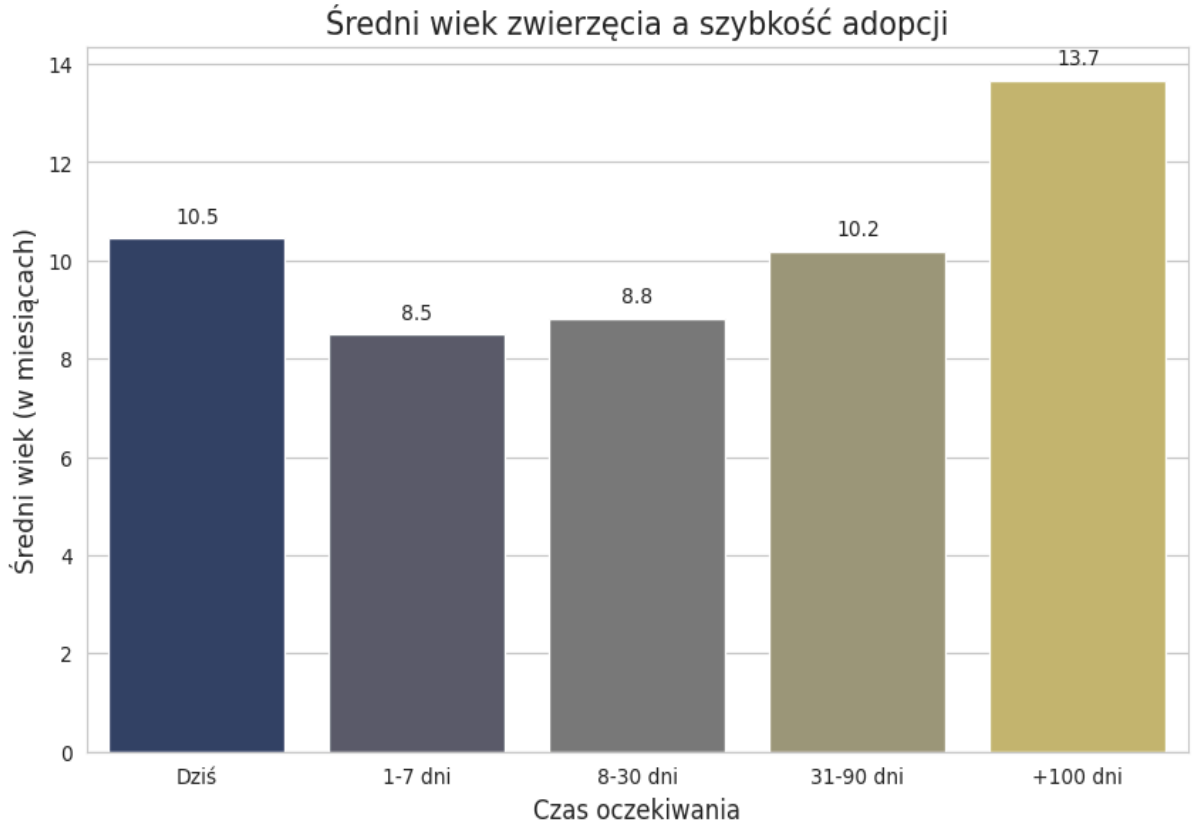


W kategoriach "Dziś" oraz "1-7 dni" zauważalna jest przewaga kotów. Sugeruje to, że decyzja o adopcji kota jest często podejmowana szybciej niż w przypadku psa.

Psy dominują w kategoriach średnioterminowych (od 8 do 90 dni), co może wynikać z konieczności dłuższego przygotowania się opiekunów do adopcji (spacery, warunki mieszkaniowe)..

Mimo tych początkowych różnic, dla obu gatunków najliczniejszą grupą pozostaje ostatnia kategoria ("+100 dni"). Oznacza to, że jeśli zwierzę nie znajdzie domu w pierwszych miesiącach, ryzyko długotrwałego pobytu w schronisku jest wysokie niezależnie od gatunku.

Drugim ważnym czynnikiem okazał się wiek. Aby sprawdzić, czy młodość gwarantuje szybszą adopcję, zestawiono średni wiek zwierząt w każdej grupie czasowej.

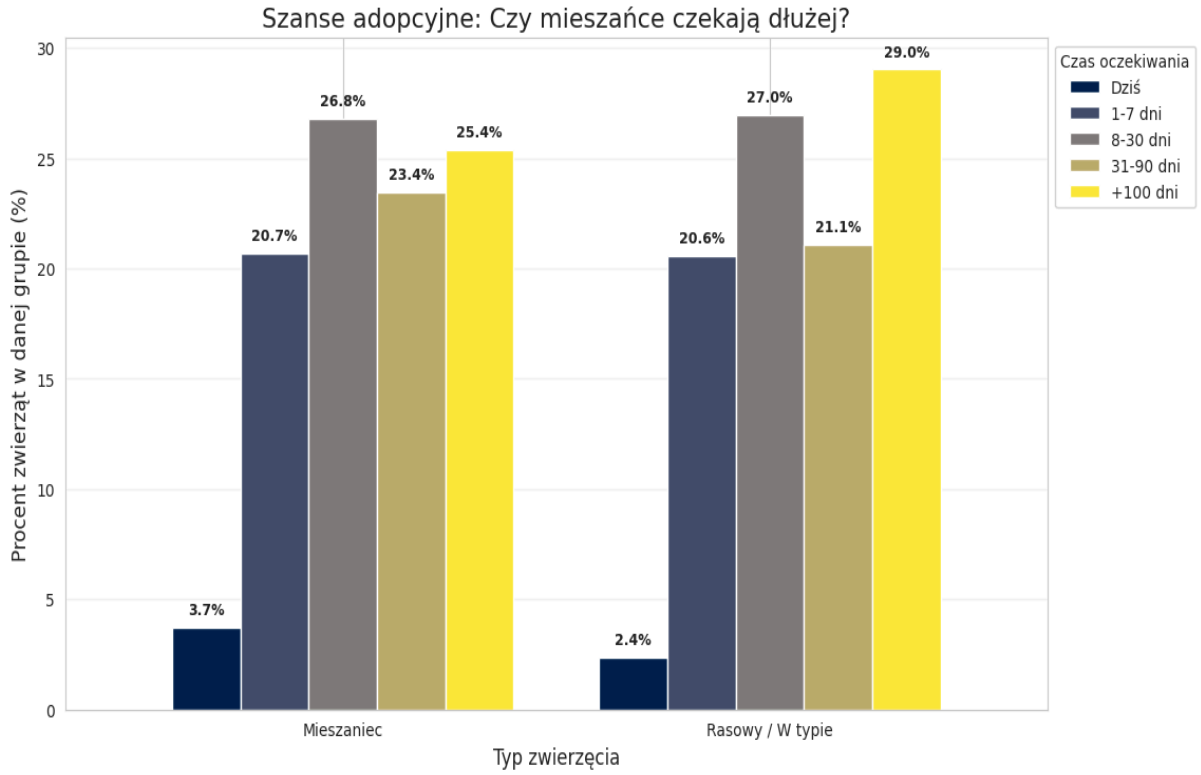


Wykres potwierdza istnienie zależności między wiekiem a czasem oczekiwania na dom. Generalny trend jest wzrostowy – im wyższa kategoria opóźnienia, tym wyższa średnia wieku zwierząt. Potwierdza to, że seniorzy są grupą najtrudniej adopcyjną.

Warto jednak odnotować obserwację w kategorii najszybszej. Średni wiek zwierząt adoptowanych "Dziś" jest wyższy niż tych z kategorii "1-7 dni" czy “8-30 dni”. Uważam, że może wynikać z procedur schroniskowych – bardzo młode zwierzęta często wymagają kwarantanny lub szczepień/odrobaczeń, co uniemożliwia ich wydanie w dniu przyjęcia

(przesuwając je do kolejnej kategorii). Z kolei w grupie "Dziś" mogą znajdować się dorosłe, zdrowe psy, które są gotowe do adopcji od ręki.

3.3. Szanse adopcyjne: Czy mieszańce czekają dłużej? Istotnym elementem analizy było zweryfikowanie hipotezy dotyczącej pochodzenia zwierząt. Powszechnie uważa się, że zwierzęta rasowe (Pure Breed) znajdują domy szybciej niż mieszańce (Mixed Breed). Aby to sprawdzić, porównano średnią szybkość adopcji dla obu tych grup.



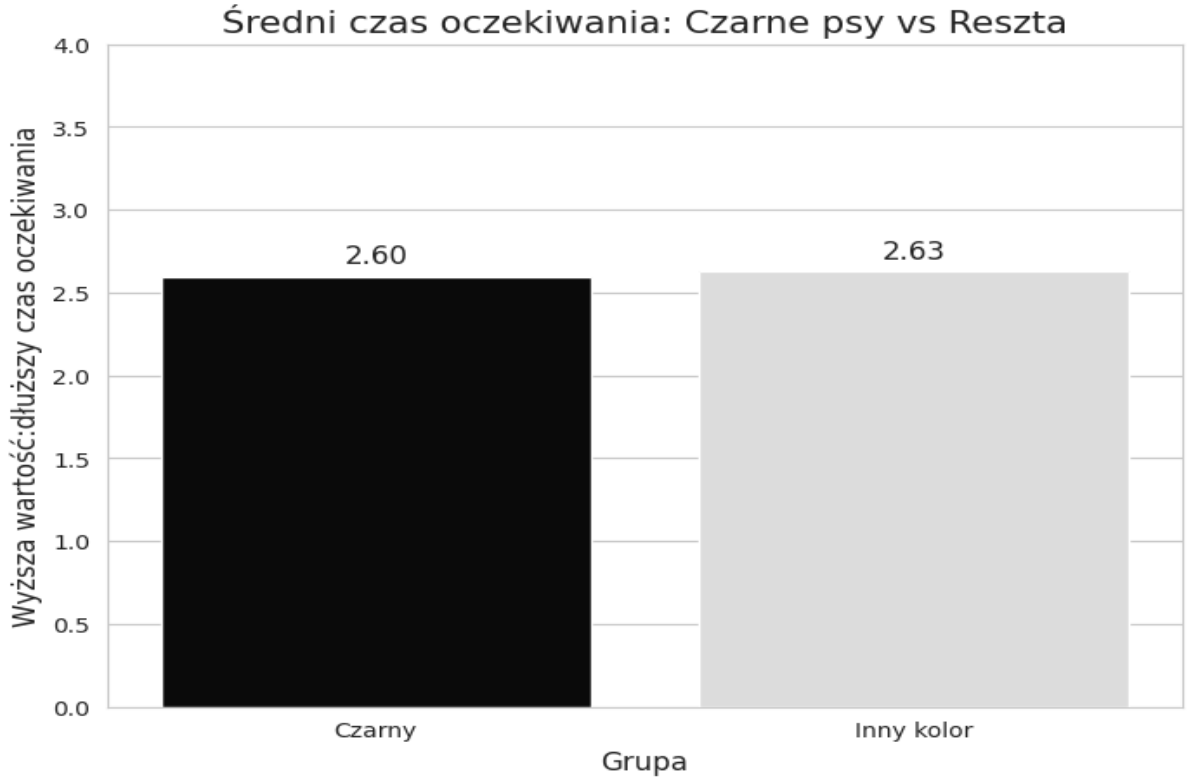
Analiza danych przyniosła nieoczywiste rezultaty, przeczące początkowej hipotezie:

- Adopcje błyskawiczne (Kategoria "Dziś"): Wbrew oczekiwaniom, to w grupie mieszańców odnotowano wyższy odsetek adopcji w dniu przyjęcia (3.7%) w porównaniu do zwierząt rasowych (2.4%). Oznacza to, że mieszańce częściej znajdują dom natychmiastowo.
- Długotrwały pobyt (Kategoria "+100 dni"): Również w przypadku najgorszego scenariusza statystyki są na niekorzyść zwierząt rasowych. Aż 29.0% z nich pozostaje w schronisku powyżej 100 dni, podczas gdy w grupie mieszańców odsetek ten jest niższy i wynosi 25.4%.

Wniosek: Dane wskazują, że status psa rasowego (lub w typie rasy) w warunkach schroniskowych nie gwarantuje szybszej adopcji. Może to wynikać z faktu, że rasowe psy

trafiające do schronisk często są starsze lub obciążone problemami zdrowotnymi/behawioralnymi, podczas gdy w grupie mieszańców częściej znajdują się młode, bezproblemowe psy, które szybciej zyskują sympatię adoptujących.

3.4. Weryfikacja "Syndromu Czarnego Psa" Ostatnim elementem analizy była weryfikacja hipotezy, czy czarne psy są dyskryminowane w procesie adopcyjnym. W tym celu wyizolowano grupę psów (z wyłączeniem kotów) i porównano średnią szybkość adopcji dla zwierząt o dominującym kolorze czarnym z pozostałymi.



Wyniki analizy są zaskakujące i nie potwierdzają powszechnego mitu o "Syndromie Czarnego Psa". Średni wskaźnik szybkości adopcji dla psów o czarnym umaszczeniu jest zbliżony (lub nawet niższy) w porównaniu do grupy kontrolnej. Oznacza to, że w badanym zbiorze danych kolor sierści nie stanowi bariery wydłużającej czas oczekiwania na nowy dom.

4. Przygotowanie danych (Preprocessing)

Przed przystąpieniem do budowy modelu, surowe dane musiały zostać odpowiednio przetworzone. Proces ten obejmował selekcję istotnych zmiennych oraz podział zbioru na część uczącą i testową.

4.1. Selekcja cech. Z oryginalnego zbioru danych usunięto kolumny, które nie niosły wartości analitycznej dla przyjętego algorytmu (takie jak imiona zwierząt, identyfikatory

`PetID` czy opisy tekstowe). Do ostatecznego modelowania wybrano zestaw 17 kluczowych cech numerycznych i kategorycznych:

- Cechy demograficzne: Typ (Pies/Kot), Wiek, Płeć, Ilość zwierząt w ogłoszeniu.
- Cechy fizyczne: Rasa (Breed1, Breed2), Kolor (Color1, Color2), Rozmiar (MaturitySize), Długość sierści (FurLength).
- Cechy zdrowotne: Szczepienie, Odrobaczenie, Sterylizacja, Ogólny stan zdrowia.
- Cechy ogłoszenia: Opłata (Fee), Liczba zdjęć, Liczba wideo.

Tak przygotowana macierz danych (`X`) posłużyła jako wsad do modelu, a wektorem wynikowym (`y`) stała się kolumna `AdoptionSpeed`.

4.2. Podział na zbiór treningowy i testowy. Aby rzetelnie ocenić skuteczność modelu, zastosowano standardową procedurę walidacji. Zbiór danych został podzielony w proporcji:

- 80% – Zbiór treningowy: Na tych danych model uczył się wzorców i zależności.
- 20% – Zbiór testowy: Te dane zostały ukryte przed modelem podczas nauki i posłużyły wyłącznie do sprawdzenia jego ostatecznej skuteczności.

Podziału dokonano z użyciem ziarna losowości (`random_state=42`), co zapewnia powtarzalność eksperymentu.

5. Budowa modelu i wyniki

Celem etapu modelowania było stworzenie algorytmu, który na podstawie cech zwierzęcia przewidzi szybkość jego adopcji. Pozwoliło to również wyłonić te atrybuty, które mają największy wpływ na decyzję adoptujących.

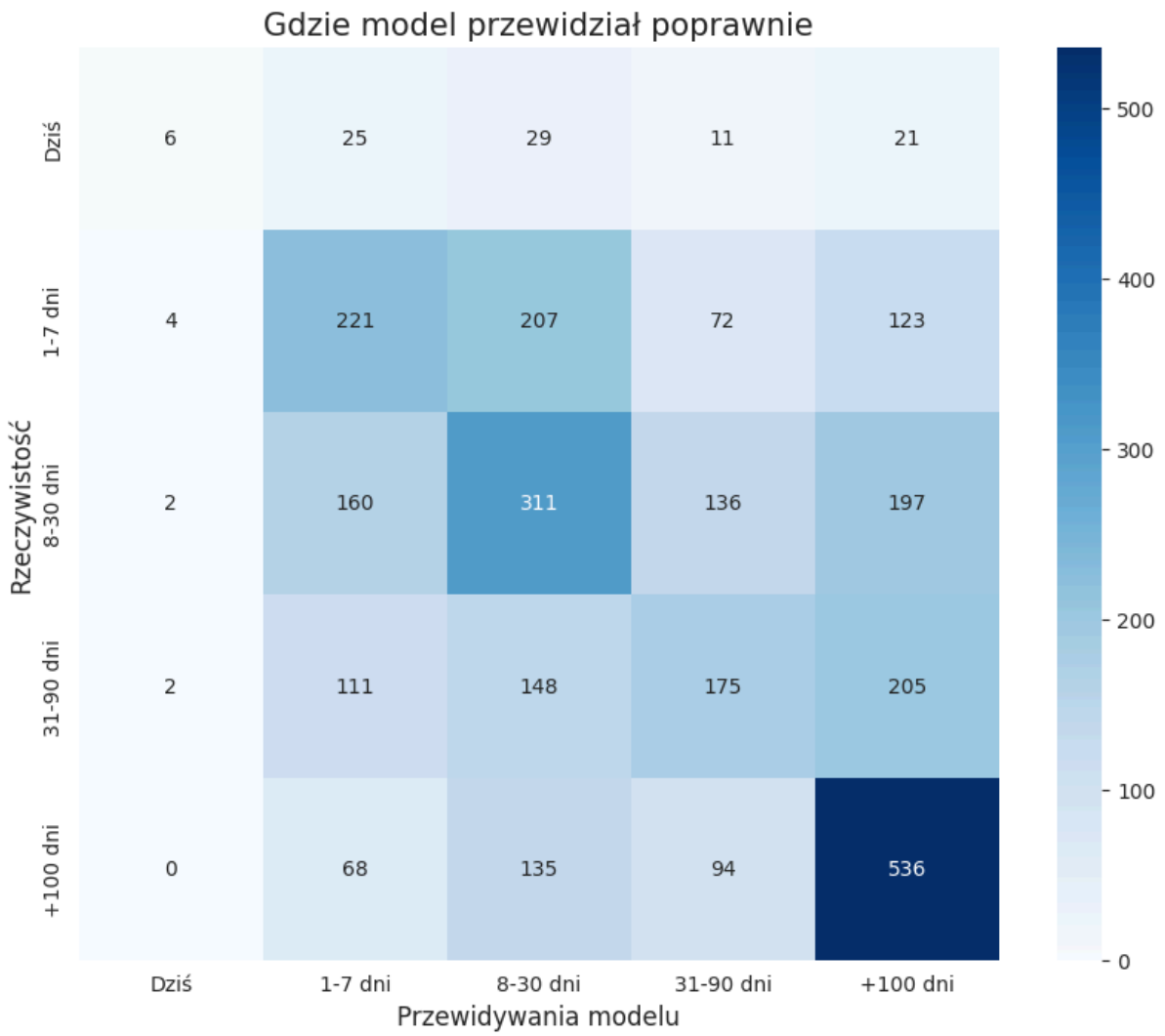
5.1. Wybór algorytmu. Do rozwiązania problemu klasyfikacji wybrano algorytm Random Forest Classifier (Las Losowy). Jest to metoda zespołowa (ensemble learning), która buduje wiele drzew decyzyjnych i uśrednia ich wyniki.

5.2. Wyniki i ocena skuteczności Po wytrenowaniu algorytmu na zbiorze uczącym, przeprowadzono ewaluację na zbiorze testowym (stanowiącym 20% wszystkich danych). Model osiągnął dokładność (Accuracy) na poziomie 41.65%.

Choć wartość ta może wydawać się umiarkowana, należy wziąć pod uwagę złożoność problemu (przewidywanie subiektywnych decyzji ludzkich) oraz fakt, że klasyfikacja

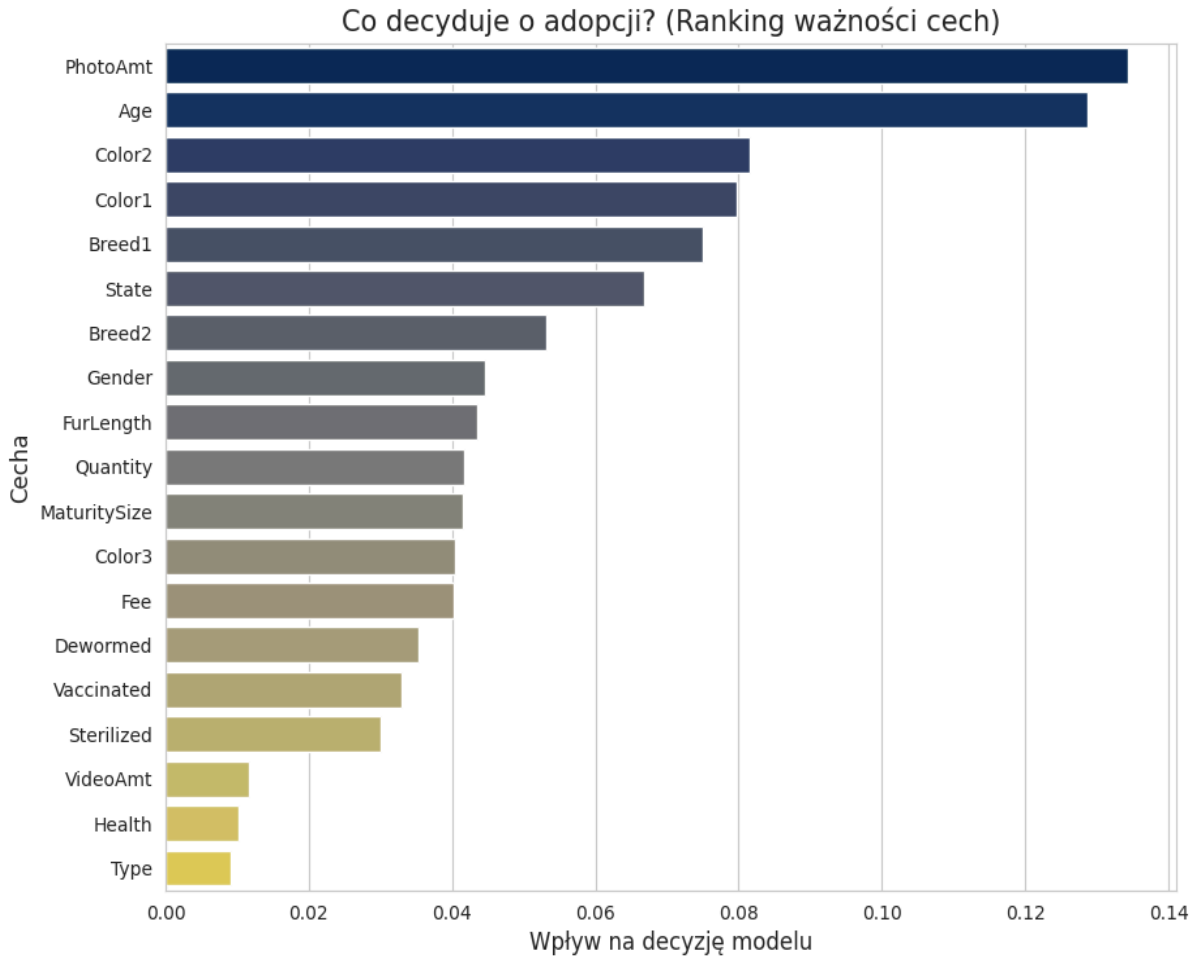
odbywała się w 5-stopniowej skali. Wynik ten jest ponad dwukrotnie lepszy od losowego przyporządkowania klas (które dałoby ok. 20% skuteczności).

Aby dokładniej przeanalizować błędy, wygenerowano Macierz Pomyłek (Confusion Matrix):



Macierz pokazuje, że model najczęściej myli ze sobą klasy sąsiadujące (np. kategorię 2 z 3). Oznacza to, że algorytm poprawnie wykrywa ogólny trend (szybka vs wolna adopcja), a błędy dotyczą głównie precyzyjnego określenia granicy dniowej.

5.3. Ranking ważności cech. Po wytrenowaniu modelu przeanalizowano, które zmienne miały największy wpływ na predykcję. Poniższy wykres prezentuje ranking najważniejszych czynników decydujących o szybkości adopcji.



Wyniki są jednoznaczne i wskazują na ogromną rolę jakości ogłoszenia. Zmienna PhotoAmt (Liczba zdjęć) okazała się najsilniejszym predyktorem w modelu, wyprzedzając nawet wiek zwierzęcia (Age). Oznacza to, że sposób prezentacji zwierzęcia w internecie ma kluczowe znaczenie dla decyzji adoptujących – im więcej zdjęć, tym większa szansa na szybką adopcję.

6. Wnioski:

- Marketing jest kluczem: Najważniejszym czynnikiem wpływającym na szybkość adopcji okazała się liczba zdjęć dołączonych do ogłoszenia. Potwierdza to, że atrakcyjna wizualizacja jest ważniejsza nawet od cech biologicznych.
- Wiek zajął drugie miejsce w rankingu ważności. Szczenięta i kocięta znajdują dom znacznie szybciej.

- Preferencje rasowe: Analiza przyniosła zaskakujące rezultaty. Wbrew powszechnej opinii, bycie psem rasowym (lub w typie rasy) nie gwarantuje szybszej adopcji. Dane wykazały, że w kategorii adopcji błyskawicznych to mieszańce radzą sobie nieco lepiej, a psy w typie rasy częściej pozostają w schronisku powyżej 100 dni.
- Mit "Syndromu Czarnego Psa" obalony: Wbrew powszechnym stereotypom, przeprowadzona analiza nie wykazała, aby czarne psy były rzadziej zabierane.

  Ich średni czas oczekiwania na adopcję był zbliżony do psów o innym umaszczeniu, co oznacza, że w badanym zbiorze kolor nie był barierą adopcyjną.

6.2. Rekomendacje dla schronisk. Na podstawie wyników można sformułować następujące zalecenia dla placówek::

1. Profesjonalizacja ogłoszeń: Zwiększenie liczby i jakości zdjęć jest kluczowe dla zwiększenia zasięgów ogłoszeń.
2. Zmiana strategii marketingowej: Skoro kolor nie jest problemem, a wiek tak – schroniska nie powinny skupiać się na promowaniu czarnych psów, ale powinny przenieść wysiłek na promocję seniorów.
3. Programy wsparcia seniorów: Rekomenduje się tworzenie specjalnych pakietów zachęt (np. darmowa opieka weterynaryjna) dedykowanych wyłącznie osobom adoptującym starsze zwierzęta, aby zniwelować główną barierę adopcyjną, jaką jest wiek.