

UNIWERSYTET ZIELONOGÓRSKI
WYDZIAŁ MATEMATYKI INFORMATYKI I EKONOMETRII

kierunek: Inżynieria Danych
specjalność: Systemy Eksploracji Danych

EWELINA KAMROWSKA

**WYKORZYSTANIE METOD UCZENIA
MASZYNOWEGO DO ROZPOZNAWANIA STANÓW
PORUSZANIA SIĘ OSOBY NA PODSTAWIE
DANYCH Z SYSTEMU GPS**

**THE USAGE OF MACHINE LEARNING METHODS FOR
IDENTIFICATION THE STATE OF HUMAN MOVEMENT
ON THE BASIS OF DATA FROM THE GPS SYSTEM**

Promotor pracy magisterskiej
DR MACIEJ NIEDZIELA

ZIELONA GÓRA, 2019

Spis treści

Wstęp	2
1 Regresja logistyczna	3
1.1 Funkcja logistyczna	3
1.2 Metoda estymacji największej wiarygodności	6
1.3 Prawdopodobieństwo przynależności obserwacji do klas	6
2 Ocena jakości algorytmów klasyfikacyjnych	8
2.1 Miary jakości klasyfikatora	8
2.2 Charakterystyka krzywej ROC	10
2.3 Badanie istotności zmiennych	11
2.4 Kryteria doboru rzędu zredukowanego modelu	11
3 Rozpoznawanie stanów poruszania się osoby na podstawie danych z systemu GPS	13
3.1 Opis problemu	13
3.2 Dane pomiarowe z odbiornika GPS	13
3.3 Analiza zachowań obserwacji z lokalizatora GPS	15
3.3.1 Klasyfikacja binarna przypadek dwóch klas (model 1)	16
3.3.2 Klasyfikacja w przypadku trzech klas (model 2)	23
4 Proces klasyfikacji danych z odbiornika GPS	30
4.1 Klasyfikacja w przypadku dwóch klas (model 1)	30
4.1.1 Klasyfikacja binarna I stopnia	30
4.1.2 Klasyfikacja binarna II stopnia	34
4.2 Klasyfikacja w przypadku trzech klas (model 2)	40
4.2.1 Klasyfikacja rodzaju drogi przemieszczania się autem	40
4.2.2 Klasyfikacja sposobu przemieszczania się osoby	44
4.3 Podsumowanie	47

Wstęp

Na przestrzeni ostatnich kilku lat miał miejsce ogromny postęp w obrębie dziedziny uczenia maszynowego (ang. *machine learning*). Uczenie maszynowe jest analizą procesów uczenia się oraz tworzeniem systemów, które doskonala swoje działanie na podstawie doświadczeń z przeszłości [11]. Uczenie maszynowe umożliwia pozyskanie wiedzy na podstawie analizy zachowań danych doświadczalnych, tj. przykładów uczących. Wiedza otrzymana metodami uczenia maszynowego może okazać się bardziej przydatna niż wiedza bezpośrednio wydedukowana przez ludzi. Czasami jest jedyną drogą budowy modeli, jeśli wiedza nie jest znana lub nie można jej pozyskać. Do najpopularniejszych i nieustannie rozwijających się metod uczenia maszynowego należą metody uczenia z nadzorem oraz bez nadzoru. Założeniem uczenia z nadzorem jest obecność ludzkiego nadzoru nad budową funkcji odwzorowującej wejście systemu na jego wyjście. Nadzór polega na stworzeniu zestawu danych uczących, tj. par: wejściowego obiektu uczącego i pożądanej przez użytkownika (nadzorce) odpowiedzi (będącej przykładowo konkretną wartością liczbową). Zadaniem systemu jest nauka predykcji prawidłowej odpowiedzi na zadane pobudzenie oraz generalizacja przypadków wyuczonych na przypadki, z którymi system jeszcze się nie zetknął. Uczenie bez nadzoru zakłada natomiast brak obecności wyjścia w danych uczących.

W niniejszej pracy podjęto próbę wykorzystania jednej z nadzorowanych metod uczenia maszynowego w celu rozpoznawania stanów przemieszczania się osoby na podstawie danych z systemu GPS. Rozdział pierwszy pracy opisuje użytą nadzorowaną metodę regresji logistycznej uczenia maszynowego oraz charakterystykę jej przypadków. Rozdział drugi został poświęcony opisowi zagadnień związanych z badaniem jakości zbudowanych modeli klasyfikacyjnych. Trzeci rozdział zawiera analizy zachowań obserwacji z lokalizatora GPS w różnych stanach poruszania się oraz opis proponowanego podejścia klasyfikacji danych. Czwarty rozdział poświęcono procesowi klasyfikacji obserwacji oraz przedstawieniu jego wyników. Na potrzeby rozwiązania postawionego problemu, urządzenie rejestrujące dane zostało udostępnione przez zielonogórską firmę *Hertz Systems Ltd.*

Rozdział 1

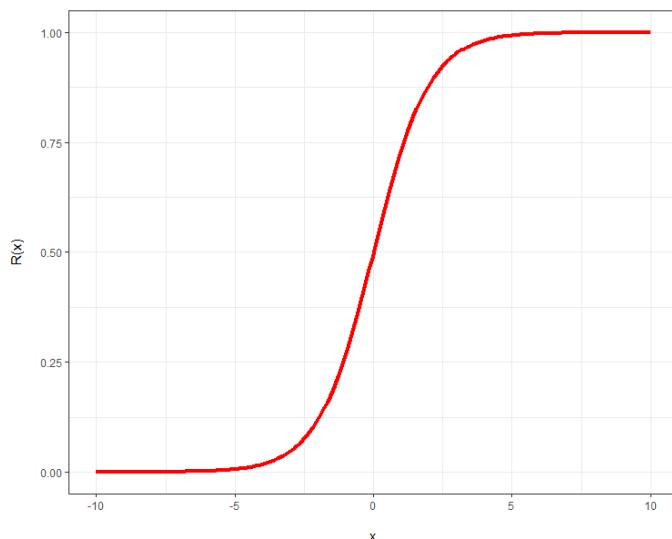
Regresja logistyczna

1.1 Funkcja logistyczna

W zagadnieniach klasyfikacji w wielu pracach badawczych, zmienna objaśniana nie ma rozkładu ciągłego lecz charakter binarny. Zmienne kategorialne dotyczące przykładowego problemu klasyfikacji *pacjent zdrowy/pacjent chory* nie powinno analizować się za pomocą regresji liniowej, gdyż błędne jest wówczas założenie o liniowym związku pomiędzy zmienną objaśnianą, a występującymi predyktorami. Sensowniejszym rozwiązaniem problemu takiej klasyfikacji jest zastosowanie regresji logistycznej, która zakłada nieliniowy związek pomiędzy zmienną objaśnianą, a zmiennymi wejściowymi (predyktorami) [3]. Niech $E(\hat{y}|x)$ oznacza szacowaną wartość oczekiwana zmiennej zależnej \hat{y} dla danej wartości predyktora x . Dla uproszczenia zapisu, przyjmijmy $E(\hat{y}|x) = R(x)$. Szacowana wartość oczekiwana zmiennej zależnej \hat{y} dla regresji logistycznej, ma postać

$$R(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}, \quad (1.1)$$

gdzie β_0 oznacza stałą regresji dla regresji logistycznej, natomiast $\beta_1, \beta_2, \dots, \beta_n$ to współczynniki regresji logistycznej dla kolejnych predyktorów x_1, x_2, \dots, x_n .



Rysunek 1.1: Wykres funkcji logistycznej $R(x)$.

Krzywe określone wzorem (1.1) noszą nazwę wykresów sigmoidalnych, ze względu na przypominający swoim wyglądem kształt litery S . Funkcja $R(x)$ jest nieliniowa, z minimum osiąganym w $\lim_{c \rightarrow -\infty} e^c/(1 + e^c) = 0$, a maksimum w $\lim_{c \rightarrow \infty} e^c/(1 + e^c) = 1$ [3].

Funkcja logistyczna $R(x)$ może być interpretowana jako prawdopodobieństwo wystąpienia wyniku pozytywnego p (*pacjent chory*) dla rekordów prezentowanych przez $X = R(x)$, oraz jako prawdopodobieństwo wyniku negatywnego $1 - p$ (*pacjent zdrowy*) gdy $X = 1 - R(x)$, przy założeniu $0 \leq R(x) \leq 1$. Zauważmy, że modele regresji liniowej zakładają $\hat{y} = \beta_0 + \beta_1 x + \epsilon$, gdzie współczynnik błędu ϵ ma rozkład normalny ze średnią zero i stałą wariancją. W przypadku użycia regresji logistycznej założenia są inne. Ponieważ zmienna zależna \hat{y} przyjmuje wartości binarne, błędy mogą przyjąć jedną z dwóch form. Jeśli $\hat{y} = 1$ (*pacjent chory*) występuje z prawdopodobieństwem $R(x)$, wówczas błąd jest postaci $\epsilon = 1 - R(x)$, a odległość wzdłuż osi pionowej pomiędzy punktem $\hat{y} = 1$, a krzywą $R(x)$ jest dokładnie pod nią dla $X = x$. Rozważając natomiast przypadek $\hat{y} = 0$ (*pacjent zdrowy*) z prawdopodobieństwem $1 - R(x)$, błąd przyjmuje postać $\epsilon = 0 - R(x) = -R(x)$ oraz odległość w pionie między punktem $\hat{y} = 0$, a krzywą logistyczną $R(x)$ jest dokładnie pod nią również dla $X = x$. Wariancję współczynnika błędu ϵ można zatem opisać jako iloczyn $R(x)[1 - R(x)]$. Jest to wariancja dla rozkładu dwumianowego, a szacowana wartość oczekiwana zmiennej zależnej $\hat{y} = R(x) + \epsilon$ zakłada reprezentację rozkładu dwumianowego z prawdopodobieństwem $R(x)$ [3].

Miara ilościowa, określająca przynależność obserwacji do danej klasy na podstawie prawdopodobieństw p oraz $1 - p$ nosi nazwę szansy (ang. **odds**). Przyjmując binarne oznaczenie dla dwóch niezależnych klas (wynik pozytywny $y = 1$ oraz wynik negatywny $y = 0$), miarę szansy można zdefiniować jako stosunek prawdopodobieństwa p przynależności do klasy $y = 1$, do prawdopodobieństwa $1 - p$ przynależności do klasy alternatywnej według wzoru

$$szansa = \frac{p}{1 - p}. \quad (1.2)$$

Miarę szansy można interpretować jako relację, pomiędzy prawdopodobieństwem sukcesu ($y = 1$), a prawdopodobieństwem porażki ($y = 0$). Równa szansa wystąpienia zarówno sukcesu, jak i porażki, ma miejsce wówczas, gdy $p = \frac{1}{2}$ (miara *szansy* jest równa 1). Ponadto, dla wartości $szansa > 1$ prawdopodobieństwo sukcesu jest większe niż prawdopodobieństwo porażki [1]. Znajomość miary *szansy* pozwala wyznaczyć prawdopodobieństwo p w następujący sposób

$$p = \frac{szansa}{1 + szansa}. \quad (1.3)$$

Większa wartość miary szansy, to większa szansa na sukces. Uwzględniając zależność określona wzorem (1.1), można prawdopodobieństwo p oraz $p - 1$ wyrazić wzorami

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}, \quad (1.4)$$

$$1 - p = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}. \quad (1.5)$$

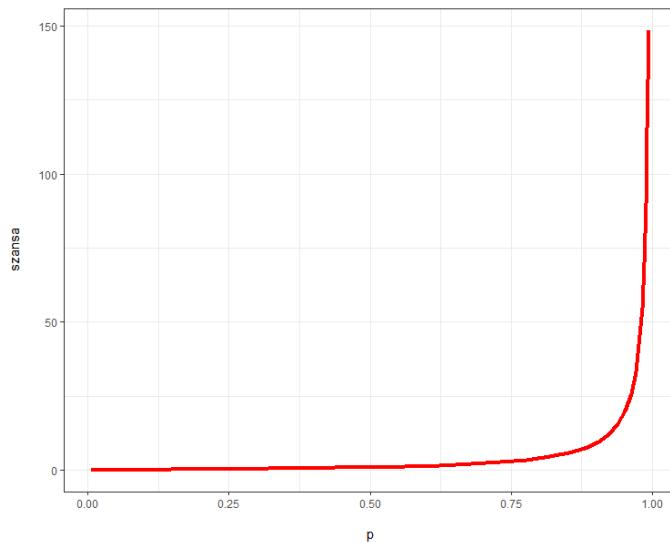
Uwzględniając powyższe zależności, miarę *szansy* można wyrazić w uproszczony sposób, jako funkcję wykładniczą predyktorów x_i postaci

$$szansa = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}. \quad (1.6)$$

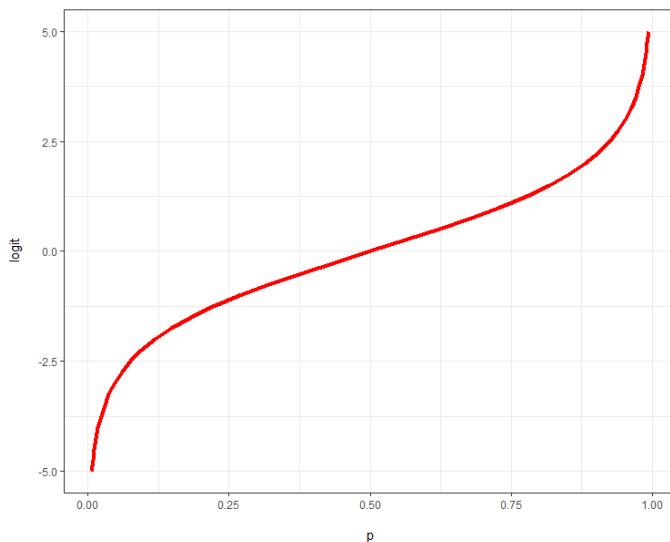
Na bazie powyższej funkcji, można zdefiniować **funkcję logitową**, określoną jako logarytm naturalny zmiennej *szansa*

$$\text{logit} = \ln(\text{szansa}) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n. \quad (1.7)$$

Warto zauważyć, że zmienna *logit* dla predyktorów x_i jest ich funkcją liniową. W funkcji zdefiniowanej w ten sposób, wartości zmieniają się w przedziale $[-\infty, \infty]$. Na rysunku 1.2 oraz 1.3 przedstawiono kolejno graficznie zależność zmiennej *szansy* i *logit* od wartości prawdopodobieństwa p . Zmienna *szansy* może przyjmować jedynie wartości dodatnie, zmienna *logit* natomiast, dopuszcza wartości dodatnie oraz ujemne.



Rysunek 1.2: Zależność funkcji *szansy* od wartości prawdopodobieństwa p .



Rysunek 1.3: Zależność funkcji *logit* od wartości prawdopodobieństwa p .

Funkcja *logit* jest liniowo zależna od kolejnych zmiennych objaśniających x_1, x_2, \dots, x_n , dzięki czemu możemy łatwo interpretować współczynniki regresji $\beta_1, \beta_2, \dots, \beta_n$. Jeżeli

- a) $e^{\beta_i} > 1$, to czynnik opisywany przez zmienną x_i ma istotny wpływ na opisywane zdarzenie,
- b) $e^{\beta_i} < 1$, to czynnik opisywany przez zmienną x_i działa ograniczająco,
- c) $e^{\beta_i} = 1$, to czynnik opisywany przez zmienną x_i nie ma istotnego wpływu na wystąpienie badanego zjawiska.

1.2 Metoda estymacji największej wiarygodności

W metodzie regresji liniowej, optymalne wartości współczynników regresji można otrzymać stosując metodę najmniejszych kwadratów. Dla metody regresji logistycznej natomiast, stosuje się metodę estymacji największej wiarygodności (ang. *maximum likelihood*). Niech $l(\beta|x)$ będzie funkcją wiarygodności parametrów $\beta = \beta_0, \beta_1, \dots, \beta_n$, która wyraża prawdopodobieństwa otrzymania obserwowanych zmiennych niezależnych x . Znalezienie wartości współczynników $\beta = \beta_0, \beta_1, \dots, \beta_n$, które maksymalizują funkcję $l(\beta|x)$, pozwala uzyskać estymatory największej wiarygodności, które są najkorzystniejszymi wartościami parametrów dla obserwowanych zmiennych x . Zakładając, że wszystkie obserwacje są od siebie niezależne, wiarygodność, to iloczyn prawdopodobieństw pojawiienia się poszczególnych obserwacji z próby przy danych parametrach modelu [3]

$$l(\beta|x) = \prod_{i=1}^n [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i}. \quad (1.8)$$

Dla uproszczenia obliczeń, na podstawie (1.8), wprowadza się logarytm największej wiarygodności postaci

$$L(\beta) = \ln[l(\beta|x)] = \sum_{i=1}^n y_i \ln[p(x_i)] + (1 - y_i) \ln[1 - p(x_i)]. \quad (1.9)$$

Estymatory największej wiarygodności można znaleźć, maksymalizując funkcję $L(\beta|x)$ przez poszukiwanie zera jej pochodnej względem wszystkich estymowanych parametrów [2]. Niestety, w przeciwieństwie do regresji liniowej, rozwiązanie w postaci analitycznej nie istnieje. Problem ten rozwiązuje się wówczas za pomocą metod iteracyjnych, stosując np. algorytm Newtona-Raphsona lub metody iteracyjne ważone najmniejszych kwadratów.

1.3 Prawdopodobieństwo przynależności obserwacji do klas

Rozpatrzmy zagadnienie klasyfikacyjne dla dwóch klas zdefiniowanych binarnie. Wówczas prawdopodobieństwo p przynależności obserwacji x do kolejno pierwszej i drugiej klasy ($i = 1, 2$) można zapisać w postaci

$$p_1 = p(kl = 1|x) = \frac{\exp(\beta_{10} + \bar{\beta}_1^T x)}{1 + \exp(\beta_{10} + \bar{\beta}_1^T x)}, \quad (1.10)$$

$$p_2 = p(kl = 2|x) = \frac{\exp(\beta_{20} + \bar{\beta}_2^T x)}{1 + \exp(\beta_{10} + \bar{\beta}_1^T x) + \exp(\beta_{20} + \bar{\beta}_2^T x)}, \quad (1.11)$$

gdzie $\bar{\beta}_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{in})$ dla $i = 1, 2$. Wartość teoretyczną zmiennej objaśnianej \hat{y} można wówczas wyznaczyć według standardowej zasady prognozy

$$\hat{y}_i = \begin{cases} 1 & \text{gdy } 0.5 < p_i \leq 1 \\ 2 & \text{gdy } 0 < p_i \leq 0.5 \end{cases}, \quad (1.12)$$

gdzie p_i oznacza prawdopodobieństwo teoretyczne uzyskane na podstawie (1.10) oraz (1.11). W przypadku niezbilansowania liczby wystąpień przypadków obu klas w próbie, do szacowania teoretycznej zmiennej objaśnianej można posłużyć się modyfikacją standardowej zasady

$$\hat{y}_i = \begin{cases} 1 & \text{gdy } x_{cut} < p_i \leq 1 \\ 2 & \text{gdy } 0 < p_i \leq x_{cut} \end{cases}, \quad (1.13)$$

gdzie parametr x_{cut} jest tak zwanym optymalnym punktem odcięcia. Jest to wartość zmiennej diagnostycznej, która optymalnie dzieli badaną zbiorowość na grupy należące do klasy 1 oraz 2.

Klasyfikację metodą regresji logistycznej można uogólnić na przypadek większej liczby klas. Założymy, że chcemy dokonać klasyfikacji dla trzech, jednakowo prawdopodobnych w wyborze klas ($kl = \{1, 2, 3\}$). Prawdopodobieństwo wystąpienia każdej z nich oznaczmy kolejno przez $p(kl = 1)$, $p(kl = 2)$ i $p(kl = 3)$. Przyjmijmy klasę 3 jako referencyjną. Funkcja logitowa dla klas 1 i 2 ma wówczas postać

$$\begin{aligned} \text{logit}(kl = 1) &= \ln \frac{p(kl = 1)}{p(kl = 3)} = \beta_{10} + \bar{\beta}_1^T x \\ \text{logit}(kl = 2) &= \ln \frac{p(kl = 2)}{p(kl = 3)} = \beta_{20} + \bar{\beta}_2^T x. \end{aligned}$$

Po przeprowadzeniu estymacji współczynników β_i , można przystąpić do wyznaczenia prawdopodobieństwa a posteriori przynależności wektora x do poszczególnych klas

$$\begin{aligned} p_1 = p(kl = 1) &= \frac{\exp(\beta_{10} + \bar{\beta}_1^T x)}{1 + \exp(\beta_{10} + \bar{\beta}_1^T x) + \exp(\beta_{20} + \bar{\beta}_2^T x)} \\ p_2 = p(kl = 2) &= \frac{\exp(\beta_{20} + \bar{\beta}_2^T x)}{1 + \exp(\beta_{10} + \bar{\beta}_1^T x) + \exp(\beta_{20} + \bar{\beta}_2^T x)} \\ p_3 = p(kl = 3) &= 1 - p(kl = 1) - p(kl = 2). \end{aligned}$$

Dla danej obserwacji, po wyznaczeniu wartości liczbowych poszczególnych prawdopodobieństw, przypisujemy ją do klasy charakteryzującej się największym prawdopodobieństwem.

Rozdział 2

Ocena jakości algorytmów klasyfikacyjnych

Badanie jakości klasyfikatora w problemach klasyfikacji oparte jest o zgodność estymowanej etykiety klasy, w zestawieniu z etykietą rzeczywistą obserwacji. Reprezentację wyników predykcji najczęściej przedstawia się za pomocą tak zwanej macierzy rozkładu klas (macierzy pomyłek), która w pracy oznaczamy przez A . Elementy wierszy reprezentują liczbę wzorców pochodzących z kolejnych klas, elementy kolumn natomiast wskazują liczbę wzorców wyestymowanych przez klasyfikator jako dana klasa.

	klasa I	klasa II	klasa III
klasa I	25	3	0
klasa II	2	48	1
klasa III	5	4	56

Tabela 2.1: Przykład macierzy pomyłek dla przypadku klasyfikacji trzech klas.

Elementy leżące na głównej przekątnej macierzy A reprezentują liczbę poprawnie rozpoznanych wzorców. Każdy element pozadiagonalny macierzy obrazuje błędna klasyfikację. Element (i, j) macierzy pomyłek określa liczbę wystąpień klasy i -tej, rozpoznanych jako klasa j -ta.

2.1 Miary jakości klasyfikatora

Bazując na zbudowanej macierzy rozkładu klas, można określić kilka miar jakości klasyfikatora. Jednym z nich jest tak zwany średni błąd względny, który reprezentuje stosunek liczby klasyfikacji błędnych elementów, do liczby wszystkich przypadków poddanych klasyfikacji. Niech a_{ij} będą elementami macierzy pomyłek. Wówczas błąd względny, dla całego zbioru obserwacji, można zapisać jako

$$\delta_w = \frac{\sum_{i \neq j} a_{ij}}{\sum_{i,j} a_{ij}}. \quad (2.1)$$

Warto zaznaczyć, że wskaźnik (2.1) nie odzwierciedla problemów związanych z niezrównoważoną liczebnością wystąpień poszczególnych klas w badanej grupie. Dla przykładu założmy, że w zbiorze danych 99% obserwacji należy do klasy pierwszej (pozytywnej), natomiast pozostały 1% do klasy przeciwniej (negatywnej). Wówczas przy całkowitym rozpoznaniu klasy pierwszej i zerowej skuteczności rozpoznania

klasy drugiej, średni błąd $\delta_w = 1\%$. Z ogólnego punktu widzenia jest to wynik satysfakcjonujący, jednak miara błędu nie odzwierciedla problemu, gdyż może się zdarzyć, że 1% klasy drugiej może reprezentować te przypadki w populacji, na których wykryciu nam zależy (np. wykrycie guza mózgu wśród grupy badanych pacjentów). Mając na względzie wystąpienie dużego zróżnicowania liczebności klas, warto zastosować alternatywne podejście do zdefiniowania jakości klasyfikatora i jego oceny.

Rozważmy przykładową binarną klasyfikację zbioru danych. Klasę pozytywną oznaczamy przez Klasa 1, a negatywną przez Klasa 2. Przy takich oznaczeniach, macierz A rozkładu klas przyjmuje postać

	Klasa 1	Klasa 2
Klasa 1	e_{++} (TP)	e_{+-} (FN)
Klasa 2	e_{-+} (FP)	e_{--} (TN)

Tabela 2.2: Symboliczny zapis macierzy pomyłek dla klasyfikacji binarnej.

Elementy macierzy pomyłek należy interpretować następująco: [4]

- a) e_{++} oznacza liczbę przypadków pozytywnych, poprawnie sklasyfikowanych jako pozytywne (ang. *True Positive* - TP),
- b) e_{+-} oznacza liczbę przypadków należących do klasy pozytywnej, sklasyfikowanych jako należące do klasy negatywnej (ang. *False Negative* - FN),
- c) e_{-+} oznacza liczbę przypadków należących do klasy negatywnej, sklasyfikowanych jako należące do klasy pozytywnej (ang. *False Positive* - FP),
- d) e_{--} oznacza liczbę przypadków negatywnych, poprawnie sklasyfikowanych jako negatywne (ang. *True Negative* - TN).

Przy tak przyjętych oznaczeniach możliwe jest określenie wskaźników definiujących jakość klasyfikacji, uwzględniających nieproporcjonalność obserwacji należących do dwóch klas. Do najpopularniejszych z nich należy [6]

- a) czułość (ozn. *TPR* ang. *True Positive Rate*) - określa stosunek liczby poprawnie sklasyfikowanych przypadków należących do klasy pozytywnej, do liczby wszystkich przypadków należących do tej klasy

$$TPR = \frac{TP}{TP + FN} \quad (2.2)$$

- b) specyficzność (ozn. *TNR* ang. *True Negative Rate*) - określa stosunek liczby poprawnie sklasyfikowanych przypadków należących do klasy negatywnej, do liczby wszystkich przypadków należących do tej klasy

$$TNR = \frac{TN}{TN + FP} \quad (2.3)$$

- c) wskaźnik rozpoznań fałszywie pozytywnych - wskaźnik fałszywych alarmów (ozn. *FPR* ang. *False Positive Rate*) - określa stosunek liczby przypadków należących do klasy negatywnej sklasyfikowanych jako należące do klasy pozytywnej, do liczby wszystkich przypadków należących do klasy negatywnej. Wskaźnik *FPR* jest ściśle powiązany ze specyficznością, to znaczy

$$FPR = \frac{FP}{FP + TN} = 1 - TNR \quad (2.4)$$

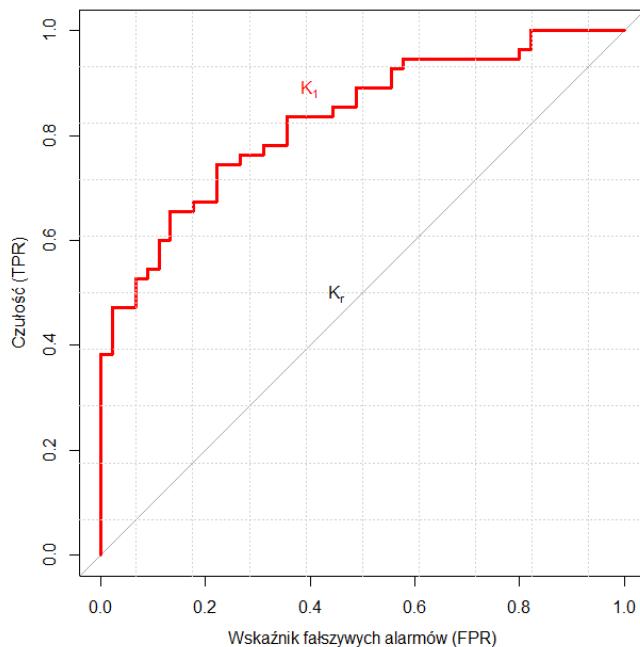
- d) dokładność (ozn. ACC ang. *Accuracy*) - określa procent poprawnie sklasyfikowanych obserwacji (należących do klasy pozytywnej oraz negatywnej) do wszystkich obserwacji poddanych klasyfikacji

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \quad (2.5)$$

Powyższe miary jakości ułatwiają ocenę oraz zauważenie pożądanych własności badanego klasyfikatora. Należy jednak pamiętać, że maksymalizacja jednego wskaźnika często jest wiążąca z pogorszeniem wartości innego. Jest to natomiast pomocne w budowie bilansu zysków i strat ogólnej klasyfikacji.

2.2 Charakterystyka krzywej ROC

Do oceny klasyfikatora często wykorzystuje się graficzną prezentację wyników klasyfikacji w postaci charakterystyki ROC (ang. *Receiver Operating Characteristics*). Krzywa ROC przedstawia zależność między wskaźnikiem TPR (oś y), a miarą FPR (oś x). Każdy punkt należący do krzywej określa inny dobór parametrów modelu klasyfikatora. Krzywa ROC jest graficzną reprezentacją efektywności modelu predykcyjnego poprzez wykreślenie charakterystyki jakościowej klasyfikatorów binarnych, powstały z modelu przy zastosowaniu wielu różnych punktów odcięcia x_{cut} . Mówiąc inaczej, każdy punkt krzywej ROC odpowiada innej macierzy rozkładu klas, uzyskanej przez modyfikowanie wielu różnych punktów odcięcia. [1]. Zbadanie większej liczby różnych punktów odcięcia skutkuje wykreśleniem większej liczby punktów na krzywej ROC. Rysunek 2.1 przedstawia dwie krzywe, jedną przykładowego klasyfikatora K_1 oraz drugą, klasyfikatora losowego K_r .



Rysunek 2.1: Wykres przykładowej krzywej ROC.

Z punktu oceny klasyfikatora, krzywa ROC zawiera kilka charakterystycznych cech. Gdy wartości obu współczynników TPR i FPR są równe 0, wówczas klasyfikator przyporządkowuje wszystkie obserwacje do klasy pozytywnej. Ponadto, gdy obie miary są równe 1 klasyfikator kwalifikuje wszystkie obserwacje do klasy negatywnej. Punkty położone na prostej diagonalnej, oznaczają klasyfikację całkowicie losową (niskiej jakości), natomiast gdy TPR jest równy 1, a współczynnik FPR równy 0, mamy do czynienia z klasyfikatorem idealnym, rozpoznającym bezbłędnie wszystkie obserwacje [1].

Podsumowując, dobry klasyfikator to taki, dla którego krzywa ROC położona jest możliwie najbliżej lewego górnego punktu układu współrzędnych $(0, 1)$. Możliwość wykreślenia krzywej, pozwala na ocenę jakości klasyfikatora w postaci metryki pola powierzchni AUC (ang. *area under curve*), powstałego pod utworzoną krzywą, gdzie $AUC \in [0, 1]$. Wartość utworzonego pola $AUC = 1$ oznacza klasyfikację bezbłędną, natomiast wartość $AUC = 0.5$ oznacza klasyfikację losową (decyzja podejmowana na podstawie modelu jest tak samo dobra, jak losowy wybór klasy dla danej obserwacji). Im wartość utworzonego pola AUC jest bliższa 1, tym lepsza jest ocena klasyfikacji utworzonego modelu [9].

2.3 Badanie istotności zmiennych

Przy budowie modelu klasyfikacji należy zbadać istotność poszczególnych zmiennych (predyktorów) wchodzących w skład modelu. W tym celu można skorzystać z testu Walda, sprawdzającego istotność statystyczną poszczególnych współczynników regresji β_i dla modelu, gdzie $i = 1, 2, \dots, n$. Test Walda opiera się na hipotezach [12]

$$H_0 : \beta_i = 0, \quad H_1 : \beta_i \neq 0.$$

Statystykę testową wyliczamy według wzoru

$$\chi^2 = \left(\frac{\beta_i}{SE(\beta_i)} \right)^2,$$

gdzie β_i jest współczynnikiem regresji dla predyktora zmiennej i , natomiast $SE\beta_i$ jest błędem standardowym dla predyktora zmiennej i . Statystyka ta ma asymptotycznie (dla dużych liczebności) rozkład chi-kwadrat z 1 stopniem swobody [10].

Wyznaczoną na podstawie statystyki testowej wartość p porównujemy z poziomem istotności α [12]

jeżeli $p \leq \alpha \Rightarrow$ odrzucamy H_0 przyjmując H_1 ,

jeżeli $p > \alpha \Rightarrow$ nie ma podstaw, aby odrzucić H_0 .

2.4 Kryteria doboru rzędu zredukowanego modelu

Istotną rolę przy ocenie jakości klasyfikacji odgrywa wymiar modelu. Skuteczny model powinien być jak najprostszy, to znaczy zawierać minimalną liczbę parametrów ale również powinien stosunkowo poprawnie odtwarzać dane oryginalne (wejściowe). Zakładając istnienie n obserwacji w postaci wektorów danych x_i ($i = 1, 2, \dots, n$) proces rzeczywisty można zapisać w postaci $f(x_1, x_2, \dots, x_n | \beta)$, gdzie β oznacza zbiór parametrów predykcyjnych $\beta = (\beta_1, \beta_2, \dots, \beta_n)$. Ograniczając model

do K składowych (dokonując redukcji liczby predyktorów), model rzeczywisty można zapisać w postaci $f(x_1, x_2, \dots, x_n | \beta_k)$. Zwiększać liczbę parametrów modelu zmniejsza się błąd systematyczny dopasowania (tzw. błąd obciążenia), lecz powiększa się jego wariancja [1]. Wartość K przyjęto nazywać rzędem lub wymiarem zredukowanego modelu. Wektor β_k opisuje rzutowanie oryginalnego wektora β w przestrzeń o niższym wymiarze K . Należy pamiętać, że dobór predytorów powinien zapewniać maksymalną wiarygodność modelu względem modelowanego procesu. Dylemat doboru właściwej wartości K polega na szukaniu kompromisu pomiędzy obciążeniem, a wariancją [1]. Dobór optymalnej wartości K może być wyznaczany według różnych kryteriów. Najczęściej stosowane są [9]

- a) kryterium informacyjne Akaike AIC (ang. *Akaike Information Criterion*) wyrażone wzorem

$$AIC = -2 \ln L(\beta_k) + 2K, \quad (2.6)$$

gdzie czynnik $L(\beta_k)$ wyraża maksymalizację funkcji wiarygodności (1.9) modelu pełnego (ze wszystkimi zmiennymi).

- b) współczynnik McFaddena R_p^2 - mierzy łączną istotność oszacowanych parametrów modelu, w stosunku do hipotezy zerowej, w taki sposób że wszystkie parametry modelu są równe zero. Wartość statystyki obliczana jest na postawie wartości funkcji logarytmu wiarogodności i wyrażona jest wzorem

$$R_p^2 = 1 - \frac{L(\beta_k)}{L(\beta_0)}, \quad (2.7)$$

gdzie czynnik $L(\beta_0)$ oznacza maksimum funkcji wiarygodności (1.9) modelu zawierającego jedynie wyraz wolny. Pomimo tego że statystyka R_p^2 przyjmuje wartości z przedziału $(0, 1)$, nie ma ona takiej interpretacji jak statystyka R^2 dla modelu regresji liniowej. Nawet gdy model jest idealnie dopasowany do danych, metryka R_p^2 przyjmuje wartości znacznie mniejsze od 1. Wobec tego nie jest ona optymalną miarą dopasowania modelu. Na jej podstawie można natomiast stwierdzić, który z wyestymowanych modeli na tej samej próbie danych, z tymi samymi zmiennymi objaśniającymi, jest lepiej do nich dopasowany. Im wyższa wartość współczynnika McFaddena, tym lepsze jest dopasowanie modelu.

Rozdział 3

Rozpoznawanie stanów poruszania się osoby na podstawie danych z systemu GPS

3.1 Opis problemu

We wstępnie pracy wspomniano, że uczenie maszynowe stanowi szereg analiz oraz procesów, służących doskonaleniu swoich działań na podstawie doświadczeń z przeszłości [12]. Jednym z najpopularniejszych problemów rozwiązywanych przez systemy uczące się są problemy dotyczące procesu klasyfikacji. Przykład może tutaj stanowić uczenie rozpoznawania chorób na podstawie występujących symptomów, rozpoznawanie spamu w skrzynkach pocztowych odbiorców internetowych czy przewidywanie trendów w danych finansowych. To tylko kilka z licznych i coraz bardziej popularnych praktycznych zastosowań uczenia maszynowego.

Do popularnych metod klasyfikacji danych należą drzewa decyzyjne, lasy losowe, sieci neuronowe czy metoda regresji logistycznej. W niniejszej pracy, w celu rozpoznawania stanów poruszania się osoby z lokalizatorem GPS, posłużono się ostatnią z wymienionych metod, regresją logistyczną.

W odniesieniu do autorskiej pracy inżynierskiej pt. *Wykrywanie anomalii w pozycjonowaniu lokalizatora GPS*, przedstawiony algorytm wykrywania anomalii opierał się o podział danych na klasy decyzyjne, wykorzystując dalej binarną funkcję decyzyjną, złożoną z wielu prostych warunków logicznych. Niekiedy jednak, problemy klasyfikacji są na tyle złożone, że system powinien dynamicznie dostosowywać się do zmieniających się warunków i nie zawsze problemy tego rodzaju są możliwe do rozwiązania przez powiązanie ze sobą wielu warunków logicznych. Niniejsza praca stanowi zatem kontynuację propozycji rozwiązań problemów klasyfikacji danych pomiarowych z systemu GPS, wykorzystując przy tym metodę regresji logistycznej, która obecnie cieszy się dużą popularnością wśród nadzorowych metod klasyfikacji uczenia maszynowego.

3.2 Dane pomiarowe z odbiornika GPS

Do klasyfikacji obserwacji posłużono się danymi magazynowanymi przez lokalizator *HP-500* w formacie tabelarycznym *csv*. Każdy z arkuszy zawierał zestaw informacji o przebytej przez osobę z odbiornikiem trasie przejazdu. Zawarte w arkuszach obserwacje zostały scharakteryzowane przez konkretne wielkości, takie jak:

- a) **Czas GPS** t_n - czas, w którym urządzenie wyznacza pozycję osoby z lokalizatorem GPS,
- b) **Szerokość geograficzna** φ_n (ang. *Longitude*) - wyznaczona przez urządzenie szerokość geograficzna w chwili t_n ,
- c) **Długość geograficzna** λ_n (ang. *Latitude*) - wyznaczona przez urządzenie długość geograficzna w chwili t_n ,
- d) **Azymut** a_n - odchylenie od kierunku północnego mierzone w stopniach $[^\circ]$ w chwili t_n ,
- e) **GPS** N_{GPS} - liczba widocznych przez odbiornik satelitów GPS w danej chwili t_n [8].

Zanim przystąpiono do przetwarzania danych, konieczne stało się ich uporządkowanie. Odrzucono rekordy z powtarzającym się czasem t_n oraz rekordy zawierające braki wartości. Istotną metryką, którą należało uwzględnić przy oczyszczaniu danych była liczba widocznych przez odbiornik satelitów N_{GPS} . W celu wyznaczenia poprawnej lub choćby przybliżonej pozycji odbiornika na Ziemi, ważne jest aby w zasięgu jego widoczności znajdowały się przynajmniej cztery satelity. Pomiary pseudoodległości znieksztalcone są przez błąd synchronizacji, błąd zegara atomowego (satelita) oraz błąd zegara kwarcowego (urządzenie pozycjonujące) [7]. Czas zegara atomowego korygowany jest dzięki depeszy nawigacyjnej, natomiast błąd zegara kwarcowego traktowany jest jako czwarty nieznany parametr. To powoduje, że ilość nieznanych parametrów wynosi 4: trzy to współrzędne urządzenia pozycjonującego, czwarty to błąd zegara odbiornika [13]. Stąd przy porządkowaniu danych przyjęto następujący warunek $N_{GPS} \geq 4$.

W oparciu o dostępne dane, możliwe stało się wyznaczenie wielkości charakteryzujących sposób przemieszczania się osoby z lokalizatorem GPS. Na ich podstawie, w późniejszym procesie klasifikacji danych, wykorzystano następujące zmienne:

- a) **Różnica czasu** Δt_n - czas pomiędzy wyznaczeniem dwóch kolejnych pozycji GPS, wyrażony w sekundach [s]. Niech t_n oraz t_{n-1} oznaczają czas wyznaczenia pozycji lokalizatora dla dwóch kolejnych momentów. Wówczas różnicę czasu Δt_n można wyrazić wzorem [8]

$$\Delta t_n = t_n - t_{n-1}.$$

- b) **Dystans** d_{GPS} - dystans, jaki pokonała osoba posiadająca lokalizator, uwzględniając jej położenie (długość i szerokość geograficzną) w dwóch kolejnych obserwacjach. Niech $P_n = (\lambda_n, \varphi_n, h_n)$ oraz $P_{n-1} = (\lambda_{n-1}, \varphi_{n-1}, h_{n-1})$ będą dwoma kolejnymi punktami określającymi położenie odbiornika GPS w chwili t_n oraz t_{n-1} . Odległość d_{GPS} między dwoma pozycjami na kuli można wówczas wyznaczyć ze wzoru [8]

$$d_{GPS} = \arccos[\cos(90^\circ - \varphi_n) \cdot \cos(90^\circ - \varphi_{n-1}) + \sin(90^\circ - \varphi_n) \cdot \sin(90^\circ - \varphi_{n-1}) \cdot \cos(\lambda_n - \lambda_{n-1})] \cdot r,$$

gdzie r jest promieniem Ziemi.

- c) **Prędkość** v_{GPS} - prędkość, jaką osiąga osoba przemieszczająca się z lokalizatorem GPS w kolejnych obserwacjach. Uwzględniając różnicę czasów Δt_n dwóch kolejnych rekordów oraz przebyty dystans d_{GPS} , prędkość v_{GPS} można wyrazić wzorem [8]

$$v_{GPS} = \frac{d_{GPS}}{\Delta t_n}, \quad \Delta t_n \neq 0.$$

- d) **Flaga prędkość** $F(v_{GPS})$ - informacja o sfaktoryzowanej wartości prędkości

$$F(v_{GPS}) = \begin{cases} 0 & \text{gdy } v_{GPS} \leqslant 50 \frac{\text{km}}{\text{h}} \\ 1 & \text{gdy } 50 \frac{\text{km}}{\text{h}} < v_{GPS} \leqslant 90 \frac{\text{km}}{\text{h}} \\ 2 & \text{gdy } v_{GPS} > 90 \frac{\text{km}}{\text{h}} \end{cases} .$$

Powyższe progi ustalono na podstawie kodeksu ruchu drogowego o dopuszczalnej prędkości poruszania się w terenie zabudowanym oraz niezabudowanym.

- e) **Zmiana azymutu** Δa_n - informacja o wartości zmiany azymutu pomiędzy dwoma kolejnymi obserwacjami, wyrażona w stopniach [$^\circ$]. Niech a_n oraz a_{n-1} będą wielkościami azymutu kolejno w chwili t_n oraz t_{n-1} . Wówczas zmiana azymutu Δa_n wyrażona jest wzorem [8]

$$\Delta a_n = \min\{|a_n - a_{n-1}|, 360^\circ - |a_n - a_{n-1}|\}.$$

- f) **Procent zmiany azymutu** $F(\Delta a_n)$ - procent liczby obserwacji w oknie czasowym, dla których wartość zmiany azymutu Δa_n przekracza 15° . Wartość ta została ustaliona domyślnie, gdyż jednym z założeń wyznaczania kolejnych pozycji lokalizatora typu *HP-500* jest właśnie zmiana azymutu o wartość 15° . Wielkość okna czasowego ustalono natomiast na jedenaście. Po przeprowadzeniu licznych testów na dostępnych danych, wartość ta została przyjęta za domyślną, która oczywiście może być zmienna.

3.3 Analiza zachowań obserwacji z lokalizatora GPS

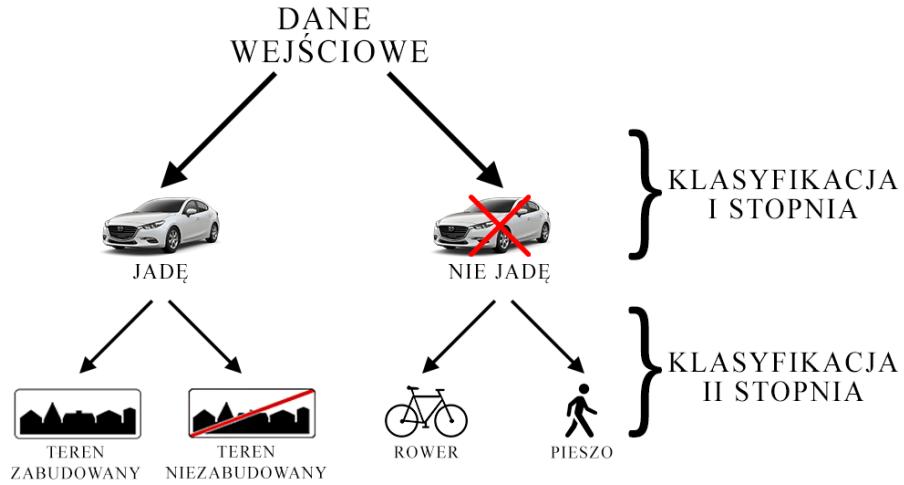
W niniejszej pracy za cel postawiono omówienie i zastosowanie metod klasyfikacji danych zbieranych przez odbiornik GPS opartych na modelu regresji logistycznej, z porównaniem uzyskanych wyników przez dwa modele. Pierwszy z nich, **model klasyfikacji binarnej**, zakłada klasyfikację dwustopniową, wyróżniającą kilka stanów przemieszczania się osoby z lokalizatorem GPS.

1. Klasyfikacja pierwszego stopnia obejmuje

- a) klasyfikację binarną przemieszczania się autem (**jadę lub nie jadę**).

2. Klasyfikacja drugiego stopnia obejmuje

- a) jeśli **jadę** - klasyfikację binarną miejsca przemieszczania się autem (**teren zabudowany lub teren niezabudowany**),
 b) jeśli **nie jadę** - klasyfikację binarną sposobu przemieszczania się środkiem innym niż auto (**rower lub pieszo**).



Rysunek 3.1: Schemat klasyfikacji binarnej (model 1), źródło: *opracowanie własne*.

Drugi z modeli uwzględnia natomiast **przypadek klasyfikacji trzech klas** dla danych pozyskanych z lokalizatora GPS. Klasyfikacja obserwacji dla tego modelu jest scharakteryzowana w nieco odmienny sposób i obejmuje dwie spośród niżej opisanych klasyfikacji.

1. Klasyfikacja **rodzaju drogi przemieszczania się autem** w trzech sytuacjach
 - a) przemieszczanie się autem **po drodze ekspresowej**,
 - b) przemieszczanie się autem **w terenie niezabudowanym**,
 - c) przemieszczanie się autem **w terenie zabudowanym**.
2. Klasyfikacja **sposobu przemieszczania się osoby** w terenie zabudowanym
 - a) przemieszczanie się **pieszo**,
 - b) przemieszczanie się **autem**,
 - c) przemieszczanie się **rowerem**.

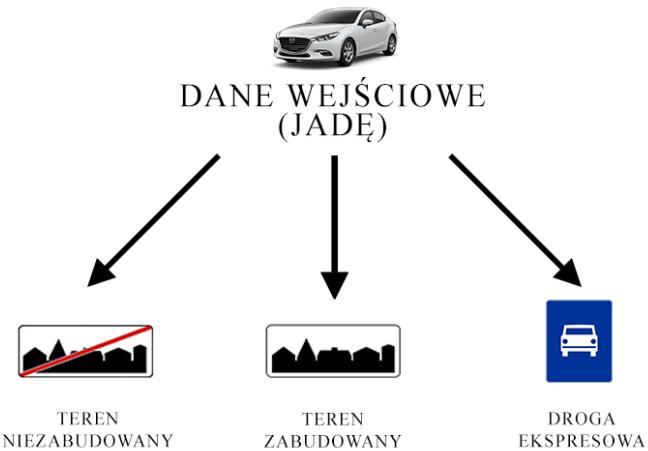
Klasyfikacja przedstawiona na rysunku 3.1 zawiera klasy wzajemnie od siebie zależne. Oznacza to, że do klasyfikacji drugiego stopnia posłuży nam zbiór danych wejściowych wykorzystany w pierwszym stopniu klasyfikacji, jednak będzie on powiększony, o obserwacje sprzeczne z ideą klasyfikacji drugiego stopnia (przykładowo dla klasyfikacji rower/pieszo, nie powinniśmy rozważać obserwacji, gdzie faktycznym stanem poruszania się była jazda autem).

Proces klasyfikacji przedstawiony na rysunku 3.3 odbywa się natomiast niezależnie od siebie. Oznacza to, że zarówno do klasyfikacji rodzaju drogi przemieszczania się autem, jak i do klasyfikacji sposobu przemieszczania się osoby z lokalizatorem GPS w terenie zabudowanym, użyto dwóch niezależnych zbiorów danych wejściowych.

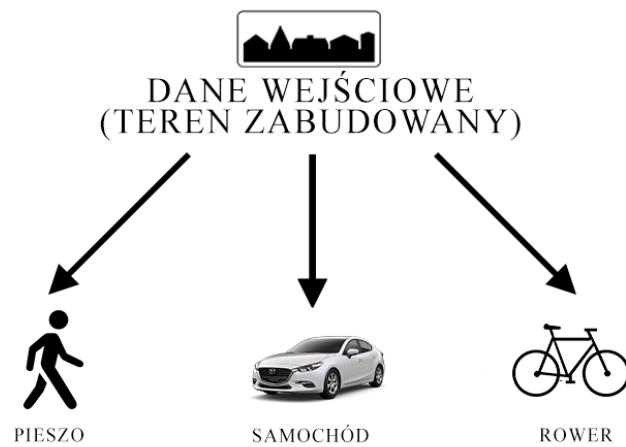
3.3.1 Klasyfikacja binarna przypadek dwóch klas (model 1)

Klasyfikacja I stopnia

Celem wykonania klasyfikacji danych, wykonano badania nad zachowaniem się niektórych parametrów opisanych w podrozdziale (3.2) i krótko je scharakteryzowano.

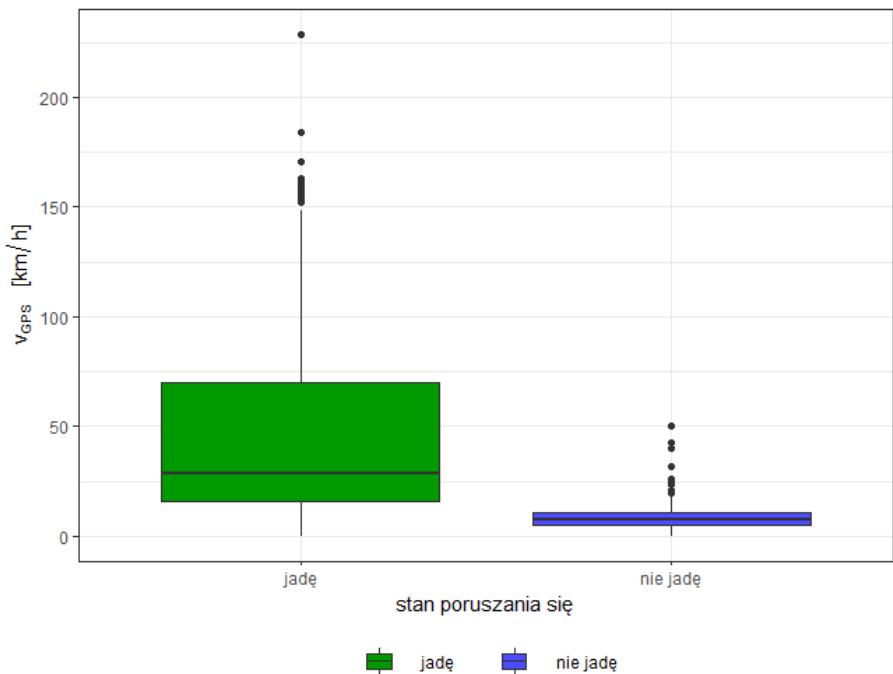


Rysunek 3.2: Schemat klasyfikacji rodzaju drogi przemieszczania się autem, źródło: *opracowanie własne*.

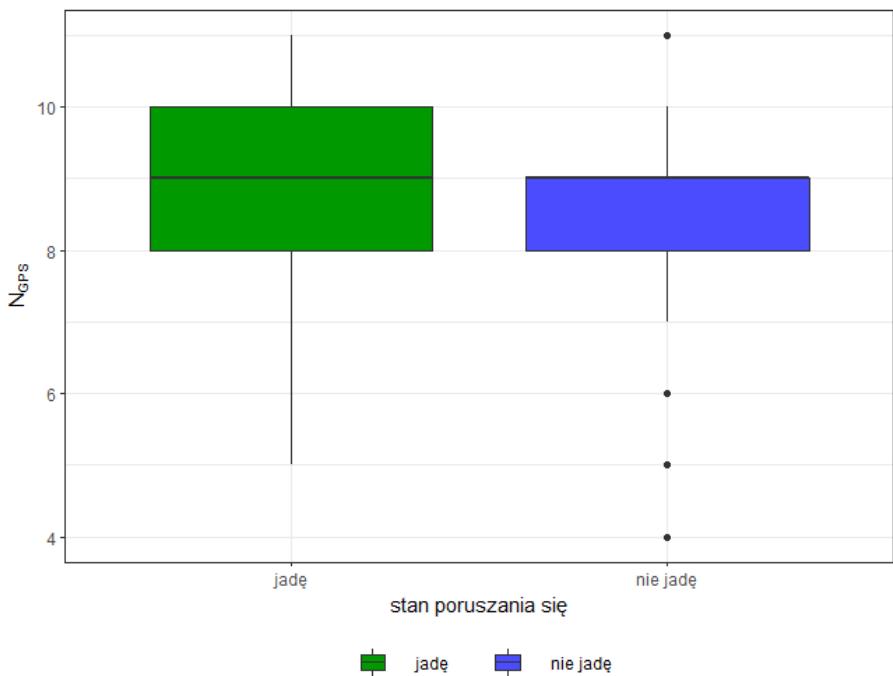


Rysunek 3.3: Schemat klasyfikacji sposobu przemieszczania się w terenie zabudowanym (model 2), źródło: *opracowanie własne*.

Rysunek 3.4 przedstawia rozkład prędkości w dwóch stanach przemieszczania się. Naturalnym zjawiskiem jest osiąganie zdecydowanie większej prędkości w przypadku jazdy autem, niż podczas poruszania się pieszo lub jazdy rowerem. Możemy zatem przypuszczać, że metryka opisująca prędkość dla tak zdefiniowanych stanów przemieszczania się, będzie determinującym czynnikiem wpływającym na jakość budowanego modelu klasyfikacji.



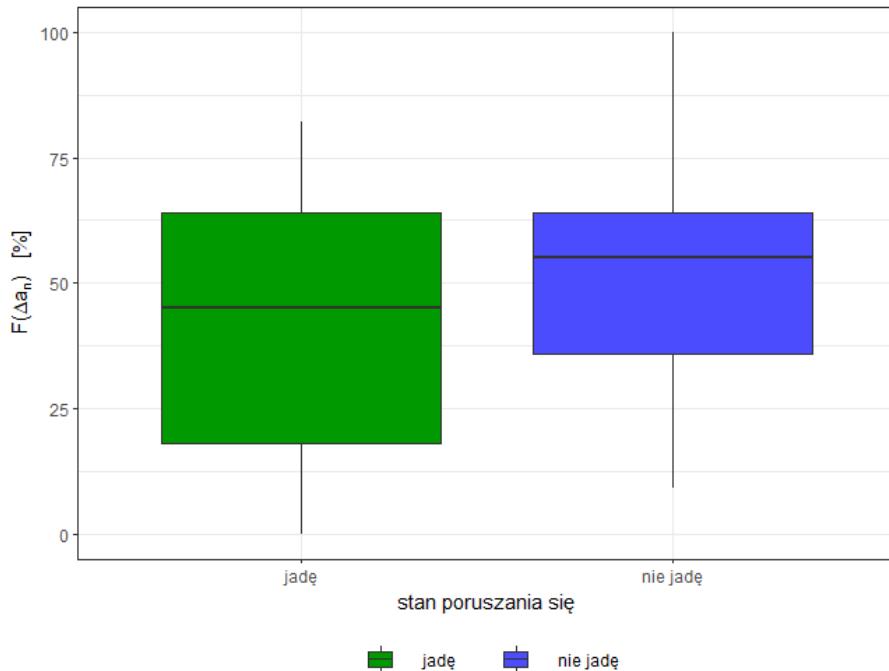
Rysunek 3.4: Wykres rozkładu prędkości v_{GPS} .



Rysunek 3.5: Wykres rozkładu widocznych przez odbiornik satelitów N_{GPS} .

Rysunek 3.5 przedstawia rozkład widocznych przez odbiornik satelitów w dwóch stanach przemieszczania się. Łatwo zauważać, że podczas przemieszczania się w sposób inny niż auto, mamy do czynienia z większym wahaniem widoczności liczby satelitów przez odbiornik GPS. Wynikać to może ze specyfiki terenu, po którym użytkownik porusza się w sposób inny jak auto. Często jest to teren zabudowany, gdzie widoczność satelitów może być ograniczona przez zabudowę wysokich bloków, wieżowców lub lasów. Inną opcją jest przemieszczanie się w budynku, gdzie zamknięta przestrzeń ogranicza nadawany przez satelity sygnał GPS. Opisaną sytuację przed-

stawia rysunek 3.5, na którym czarnymi kropkami oznaczono obserwacje odstające. Wynikać one mogą również z faktu, że odbiornik namierza swoją pozycję tym częściej, im osiągane prędkości w kolejnych obserwacjach są większe. Specyfika działania urządzenia *HP-500* zakłada wówczas bardziej regularne wyznaczanie położenia. W przypadku poruszania się w sposób inny niż samochód, regularność ta może zostać zachwiana, ze względu na osiąganie mniejszych prędkości v_{GPS} .



Rysunek 3.6: Wykres rozkładu procentowej zmiany azymutu $F(\Delta a_n)$ o więcej jak 15° .

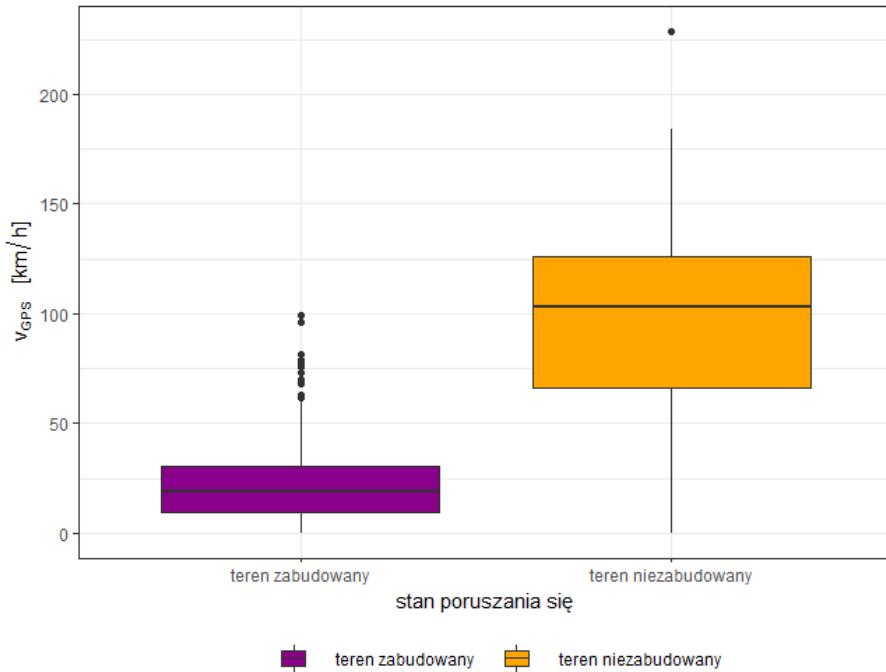
Rysunek 3.6 pokazuje, że rozkład współczynnika $F(\Delta a_n)$ dla stanu poruszania się w sposób inny jak auto, osiąga wyższe wartości. Fakt ten wynikać może ze specyfiki tego stanu poruszania się. Użytkownik przemieszczający się z lokalizatorem GPS podczas spaceru może wykonywać manewry, które nie są płynnymi zmianami pozycji GPS i dla których wartości zmiany azymutu Δa_n mogą być znacznie różnicą się. Podczas jazdy autem natomiast, przemieszczanie się jest z reguły płynne, w rezultacie czego wartości zmiany azymutu Δa_n nie zmieniają się w sposób gwałtowny.

Klasyfikacja II stopnia miejsca przemieszczania się autem

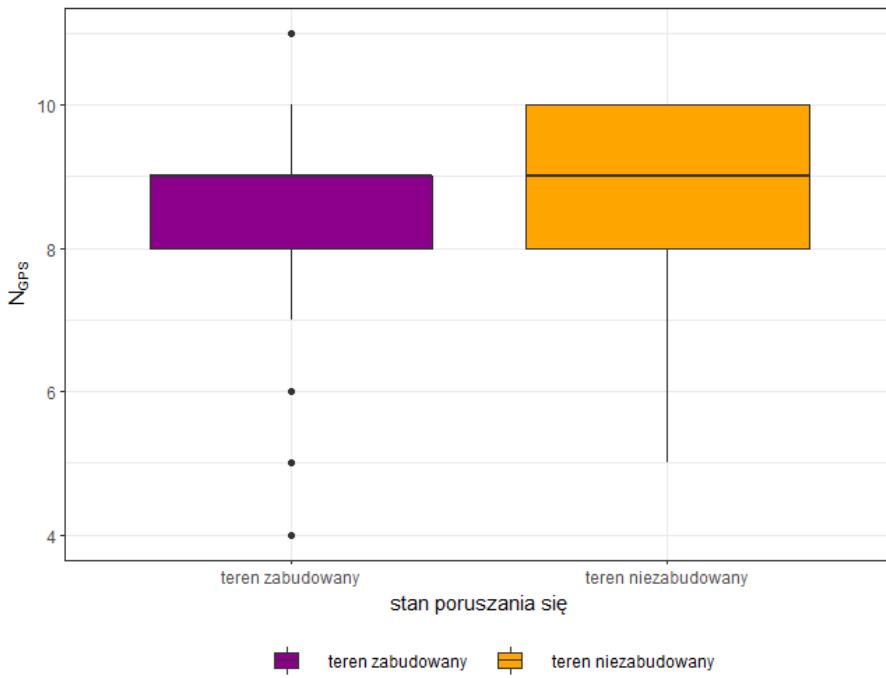
Poniższe zestawienia obejmują badania nad zachowaniami niektórych parametrów opisanych w podrozdziale 3.2, w celu klasyfikacji binarnej II stopnia. Przy założeniu przemieszczania się za pomocą auta, na podstawie cech charakterystycznych zachowań poszczególnych statystyk, postaramy się wyróżnić zmienne mogące mieć istotny wpływ na jakość budowanego modelu klasyfikacji terenu (teren zabudowany lub niezabudowany).

Rysunek 3.7 przedstawia rozkład prędkości podczas przemieszczania się autem w terenie zabudowanym oraz niezabudowanym. Mediana prędkości v_{GPS} dla terenu zabudowanego wynosi mniej niż 30 km/h, natomiast w terenie niezabudowanym około 100 km/h. Różnica w prędkościach wynika głównie z faktu istniejących, drogowych ograniczeń prędkości na poszczególnych odcinkach trasy (teren zabudowany

- ograniczenie do 50 km/h, teren niezabudowany - ograniczenie do 90 km/h). Możemy zatem przypuszczać, że zmienna v_{GPS} lub jej sfaktoryzowana postać (zmienna $F(v_{GPS})$) będą miały istotny wpływ na jakość klasyfikacji II stopnia.



Rysunek 3.7: Wykres rozkładu prędkości v_{GPS} .

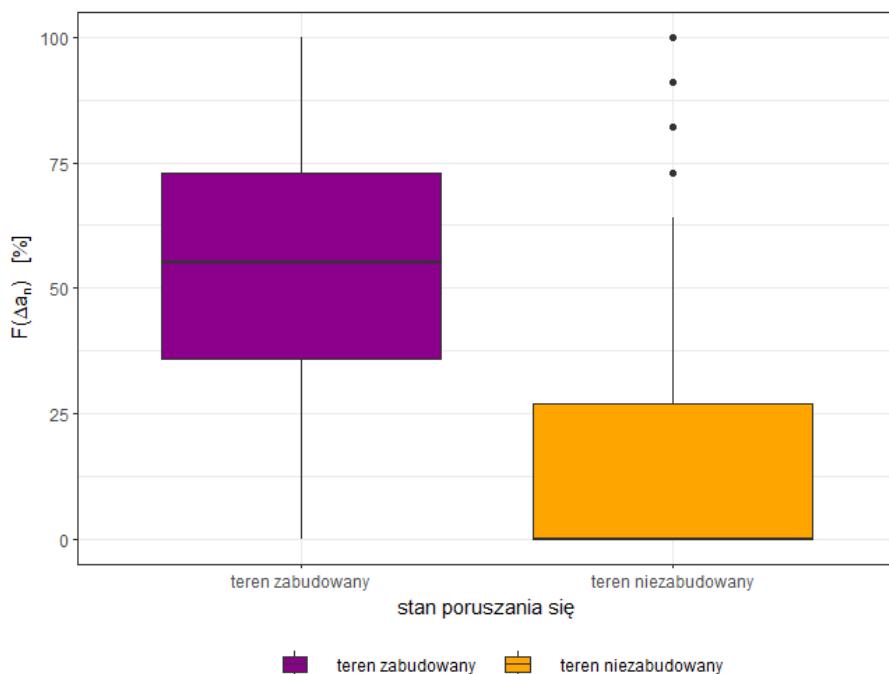


Rysunek 3.8: Wykres rozkładu widocznych przez odbiornik satelitów N_{GPS} .

Rysunek 3.8 przedstawia rozkład widocznych przez odbiornik satelitów zarejestrowanych przez odbiornik GPS podczas przemieszczania się autem w terenie zabudowanym i niezabudowanym. Ze względu na otwartą przestrzeń terenu w obu przypadkach, liczba widocznych satelitów N_{GPS} jest do siebie zbliżona (utrzymuje się na po-

ziomie 8-10 satelitów), jednak podczas przemieszczania się w terenie zabudowanym pojawiają się wartości odstające, wynikające naprawdopodobnie z przemieszczania się pomiędzy wysokimi budynkami, ograniczającymi widoczność nieba.

Rysunek 3.9 obrazuje rozkład zmiennej $F(\Delta a_n)$ podczas przemieszczania się autem w terenie zabudowanym i poza nim. Można zauważyć, że procent zmiany azymutu o więcej jak 15° jest większy w przypadku poruszania się w terenie zabudowanym, niż niezabudowanym. Mediana dla wartości statystyki $F(\Delta a_n)$ podczas przemieszczania się w terenie zabudowanym wynosi ponad 50%, podczas gdy dla terenu niezabudowanego wartość mediany nie przekracza 25%.

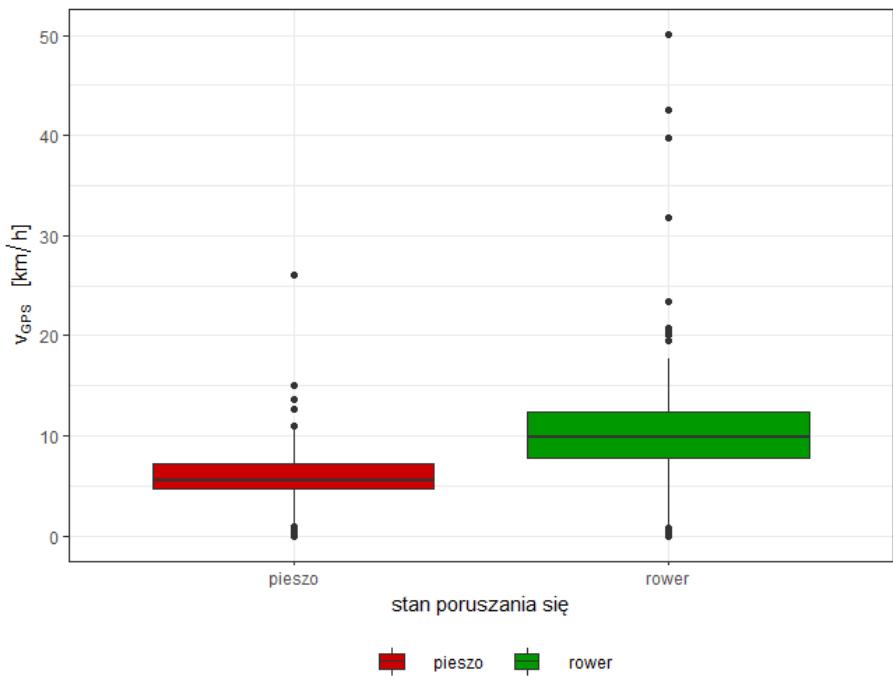


Rysunek 3.9: Wykres rozkładu procentowej zmiany azymutu $F(\Delta a_n)$ o więcej jak 15° .

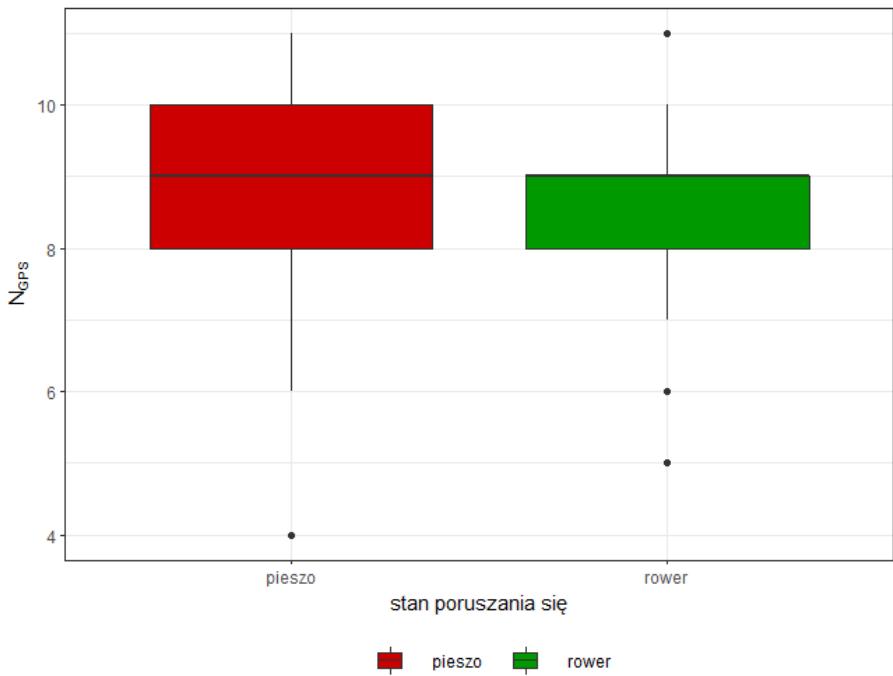
Klasyfikacja II stopnia sposobu przemieszczania się środkiem innym niż auto

Przyjrzyjmy się bliżej klasyfikacji binarnej II stopnia sposobu przemieszczania się środkiem innym niż auto. Wyróżniono tutaj przemieszczanie się użytkownika z lokalizatorem GPS rowerem lub pieszo (głównie w terenie zabudowanym).

Rysunek 3.10 przedstawia rozkład zmiennej v_{GPS} podczas przemieszczania się osoby z lokalizatorem GPS pieszo i podczas jazdy rowerem. Nieco wyższe prędkości osiągane są podczas jazdy rowerem, przy czym warto zaznaczyć, że w celach testowych przemieszczanie się rowerem miało charakter rekreacyjny. Wartości odstające mogą wynikać ze specyfiki działania samego urządzenia GPS, gdzie w sytuacji osiągania niewielkich prędkości, odbiornik nie namierza swojej pozycji w sposób regularny (ciągły).

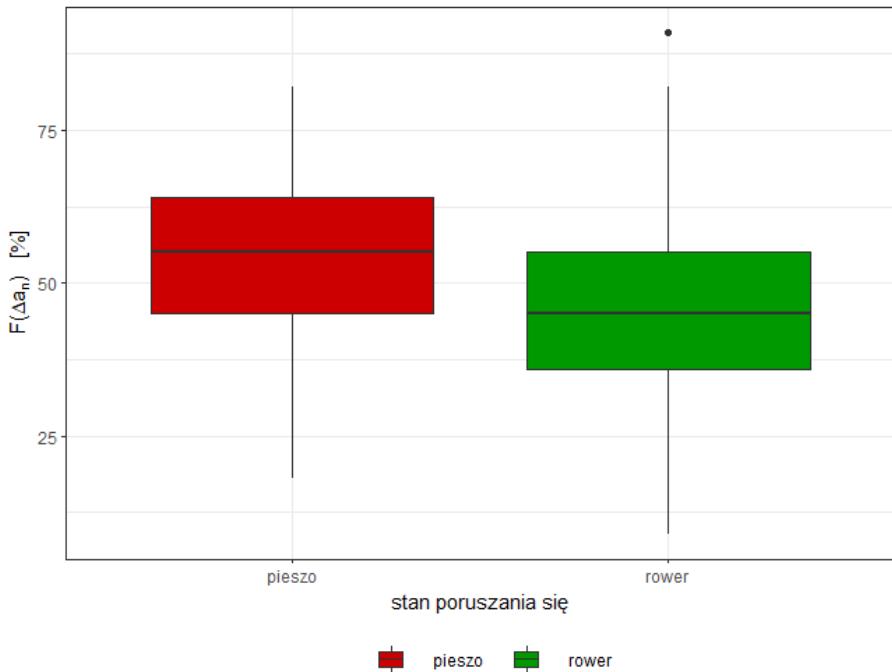


Rysunek 3.10: Wykres rozkładu przedkości v_{GPS} .



Rysunek 3.11: Wykres rozkładu widocznych przez odbiornik satelitów N_{GPS} .

Rysunek 3.11 przedstawia rozkład zmiennej N_{GPS} w sytuacji przemieszczania się pieszo oraz jazdy rowerem. Liczba satelitów w obu przypadkach utrzymuje się na poziomie 8-9 satelitów. Wartości odstające w obu przypadkach ruchu mogą pojawiać się na skutek przemieszczania się w specyficznych wówczas warunkach. W przypadku ruchu pieszego był to teren zabudowany (ograniczony wysokimi budynkami ograniczającymi widoczność nieba), natomiast w przypadku jazdy rowerem, był to najczęściej las, gdzie widoczność satelitów również może być ograniczona.



Rysunek 3.12: Wykres rozkładu procentowej zmiany azymutu $F(\Delta a_n)$ o więcej jak 15° .

Wykres 3.12 obrazuje rozkład zmiennej $F(\Delta a_n)$ w przypadku przemieszczania się osoby z odbiornikiem GPS pieszo oraz jazdy rowerem. Z wykresu możemy wnioskować, że przemieszczanie się pieszo generuje większe zmiany azymutu niż jazda rowerem. Fakt ten wynikać może ze zmniejszonej płynności ruchu podczas przemieszczania się pieszo, w porównaniu do jazdy rowerem. Możemy zatem przypuszczać, że współczynnik $F(\Delta a_n)$ może mieć istotny wpływ na późniejszą klasyfikację sposobu przemieszczania się osoby z lokalizatorem GPS.

Podsumowując, zróżnicowane zachowania statystyk prędkości v_{GPS} , liczby widocznych satelitów N_{GPS} oraz procentowej wartości zmiany azymutu $F(\Delta a_n)$ w różnych stanach poruszania się, mogą świadczyć o dalszej sensowności oraz istotności tych zmiennych w budowaniu i testowaniu modeli klasyfikacji metodą regresji.

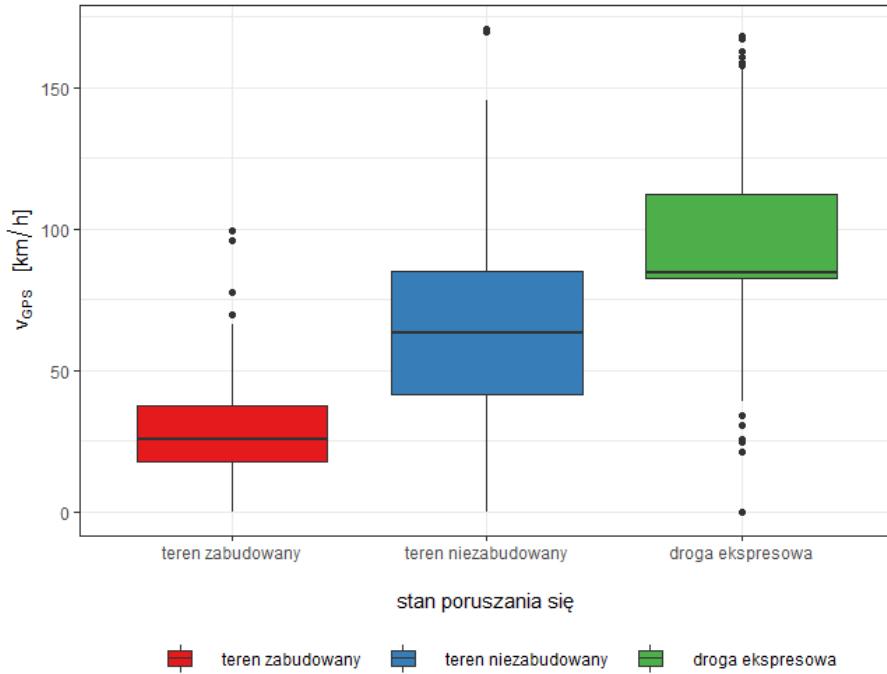
3.3.2 Klasyfikacja w przypadku trzech klas (model 2)

Klasyfikacja rodzaju drogi przemieszczania się autem

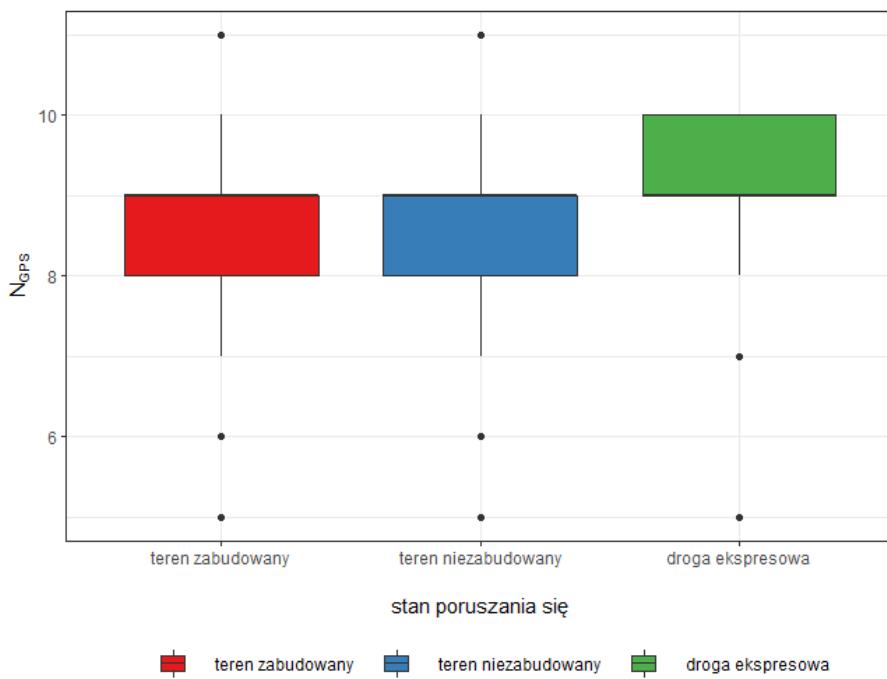
W nieniejszym podrozdziale pod względem wzęto zachowanie statystyk opisujących działanie odbiornika GPS w trzech stanach poruszania się autem. Są nimi przemieszczanie się po drodze ekspresowej, w terenie zabudowanym oraz niezabudowanym. Celem zadania jak poprzednio, jest wyróżnienie zmiennych, mogących mieć istotny wpływ na przeprowadzaną klasyfikację.

Rysunek 3.13 przedstawia rozkład prędkości v_{GPS} podczas przemieszczania się autem w trzech poszczególnych terenach. Medianą prędkości w terenie zabudowanym jest na poziomie około 25 km/h, w terenie niezabudowanym najczęściej osiąganą prędkością jest wartość zbliżona do 70 km/h, natomiast w przypadku poruszania się po drodze ekspresowej jest to wielkość na poziomie bliskim 90 km/h. Znaczące różnice w rozkładzie zmiennej v_{GPS} wynikają głównie z istniejących drogowych ograniczeń prędkości w poszczególnych stanach poruszania się (teren zabudowany - 50 km/h, teren niezabudowany - 90 km/h, droga ekspresowa - 120 km/h). Ponadto

wartości odstające pojawiające się przy poruszaniu się drogą ekspresową mogą pojawiać się na skutek nagłego hamowania, lub w przypadku terenu zabudowanego - mało płynnej jazdy autem podczas stania w korku lub podczas jazdy przez wysokie osiedla, ograniczające widoczność satelitów, a co za tym idzie, uniemożliwiające sposób regularnego namierzania pozycji odbiornika. Rozkład prędkości widoczny na rysunku 3.13 pokazuje, że budowany model klasyfikacji w przypadku trzech klas, może być zdeterminowany przez zmienną v_{GPS} .

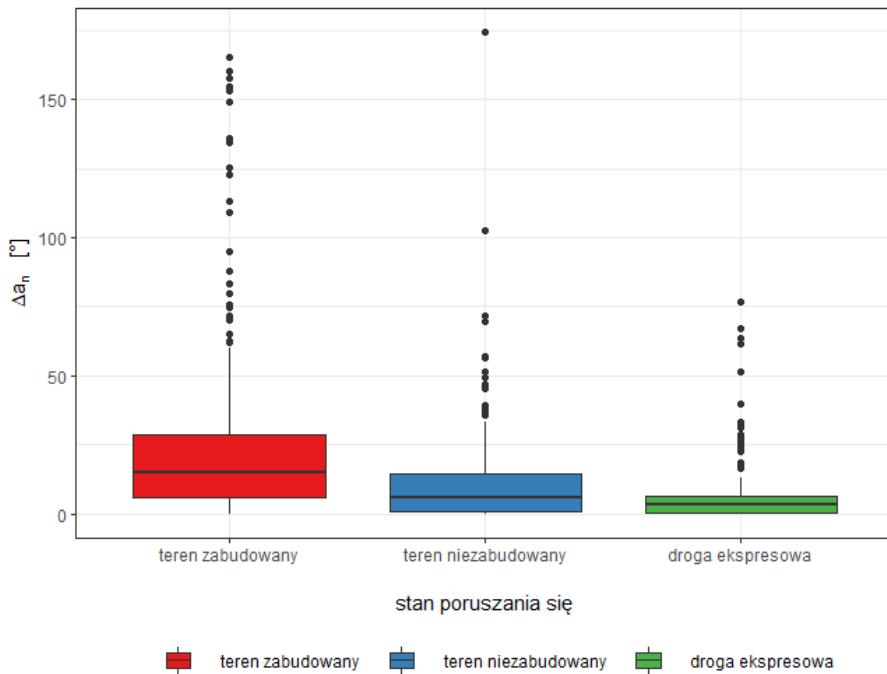


Rysunek 3.13: Wykres rozkładu prędkości v_{GPS} .



Rysunek 3.14: Wykres rozkładu widocznych przez odbiornik satelitów N_{GPS} .

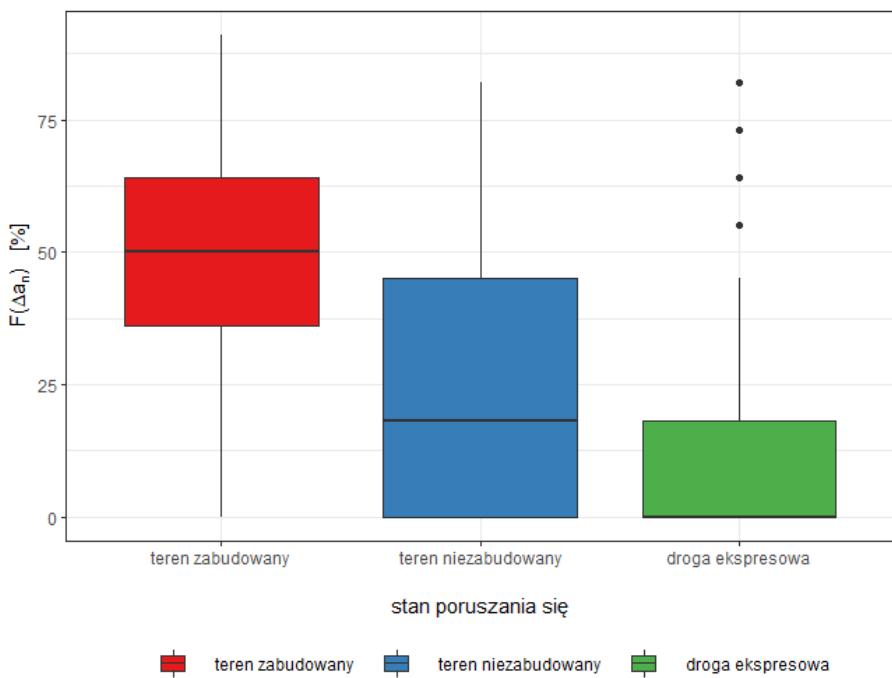
Rysunek 3.14 przedstawia rozkład widocznych przez odbiornik satelitów w trzech poszczególnych terenach przemieszczania się autem. Dla wszystkich stanów wartość czynnika N_{GPS} utrzymuje się na wysokim poziomie i osiąga wartości 8–10 satelitów. Wartości odstające pojawiające się w przypadku poruszania się w terenie zabudowanym, mogą mieć związek ze wspomnianym wcześniej problemem przemieszczania się po mieście pomiędzy wysokimi budynkami, niekiedy ograniczającymi widoczność nieba.



Rysunek 3.15: Wykres rozkładu zmian azymutu Δa_n .

Wykres 3.15 przedstawia rozkład zmiennej Δa_n w trzech poszczególnych terenach przemieszczania się autem. Najmniejsze zmiany wartości azymutu osiągane są podczas poruszania się drogą ekspresową, największe natomiast podczas jazdy autem w terenie zabudowanym. Współczynnik zmiany azymutu Δa_n związany jest ściśle z osiąganą prędkością v_{GPS} . Zauważono, że im szybciej użytkownik się porusza, tym jego pozycja jest częściej wyznaczana (w sposób bardziej regularny), co niesie za sobą mniej gwałtowne zmiany wartości azymutu.

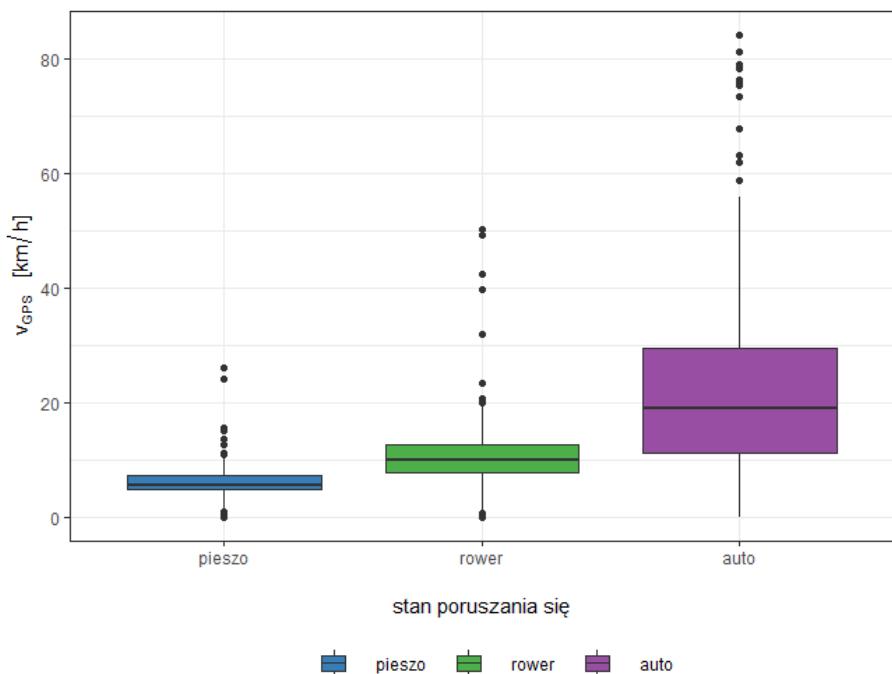
Rysunek 3.16 przedstawia rozkład zmiennej $F(\Delta a_n)$ dla trzech poszczególnych terenów przemieszczania się autem. W tym przypadku również widoczne są znaczące różnice w rozkładzie częstotliwości zmiany azymutu dla poszczególnych stanów poruszania się. Procent zmiany azymutu o więcej jak 15° w przypadku jazdy w terenie zabudowanym osiąga wartości nawet ponad 50%. W przypadku poruszania się w terenie niezabudowanym, wahania współczynnika $F(\Delta a_n)$ są największe, jednak nie przekraczają progu 50%. Podczas przemieszczania się drogą ekspresową procent wystąpień zmian azymutu o więcej jak 15° jest najmniejszy, co wynika głównie z najbardziej regularnej pracy urządzenia GPS, podczas osiągania większych prędkości. Zróżnicowanie wartości czynnika $F(\Delta a_n)$ w poszczególnych terenach przemieszczania, może świadczyć o większej istotności tego współczynnika przy badaniu jakości budowanego modelu klasyfikacji.



Rysunek 3.16: Wykres rozkładu procentowej zmiany azymutu $F(\Delta a_n)$ o więcej jak 15° .

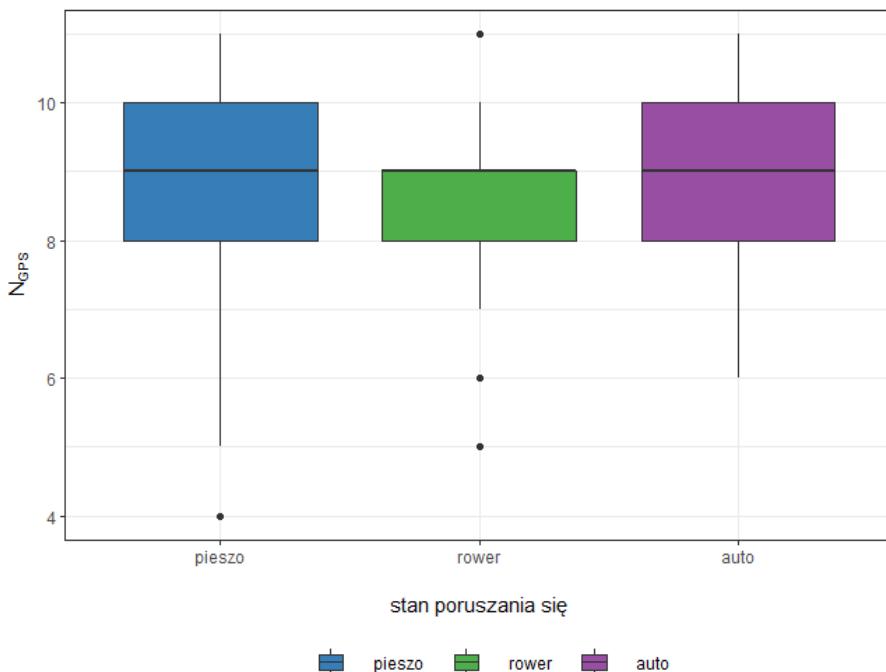
Klasyfikacja sposobu przemieszczania się osoby

Poniższe zestawienia zawierają analizę rozkładu czynników mogących mieć istotny wpływ na budowany model klasyfikacji, dotyczący sposobu przemieszczania się osoby w terenie zabudowanych, w rozróżnieniu na trzy klasy: poruszanie się autem, rowerem lub pieszo.



Rysunek 3.17: Wykres rozkładu prędkości v_{GPS} .

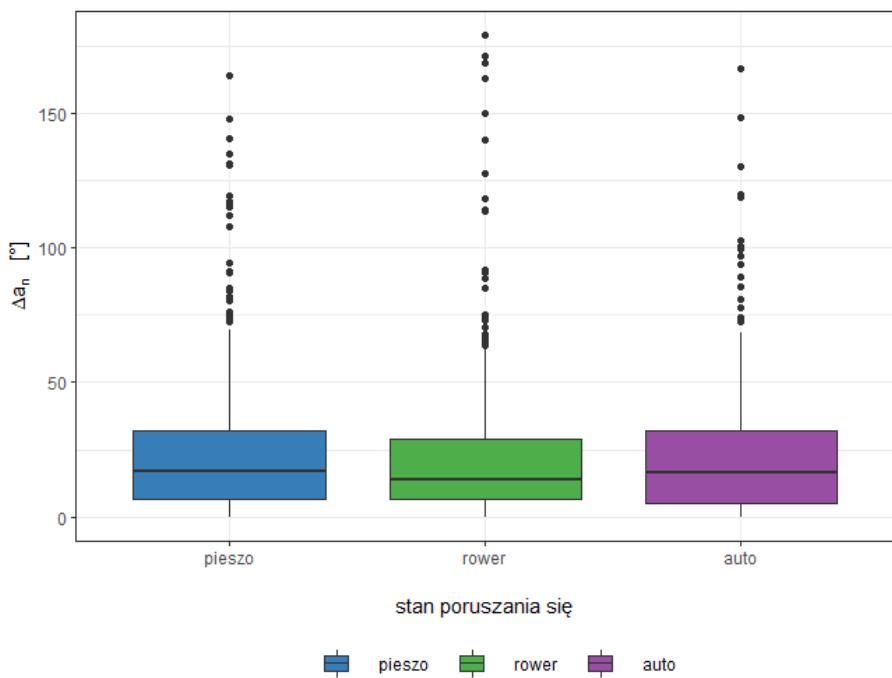
Rysunek 3.17 przedstawia rozkład zmiennej v_{GPS} podczas przemieszczania się osoby w terenie zabudowanym pieszo, rowerem oraz autem. Rozkład prędkości jest zgodny z intuicją, to znaczy, największe prędkości osiągana są podczas przemieszczania się autem (do 50 km/h), a najmniejsze podczas ruchu pieszego (najczęściej jest to około prędkość około 6 km/h). Widoczne na wykresie wartości odstające w każdym z trzech przypadków są charakterystyczne dla przemieszczania się osoby z lokalizatorem w terenie zabudowanym. Mamy wówczas do czynienia z częstymi przeszkodami ograniczającymi widoczność nieba (wysokie budynki, drzewa) lub w przypadku poruszania się autem - nieregularną jazdą spowodowaną nagłym przyspieszaniem lub hamowaniem (przed pasami lub sygnalizacją świetlną). Warto wspomnieć, że dane pomiarowe dla niniejszej klasyfikacji, były zbierane podczas weekendu (sobota/niedziela), gdzie wówczas zator uliczny był mniejszy niż w zwyczajne popołudnie w środku tygodnia.



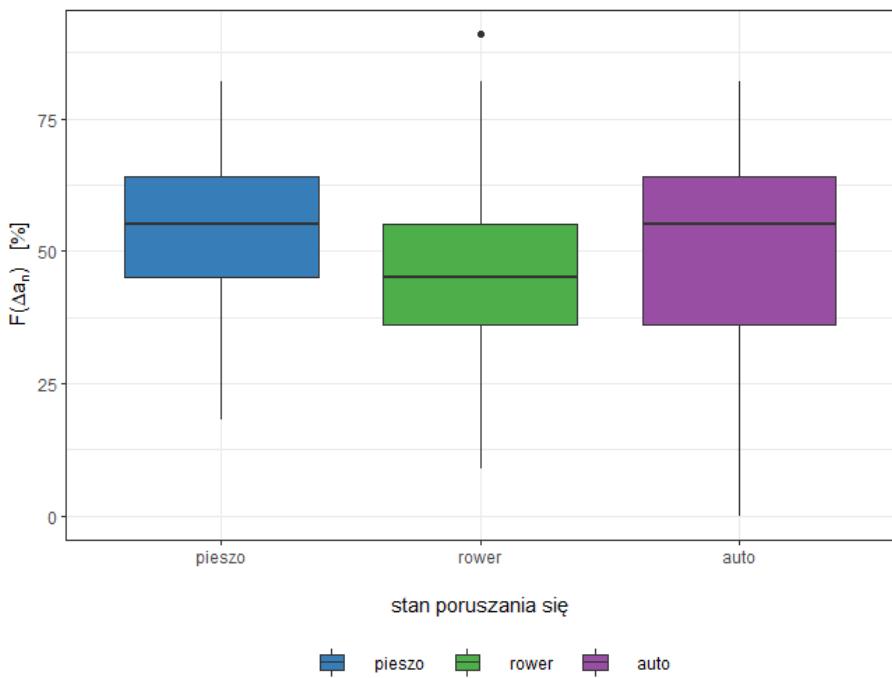
Rysunek 3.18: Wykres rozkładu widocznych przez odbiornik satelitów N_{GPS} .

Na rysunku 3.18 pokazano rozkład widocznych przez odbiornik satelitów podczas przemieszczania się osoby w terenie zabudowanym pieszo, rowerem oraz autem. Mediana współczynnika N_{GPS} dla trzech wyróżnionych sposobów poruszania się osiąga wartość 9. Powyższy wykres nie wskazuje na znaczące różnice w rozkładzie zmiennej N_{GPS} podczas poruszania się pieszo, rowerem czy też autem. Możemy zatem wnioskować, że liczba satelitów w przypadku takiej klasyfikacji, nie będzie miała istotnego wpływu na model.

Rysunek 3.19 przedstawia rozkład zmiany azymutu podczas przemieszczania się osoby w terenie zabudowanym pieszo, rowerem oraz autem. Podobnie jak w przypadku rozkładu liczby widocznych satelitów, tak i tutaj wartości zmiennej Δa_n dla poszczególnych sposobów przemieszczania się, osiągają zblżone wartości oraz zauważalna jest liczna grupa wartości odstających statystyki Δa_n . Można zatem wnioskować, że zmienna Δa_n , nie wpłynie istotnie na jakość modelu klasyfikacji danych.



Rysunek 3.19: Wykres rozkładu zmian azymutu Δa_n .



Rysunek 3.20: Wykres rozkładu procentowej zmiany azymutu $F_{\Delta a_n}$ o więcej jak 15° .

Rysunek 3.20 przedstawia rozkład współczynnika $F(\Delta a_n)$ podczas przemieszczania się osoby w terenie zabudowanym pieszo, rowerem lub autem. Mediana procentowej wartości zmiany azymutu jest najniższa podczas jazdy rowerem i osiąga wartość na poziomie 40%. Największe wahania zmiennej $F(\Delta a_n)$ zauważalne są w przypadku przemieszczania się w terenie zabudowanym autem, gdzie 25% obserwacji osiąga wartość procentowej zmiany azymutu większą lub równą około 35%, natomiast 75% obserwacji osiąga wartości mniejsze lub równe 65%.

Podsumowując, możemy przypuszczać, że zróżnicowane zachowania parametrów takich jak: prędkość v_{GPS} , liczba widocznych przez odbiornik satelitów N_{GPS} , zmiana azymutu Δa_n czy procent zmiany azymutu o więcej jak 15° wpłyną istotnie na proces klasyfikacji rozpoznawania poszczególnych stanów poruszania się. Statystyki takie jak zmiana czasu Δt_n czy dystans d_{GPS} nie zostały zestawione w powyższych analizach ze względu na brak znaczących wielkości tych parametrów podczas testowania poszczególnych stanów przemieszczania się osoby z lokalizatorem.

Rozdział 4

Proces klasyfikacji danych z odbiornika GPS

4.1 Klasyfikacja w przypadku dwóch klas (model 1)

4.1.1 Klasyfikacja binarna I stopnia

Dane wejściowe zawierają 1250 obserwacji, z rozróżnieniem na 597 rekordów przemieszczania się autem oraz 653 rekordów poruszania się w inny sposób. Dane zostały podzielone w sposób losowy w proporcji odpowiednio 80% i 20% na zbiór uczący oraz testowy.

Uwzględniając analizy parametrów danych opisanych w podrozdziale 3.3.1, do budowy modelu klasyfikacji posłużono się następującymi zmiennymi:

- a) prędkość v_{GPS} ,
- b) liczba widocznych przez odbiornik satelitów N_{GPS} ,
- c) procentowa wartość zmiany azymutu w oknie czasowym $F(\Delta a_n)$.

Model klasyfikacji został zbudowany, a następnie przetestowany przy wykorzystaniu środowiska R oraz odpowiednich pakietów i funkcji, takich jak *stats::glm*, *predict* [14] [16] czy *InformationValue::optimalCutoff*, *misClassError*. Model wytrenowano na podstawie zestawu danych uczących, natomiast jego jakość weryfikowano na podstawie zbioru testowego.

Tabela 4.1 zawiera podsumowanie zmiennych wchodzących w skład modelu. Kolumna $\hat{\beta}_i$ określa wyestymowane współczynniki wektora β_i , kolumna o nazwie $se(\hat{\beta}_i)$ zawiera wielkości błędów standardowych dla wyestymowanych współczynników β_i , natomiast ostatnia kolumna informuje o wyniku testu, dotyczącego istotności poszczególnych zmiennych wchodzących w skład modelu (tj. o odrzuceniu hipotezy zerowej o nieistotności zmiennych na poziomie istotności równym odpowiednio 0.01 (***)¹, 0.05 (**) oraz 0.1 (*)) [5]. Innymi słowy, im mniejsza wartość testu, tym większa istotność badanej zmiennej, wchodzącej w skład modelu.

	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	statystyka z	$p(\hat{\beta}_i \neq 0)$
<i>const</i>	7.291	0.995	7.322	$2.44e-13^{***}$
N_{GPS}	-0.270	0.091	-2.952	0.003**
v_{GPS}	-0.819	0.059	-13.737	< 2e-16***
$F(\Delta a_n)$	-0.029	0.005	-4.953	$7.29e-07^{***}$

Tabela 4.1: Podsumowanie zmiennych modelu regresji dla klasyfikacji binarnej I stopnia.

W oparciu o zależność określoną wzorem 1.7, model klasyfikacji I stopnia można zapisać w postaci

$$\hat{y} = \ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 \cdot N_{GPS} + \beta_2 \cdot v_{GPS} + \beta_3 \cdot F(\Delta a_n), \quad (4.1)$$

gdzie p oznacza prawdopodobieństwo, że użytkownik posiadający urządzenie GPS, przemieszcza się autem. Współczynniki $\beta_0, \beta_1, \beta_2, \beta_3$ są parametrami bezpośrednio wpływającymi na specyfikę otrzymanego modelu. Uwzględniając informacje zawarte w powyższej tabeli, model klasyfikacji można zapisać w postaci

$$\hat{y} = \ln \left(\frac{p}{1-p} \right) \approx 7.29 - 0.27 \cdot N_{GPS} - 0.82 \cdot v_{GPS} - 0.03 \cdot F(\Delta a_n), \quad (4.2)$$

gdzie prawdopodobieństwo p można wyrazić przez

$$p = \frac{\exp(\hat{y})}{1 + \exp(\hat{y})} \approx \frac{\exp(7.29 - 0.27 \cdot N_{GPS} - 0.82 \cdot v_{GPS} - 0.03 \cdot F(\Delta a_n))}{1 + \exp(7.29 - 0.27 \cdot N_{GPS} - 0.82 \cdot v_{GPS} - 0.03 \cdot F(\Delta a_n))}. \quad (4.3)$$

Model należy interpretować następująco: wraz ze wzrostem współczynnika liczby widocznych satelitów N_{GPS} , wartość funkcji $\ln(p/(1-p))$ zmniejsza się każdorazowo o 0.27. Podobnie, zwiększając wartość zmiennej v_{GPS} , wartość funkcji \hat{y} zmniejsza się o 0.82 jednostek, natomiast zwiększając wartość parametru $F(\Delta a_n)$, szansa na sukces jest pomniejszona każdorazowo o wartość 0.03.

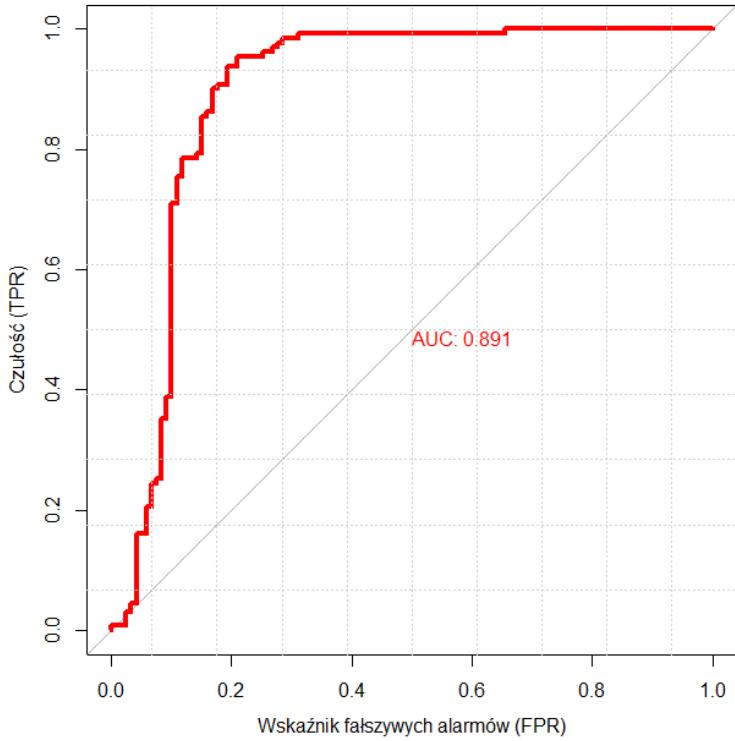
Uwzględniając zbudowany model oraz predykcję na zbiorze testowym, przy użyciu odpowiednich funkcji, możliwe stało się określenie tak zwanego optymalnego punktu odcięcia $x_{cut} \approx 0.49$. Macierz pomyłek dla modelu klasyfikacji I stopnia, przy obliczonej wartości x_{cut} , ma postać

	jadę	nie jadę
jadę	94	25
nie jadę	6	125

Tabela 4.2: Macierz rozkładu klas dla klasyfikacji binarnej I stopnia.

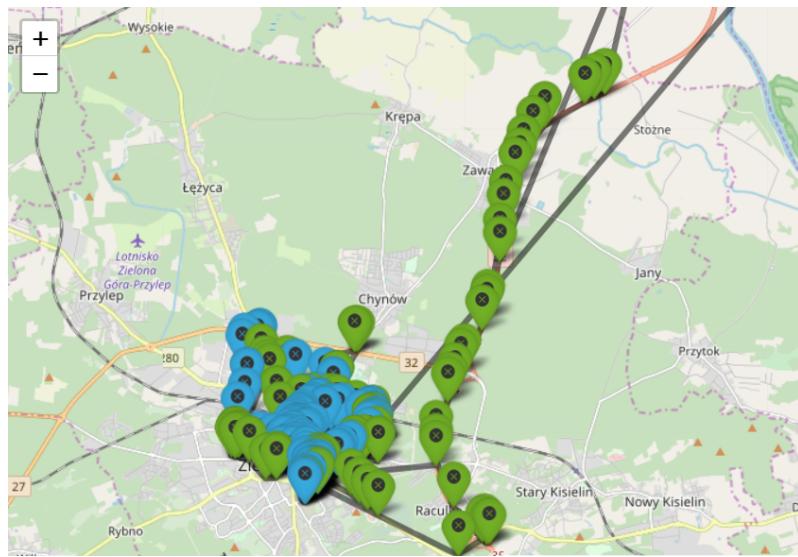
Zgodnie z oznaczeniami przyjętymi w podrozdziale 2.1, wiersze tabeli 4.2 reprezentują liczbę wzorców pochodzących z faktycznych klas zbioru testowego, elementy kolumn wskazują natomiast liczbę wzorców wyestymowanych przez klasyfikator jako dana klasa. Elementy leżące poza główną przekątną są uznawane za niepoprawne przyporządkowanie obserwacji do klasy. Błąd klasyfikacji wynosi $\delta_w \approx 12\%$.

Rysunek 4.1 przedstawia krzywą ROC wykreoloną przez wiele różnych punktów odcięcia. Pole powierzchni pod utworzoną krzywą wynosi $AUC \approx 89$, co świadczy o dobrej jakości klasyfikacji.



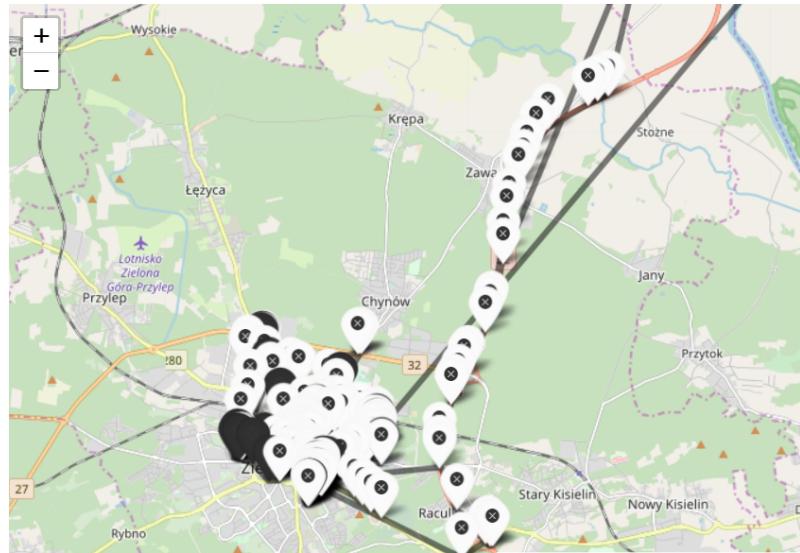
Rysunek 4.1: Wykres krzywej ROC dla klasyfikacji I stopnia.

Rysunek 4.2 obrazuje fragment zbioru testowego przed zastosowaniem algorytmu klasyfikacji. Znaczniki koloru zielonego określają przemieszczanie się autem, koloru niebieskiego natomiast, inny sposób poruszania się.



Rysunek 4.2: Fragment zbioru testowego przed zastosowaniem algorytmu klasyfikacji binarnej I stopnia, źródło: *opracowanie własne*.

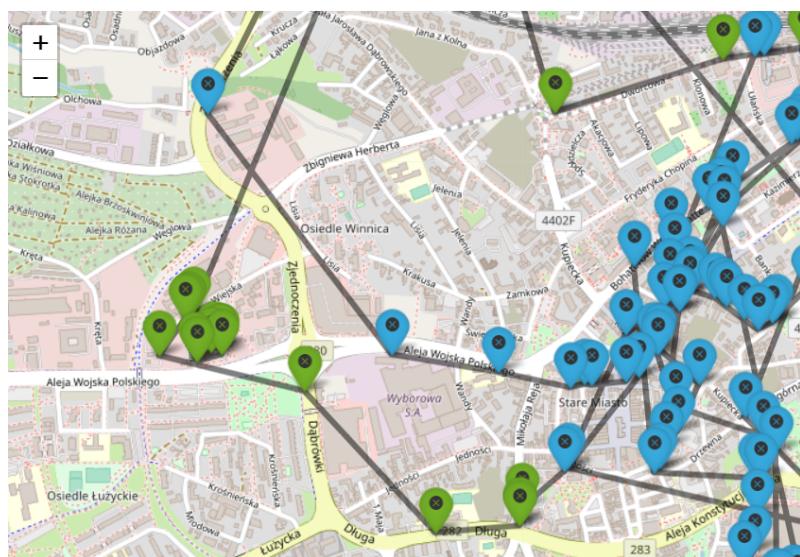
Rysunek 4.3 przedstawia graficzny wynik klasyfikacji określający przemieszczanie się autem lub jego brak. Znaczniki koloru białego określają poprawną klasyfikację obserwacji do klasy, czarne natomiast, błędne przyporządkowanie obserwacji do danej klasy.



Rysunek 4.3: Fragment wyniku klasyfikacji binarnej I stopnia, źródło: *opracowanie własne*.

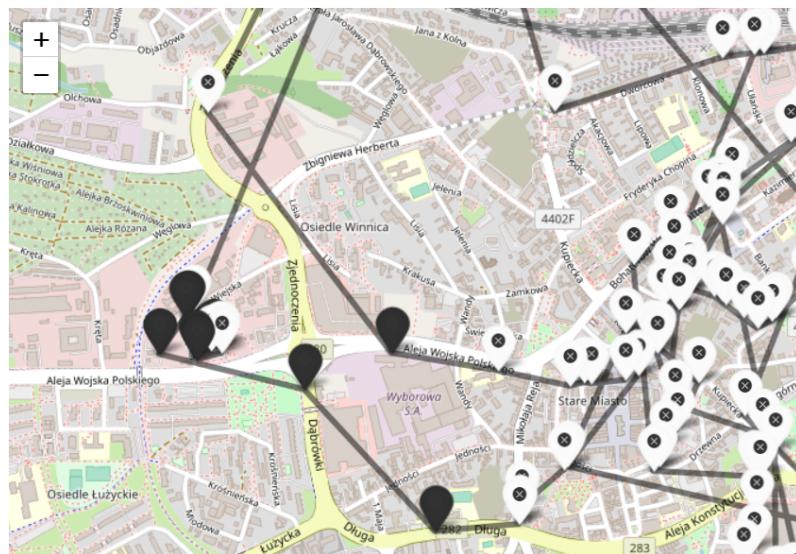
Porównując rysunki 4.2 oraz 4.3, proces klasyfikacji wydaje się być całkiem udany. Przyjrzyjmy się jednak bliżej wynikom klasyfikacji w mieście, gdzie ze względu na istniejące ograniczenie prędkości, sygnalizację świetlną oraz częste występowanie korków ulicznych, klasyfikacja stanu przemieszczania się samochodem lub jego brakiem może nie być trywialna.

Rysunek 4.4 przedstawia fragment zbioru danych testowych użytych podczas przemieszczania się użytkownika z lokalizatorem GPS w terenie zabudowanym. Dla przy pomnienia, znaczniki koloru zielonego definiują poruszanie się autem, niebieskiego natomiast, przemieszczanie się w inny sposób.



Rysunek 4.4: Fragment zbioru testowego w terenie zabudowanym, przed zastosowaniem algorytmu klasyfikacji binarnej I stopnia, źródło: *opracowanie własne*.

Rysunek 4.5 przedstawia graficzny wynik klasyfikacji zastosowanego algorytmu regresji w terenie zabudowanym. Na poniższym rysunku widać skupienie znaczników koloru czarnego, oznaczających błędą klasyfikację obserwacji. W tej sytuacji klasyfikator niepoprawnie przyporządkował obserwacje jako przemieszczanie się w sposób inny niż jazda samochodem, gdzie faktycznie użytkownik poruszał się wówczas samochodem. Warto zauważyć, że sytuacja ta ma miejsce w okolicy wjazdu i wyjazdu samochodu pomiędzy wysokie osiedla. Sygnał GPS docierający do odbiornika może być wówczas utrudniony, a namierzenie pozycji użytkownika może zostać opóźnione. W ten sposób nieprawidłowości występujące w parametrach opisanych w podrozdziale 3.2, mające bezpośredni wpływ na zbudowany model klasyfikacji mogą generować nieuniknione błędy.



Rysunek 4.5: Fragment wyniku klasyfikacji binarnej I stopnia w terenie zabudowanym, źródło: *opracowanie własne*.

4.1.2 Klasyfikacja binarna II stopnia

Do klasyfikacji miejsca przemieszczania się autem (teren zabudowany lub niezabudowany), jako zbiór danych wejściowych wybrano obserwacje wykorzystane do klasyfikacji I stopnia, jednak uwzględniając jedynie te rekordy, gdzie faktycznym stanem poruszania się była jazda autem. Zbiór danych wejściowych, po wstępny czyszczeniu, składał się z 665 rekordów, z rozróżnieniem na 473 rekordów odpowiadających przemieszczaniu się autem w terenie zabudowanym oraz 192 obserwacji odpowiadających poruszaniu się poza nim. Zbiór danych wejściowych został podzielony w sposób losowy na zbiór danych uczących i testowym, w proporcji odpowiednio 80% do 20%. Zbiór danych uczących zawierał odpowiednio 378 rekordów definiujących przemieszczanie się autem w terenie zabudowanym oraz 154 poza nim, natomiast zbiór danych testowych - odpowiednio 95 i 38 obserwacji w tych samych grupach.

Opierając się o analizę najbardziej charakterystycznych parametrów opisanych w podrozdziale 3.3.1, do modelu klasyfikacji terenu przemieszczania się autem wykorzystano informację o sfaktoryzowanej wartości prędkości $F(v_{GPS})$. Tabela 4.3 przedstawia wygenerowane podsumowanie modelu klasyfikacji.

	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	statystyka z	$p(\hat{\beta}_i \neq 0)$
$const$	2.493	0.187	13.36	$< 2e-16^{***}$
$F(v_{GPS})$	-2.905	0.243	-11.96	$< 2e-16^{***}$

Tabela 4.3: Podsumowanie zmiennych modelu regresji dla klasyfikacji miejsca przemieszczania się autem.

W oparciu o zależność określoną wzorem 1.7, model klasyfikacji II stopnia przemieszczania się autem w terenie zabudowanym lub poza nim, można zapisać w postaci

$$\hat{y} = \ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 \cdot F(v_{GPS}), \quad (4.4)$$

gdzie p oznacza prawdopodobieństwo, że użytkownik posiadający lokalizator, przemieszcza się autem w terenie zabudowanym. Uwzględniając tabelę podsumowania modelu, można zapisać go w postaci

$$\hat{y} = \ln \left(\frac{p}{1-p} \right) \approx 2.49 - 2.91 \cdot F(v_{GPS}), \quad (4.5)$$

gdzie prawdopodobieństwo p można wyrazić przez

$$p = \frac{\exp(\hat{y})}{1 + \exp(\hat{y})} \approx \frac{\exp(2.49 - 2.91 \cdot F(v_{GPS}))}{1 + \exp(2.49 - 2.91 \cdot F(v_{GPS}))}. \quad (4.6)$$

Model określony wzorem 4.5 posiada następującą interpretację. Wraz wzrostem współczynnika $F(v_{GPS})$, wartość funkcji $\ln(p/(1-p))$ maleje każdorazowo o 2.91. Optymalny punkt odcięcia dla klasyfikacji terenu przemieszczania się autem wynosi $x_{cut} \approx 0.04$. Po uwzględnieniu oszacowanej wartości parametru x_{cut} macierz rozkładu klas ma postać

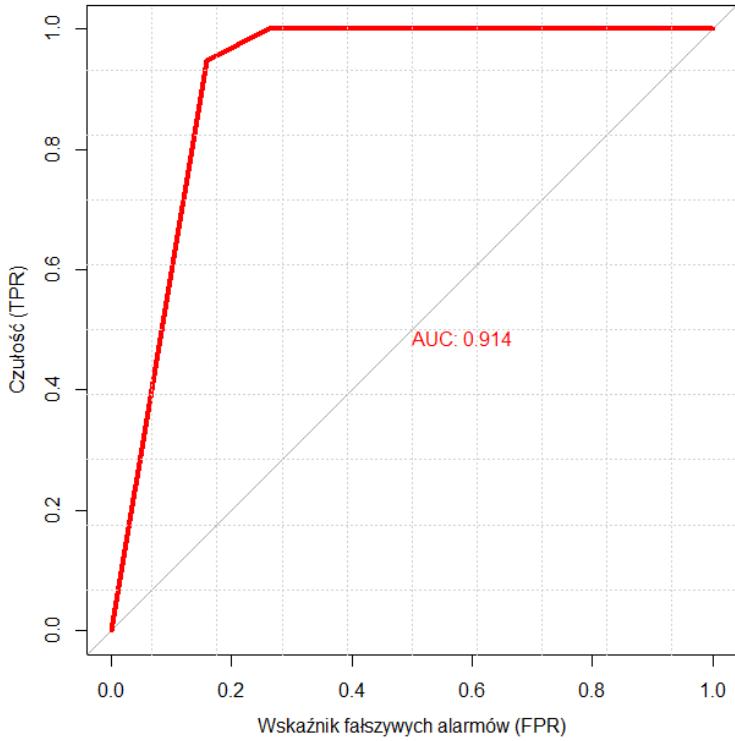
	teren niezabudowany	teren zabudowany
teren niezabudowany	28	10
teren zabudowany	0	95

Tabela 4.4: Macierz rozkładu klas dla klasyfikacji miejsca przemieszczania się autem.

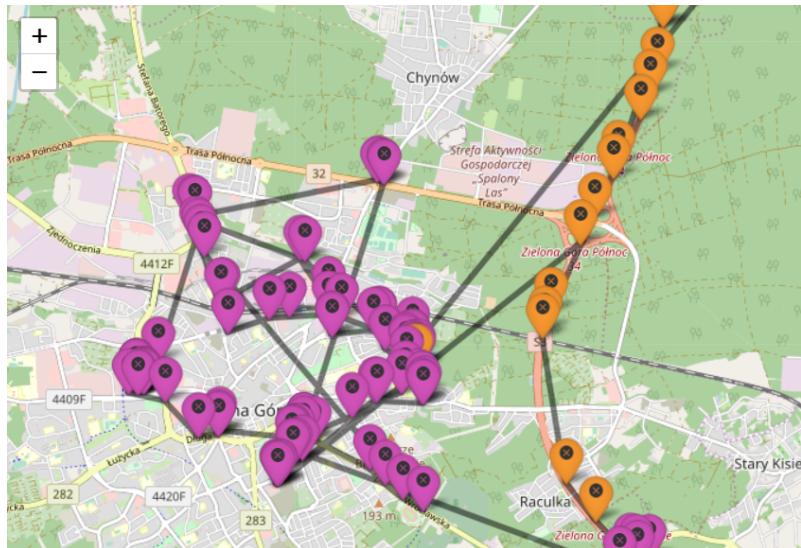
Na podstawie tabeli 4.4 widać, że niepoprawnie sklasyfikowano 10 rekordów jako ruch pojazdu w terenie zabudowanym, gdzie faktyczną klasą przynależności dla tych obserwacji był teren niezabudowany. Błąd klasyfikacji wynosi $\delta_w \approx 7\%$.

Rysunek 4.6 przedstawia wykres krzywej ROC wykreślony dla różnych punktów odcięcia. Pole powierzchni pod krzywą wynosi $AUC \approx 0.91$, co świadczy o bardzo dobrej jakości zbudowanego klasyfikatora.

Rysunek 4.7 przedstawia fragment zbioru danych testowych przed zastosowaniem klasyfikacji binarnej terenu przemieszczania się samochodem. Znaczniki koloru fioletowego oznaczają przemieszczanie się autem w terenie zabudowanym, pomarańczowego zaś, w terenie niezabudowanym.

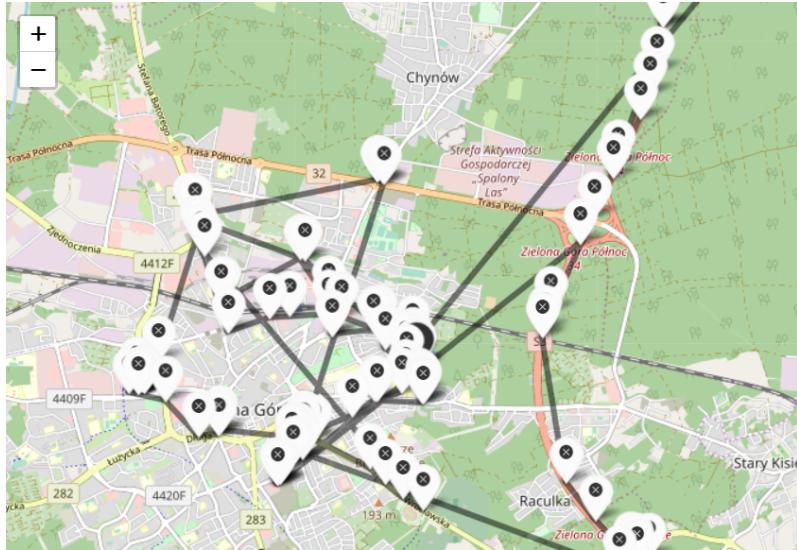


Rysunek 4.6: Wykres krzywej ROC dla klasyfikacji II stopnia. Przemieszczanie się autem w terenie zabudowanym lub niezabudowanym.



Rysunek 4.7: Fragment zbioru danych testowych przed zastosowaniem algorytmu klasyfikacji terenu przemieszczania się samochodem, źródło: *opracowanie własne*.

Rysunek 4.8 przedstawia wynik klasyfikacji terenu przemieszczania się samochodem. Porównując oba rysunki można stwierdzić, że poprawność klasyfikatora jest bardzo dobra. Wynikać to może z faktu bardzo wyraźnego podziału pomiędzy zachowaniami parametrów opisanych w podrozdziale 3.2 mających bezpośredni wpływ na zbudowany model lub z faktu nieproporcjonalności obserwacji danych wejściowych określających przemieszczanie się samochodem w terenie zabudowanym i niezabudowanym.



Rysunek 4.8: Fragment wyniku algorytmu klasyfikacji terenu przemieszczania się samochodem, źródło: *opracowanie własne*.

Do klasyfikacji binarnej sposobu przemieszczania się środkiem innym niż auto (rower lub pieszo), jako zbiór danych wejściowych wybrano obserwacje wykorzystane do klasyfikacji I stopnia, jednak uwzględniając jedynie te rekordy, gdzie nie doświadczyono jazdy autem. Po wstępny oczyszczaniu danych, zbiór składał się z 558 rekordów, z rozróżnieniem 336 rekordów poruszania się rowerem oraz 222 obserwacji przemieszczania się pieszo. Podobnie jak we wcześniejszych klasyfikacjach, zbiór danych wejściowych podzielono na zestaw danych treningowych i testowych, odpowiednio w stosunku 80% – 20%.

Mając na uwadze analizę parametrów charakterystycznych dla występowania ruchu pieszego i jazdę rowerem w podrozdziale 3.3.1, do modelu klasyfikacji wybrano zmienne najbardziej istotne

- a) prędkość v_{GPS} ,
- b) procentową zmianę wartości azymutu w oknie czasowym $F(\Delta a_n)$,
- c) liczbę widocznych przez odbiornik satelitów N_{GPS} .

Poniższa tabela przedstawia podsumowanie modelu klasyfikacji przemieszczania się rowerem lub pieszo.

	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	statystyka z	$p(\hat{\beta}_i \neq 0)$
$const$	-3.113	1.150	-2.706	0.007**
N_{GPS}	0.523	0.118	4.427	< 9.58e-06***
$F(\Delta a_n)$	0.024	0.008	3.111	0.002**
v_{GPS}	-1.443	0.161	-8.944	< 2e-16***

Tabela 4.5: Podsumowanie zmiennych modelu regresji dla klasyfikacji przemieszczania się rowerem lub pieszo.

W oparciu o zależność określoną wzorem 1.7, model klasyfikacji II stopnia przemieszczania się rowerem lub pieszo, można zapisać w postaci

$$\hat{y} = \ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 \cdot N_{GPS} + \beta_2 \cdot F(\Delta a_n) + \beta_3 \cdot v_{GPS}, \quad (4.7)$$

gdzie p oznacza prawdopodobieństwo, że użytkownik posiadający urządzenie GPS, przemieszcza się rowerem. Uwzględniając informacje zawarte w tabeli 4.5, model klasyfikacji można zapisać w postaci

$$\hat{y} = \ln \left(\frac{p}{1-p} \right) \approx -3.11 + 0.52 \cdot N_{GPS} + 0.02 \cdot F(\Delta a_n) - 1.44 \cdot v_{GPS}, \quad (4.8)$$

gdzie prawdopodobieństwo p można wyrazić przez

$$p = \frac{\exp(\hat{y})}{1 + \exp(\hat{y})} \approx \frac{\exp(-3.11 + 0.52 \cdot N_{GPS} + 0.02 \cdot v_{GPS} - 1.44 \cdot F(\Delta a_n))}{1 + \exp(-3.11 + 0.52 \cdot N_{GPS} + 0.02 \cdot v_{GPS} - 1.44 \cdot F(\Delta a_n))}. \quad (4.9)$$

Model można interpretować następująco. Zwiększając wartość liczby widocznych przez odbiornik satelitów N_{GPS} , każdorazowo zwiększeniu ulega wartość funkcji $\ln(p/(1-p))$ o 0.52 jednostki. Ponadto, zwiększając wartość parametru $F(\Delta a_n)$, funkcja \hat{y} zwiększa swoją wartość o 0.02. Zmienna określająca prędkość v_{GPS} ma natomiast ujemny wpływ na rozważany model. Dla każdej jednostki zwiększającej parametr v_{GPS} , szansa na sukces maleje każdorazowo o wartość 1.44.

Optymalny punkt odcięcia dla klasyfikacji przemieszczania się rowerem lub pieszo wynosi $x_{cut} \approx 0.52$. Na podstawie wyznaczonego x_{cut} oraz wartości predykcji na zbiorze testowym, możliwe stało się zbudowanie macierzy pomyłek

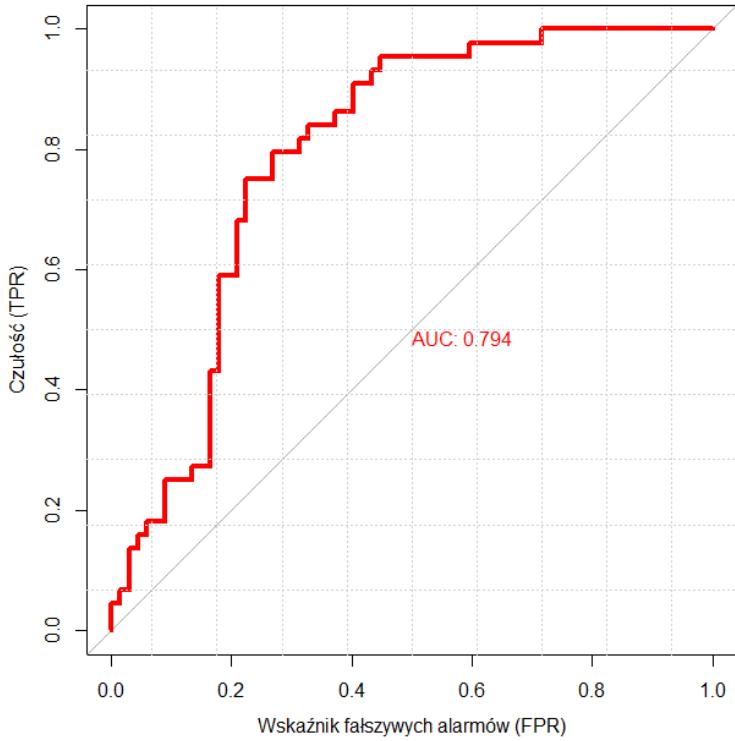
	rower	pieszo
rower	52	15
pieszo	11	33

Tabela 4.6: Macierz rozkładu klas dla klasyfikacji przemieszczania się rowerem lub pieszo.

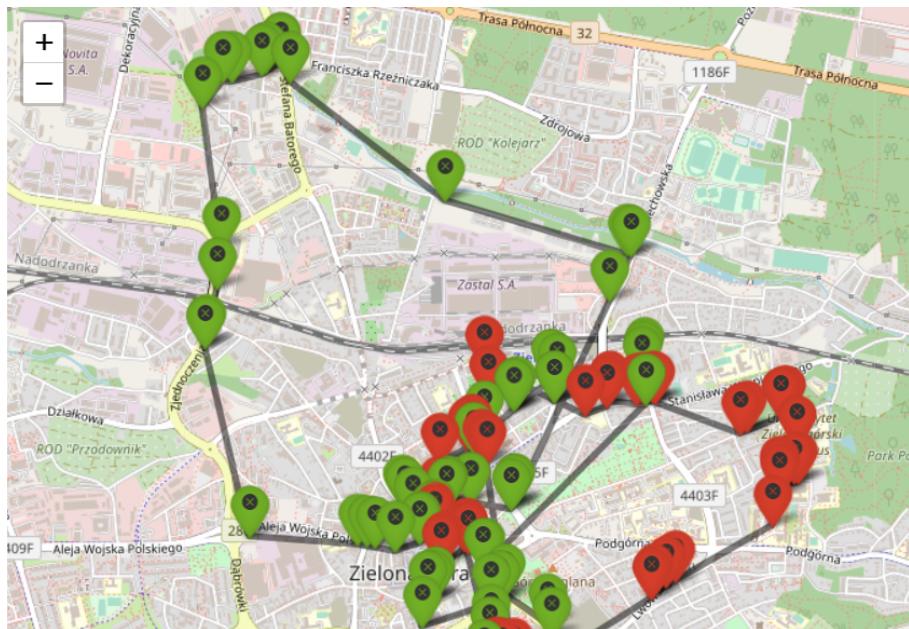
Na podstawie tabeli 4.6 widać, że 11 obserwacji zostało niepoprawnie sklasyfikowanych jako poruszanie się rowerem oraz 15 rekordów jako przemieszczanie się pieszo. Błąd klasyfikacji wynosi w tym przypadku $\delta_w \approx 23\%$.

Rysunek 4.9 przedstawia krzywą ROC wykreślona dla różnych punktów odcięcia. Pole powierzchni pod krzywą wynosi $AUC \approx 0.8$, co świadczy o poprawnej jakości zbudowanego klasyfikatora.

Rysunek 4.10 i 4.11 przedstawiają kolejno fragment zbioru danych testowych przed zastosowaniem algorytmu klasyfikacji sposobu przemieszczania się środkiem innym niż samochód oraz fragment wyniku zastosowania tego podziału klasyfikacji. Znaczniki koloru zielonego oznaczają przemieszczanie się użytkownika z lokalizatorem GPS rowerem, czerwone natomiast przemieszczenie się pieszo. Zebrany zestaw danych obrazuje przemieszczanie się w terenie zabudowanym, z zaznaczeniem powtarzania się fragmentami tej samej trasy rowerem i pieszo.



Rysunek 4.9: Wykres krzywej ROC dla klasyfikacji II stopnia. Przemieszczanie się rowerem lub pieszo.



Rysunek 4.10: Fragment zbioru danych testowych przed zastosowaniem algorytmu klasyfikacji sposobu przemieszczania się rowerem lub pieszo, źródło: *opracowanie własne*.

Na rysunku 4.11 główne skupisko błędnie sklasyfikowanych obserwacji (znaczników koloru czarnego) znajduje się w centrum miasta. Fakt ten wynikać może ze specyfiki działania samego urządzenia GPS i problemów towarzyszących namierzaniu pozycji lokalizatora w strefach ograniczających widoczność nieba i satelitów (np. między wysokimi budynkami).



Rysunek 4.11: Fragment wyniku algorytmu klasyfikacji sposobu przemieszczania się rowerem lub pieszo, źródło: opracowanie własne.

4.2 Klasyfikacja w przypadku trzech klas (model 2)

4.2.1 Klasyfikacja rodzaju drogi przemieszczania się autem

Na podstawie dostępnych danych zbudowano również model klasyfikacji dla trzech klas, celem porównania wyników z klasyfikacją binarną. Podobnie jak w przypadku klasyfikacji binarnej, budowę modelu oraz jego testowanie przeprowadzono w środowisku *R* przy wykorzystaniu pakietów i funkcji takich jak *nnet::multinom* [15], *stats::predict* [16], czy *lmtest::coeftest* [17]. Na początek skupiono się na klasyfikacji terenu poruszania się autem, wyróżniając trzy stany: przemieszczanie się autem po drodze ekspresowej, w terenie zabudowanym oraz poza nim. Dane wejściowe zawierają łącznie 835 obserwacji, z wyszczególnieniem 236 rekordów zdefiniowanych jako ruch pojazdu w terenie zabudowanym, 303 w terenie niezabudowanym oraz 296 obserwacji odpowiadających za przemieszczanie się samochodem po drodze ekspresowej.

Na potrzeby modelu, dane podzielono losowo na zbiór uczący oraz testowy, w stosunku odpowiednio 80% – 20%. W przypadku klasyfikacji trzech klas, konieczne jest określenie klasy referencyjnej (bazowej) klasyfikacji, w stosunku do której określone jest prawdopodobieństwo przynależności obserwacji do poszczególnych klas. Na potrzeby tego zadania, za klasę referencyjną przyjęto przemieszczanie się samochodem w terenie niezabudowanym.

W oparciu o analizy parametrów opisanych w podrozdziale 3.3.2, w modelu uwzględniono następujące wielkości

- a) prędkości v_{GPS} ,
 - b) procentową wartość zmiany azymutu o więcej niż 15° $F(\Delta a_n)$.

W zbiorze treningowym znalazło się odpowiednio 189, 242 i 237 obserwacji zdefiniowanych jako przemieszczanie się samochodem w terenie zabudowanym, niezabudowanym oraz po drodze ekspresowej. W takiej samej kolejności, zbiór testowy

składał się z 47, 61 i 59 obserwacji przypisanych kolejno do poszczególnych klas. Dla ułatwienia zapisu, przyjmijmy następujące oznaczenia.

$$\begin{cases} 1 \text{ gdy poruszam się autem po drodze ekspresowej} \\ 2 \text{ gdy poruszam się autem w terenie niezabudowanym} \\ 3 \text{ gdy poruszam się autem w terenie zabudowanym.} \end{cases}$$

Poniższe zestawienia (tabela 4.7 i 4.8) przedstawiają podsumowanie wyestymowanych współczynników mających bezpośredni wpływ na zmienne wchodzące w skład modelu i prawdopodobieństwa przynależności poszczególnych obserwacji do klas, z uwzględnieniem klasy referencyjnej (bazowej).

$\frac{p(1)}{p(2)}$	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	statystyka z	$p(\hat{\beta}_i \neq 0)$
$const$	-1.198	0.385	-3.113	0.002**
v_{GPS}	0.067	0.014	4.619	3.860e-06***
$F(\Delta a_n)$	-0.019	0.006	-3.193	0.001**

Tabela 4.7: Podsumowanie zmiennych modelu regresji terenu przemieszczania się autem.

$$\hat{y}_1 = \ln \frac{p(1)}{p(2)} \approx -1.19 + 0.07 \cdot v_{GPS} - 0.2 \cdot F(\Delta a_n) \quad (4.10)$$

$\frac{p(3)}{p(2)}$	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	statystyka z	$p(\hat{\beta}_i \neq 0)$
$const$	1.236	0.466	2.652	0.008**
v_{GPS}	-0.175	0.023	-7.641	2.147e-14***
$F(\Delta a_n)$	0.019	0.006	2.902	0.004**

Tabela 4.8: Podsumowanie zmiennych modelu regresji terenu przemieszczania się autem.

$$\hat{y}_2 = \ln \frac{p(3)}{p(2)} \approx 1.24 - 0.18 \cdot v_{GPS} + 0.02 \cdot F(\Delta a_n) \quad (4.11)$$

Na podstawie zbudowanych modeli określonych wzorami 4.10 i 4.11, możliwe jest wyznaczenie poszczególnych prawdopodobieństw przynależności obserwacji do klas.

$$\begin{cases} \frac{p(1)}{p(2)} = e^{\hat{y}_1} \\ \frac{p(3)}{p(2)} = e^{\hat{y}_2} \end{cases} \implies \frac{p(1)+p(3)}{p(2)} = e^{\hat{y}_1} + e^{\hat{y}_2}. \quad (4.12)$$

$$p(1) + p(2) + p(3) = 1 \implies \frac{1-p(2)}{p(2)} = e^{\hat{y}_1} + e^{\hat{y}_2}. \quad (4.13)$$

Uwzględniając zależności określone wzorami 4.12 i 4.13 bezpośrednio można zdefiniować prawdopodobieństwo poruszania się samochodem po drodze ekspresowej $p(1)$, w terenie niezabudowanym $p(2)$ oraz w terenie zabudowanym $p(3)$

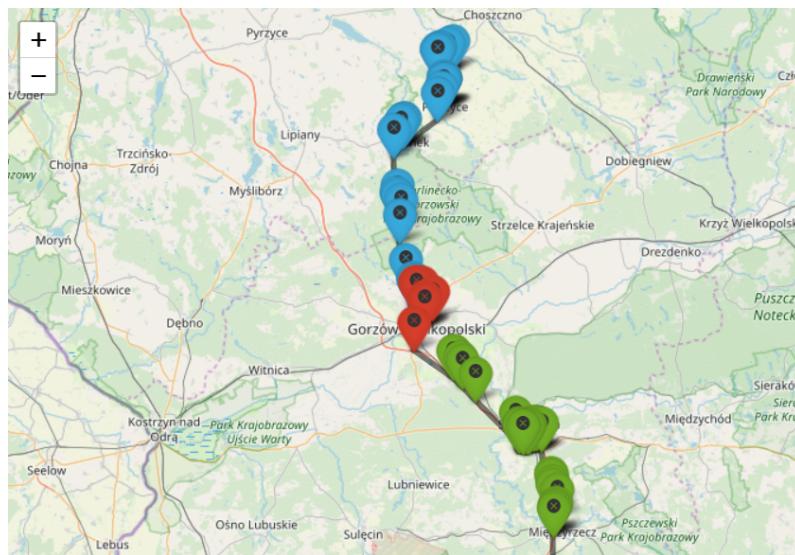
$$\begin{cases} p(1) = \frac{\exp(\hat{y}_1)}{1+\exp(\hat{y}_1)+\exp(\hat{y}_2)} \\ p(2) = \frac{1}{1+\exp(\hat{y}_1)+\exp(\hat{y}_2)} \\ p(3) = \frac{\exp(\hat{y}_2)}{1+\exp(\hat{y}_1)+\exp(\hat{y}_2)} \end{cases}$$

Na podstawie uzyskanych modeli i wyznaczeniu prawdopodobieństw przynależności obserwacji do poszczególnych klas, otrzymano macierz pomyłek, zapisaną w postaci tabeli 4.9. Uwzględniając zawarte w niej wartości, błąd klasyfikacji wynosi $\delta_w \approx 26\%$, natomiast dokładność modelu $ACC \approx 73\%$.

	klasa 1	klasa 2	klasa 3
klasa 1	51	7	1
klasa 2	15	28	18
klasa 3	0	3	44

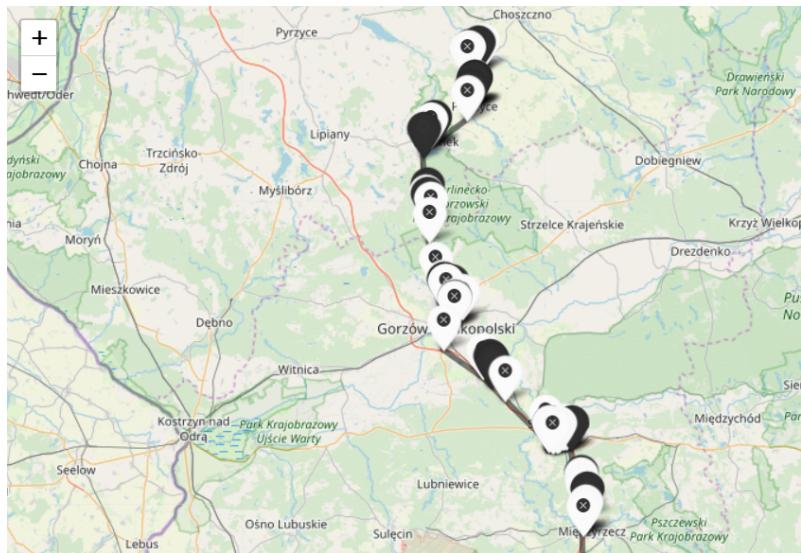
Tabela 4.9: Macierz rozkładu klas dla klasyfikacji terenu przemieszczania się autem.

Rysunek 4.12 przedstawia fragment zbioru danych testowych przed zastosowaniem algorytmu klasyfikacji terenu poruszania się samochodem. Znaczniki koloru zielonego oznaczają przemieszczanie się autem po drodze ekspresowej, koloru czerwonego w terenie zabudowanym, koloru niebieskiego - w terenie niezabudowanym.



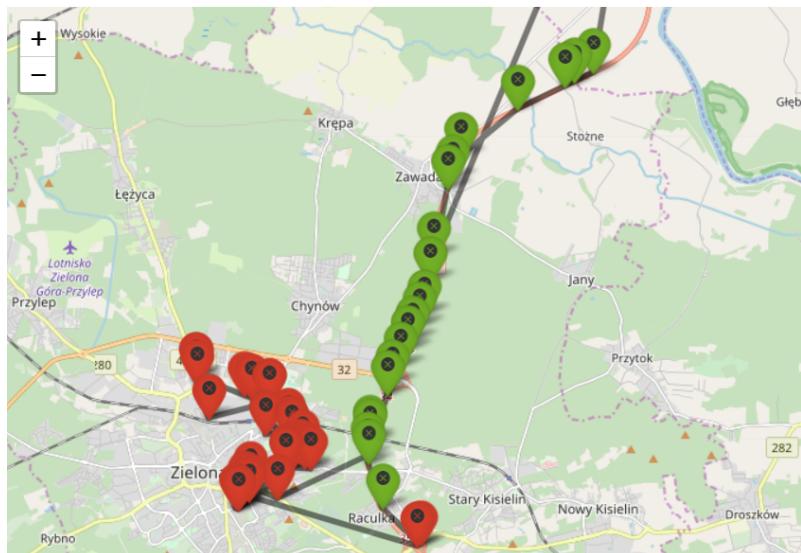
Rysunek 4.12: Fragment zbioru danych testowych przed zastosowaniem algorytmu klasyfikacji terenu poruszania się samochodem, źródło: *opracowanie własne*.

Rysunek 4.13 przedstawia fragment wyniku algorytmu klasyfikacji terenu przemieszczania się samochodem. Błędnie sklasyfikowane obserwacje pojawiają się sporadycznie na całej długości trasy i mogą wynikać z pojawiających się różnic w czasie odbioru kolejnych pozycji GPS. Zauważono bowiem, że regularność wyznaczania pozycji urządzenia wpływa na jakość parametrów opisanych w podrozdziale 3.2, które to mają bezpośredni wpływ na proces klasyfikacji oraz jego wynik.

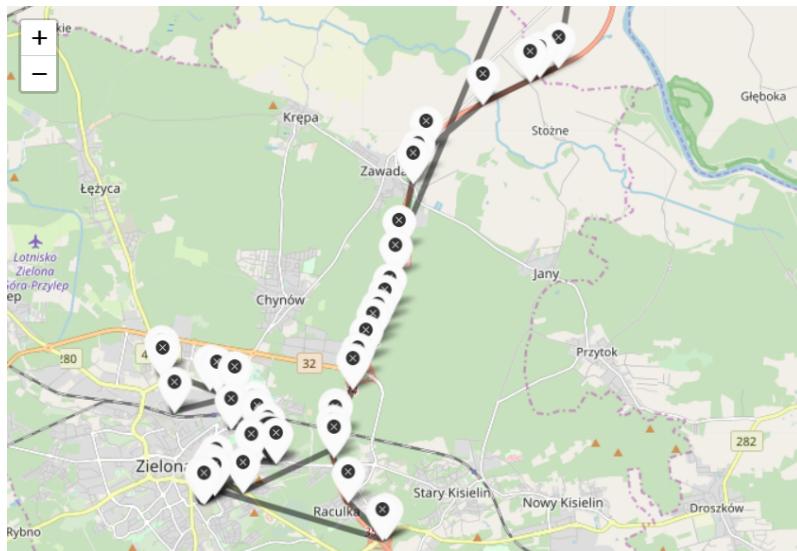


Rysunek 4.13: Fragment wyniku algorytmu klasyfikacji terenu poruszania się samochodem, źródło: *opracowanie własne*.

Rysunek 4.14 przedstawia inny fragment zbioru danych testowych przed zastosowaniem algorytmu klasyfikacji terenu przemieszczania się samochodem. Kolejne pozycje GPS wyznaczane są w sposób bardziej regularny, co istotnie wpływa na jakość otrzymanej klasyfikacji, widocznej na rysunku 4.15.



Rysunek 4.14: Fragment zbioru danych testowych przed zastosowaniem algorytmu klasyfikacji terenu poruszania się samochodem, źródło: *opracowanie własne*.



Rysunek 4.15: Fragment wyniku algorytmu klasyfikacji terenu poruszania się samochodem, źródło: opracowanie własne.

4.2.2 Klasyfikacja sposobu przemieszczania się osoby

Drugi z modeli klasyfikacyjnych dla trzech klas, dotyczy podziału danych względem sposobu przemieszczania się osoby w terenie zabudowanym, wyróżniając przemieszczanie się osoby z lokalizatorem GPS pieszo, rowerem lub autem. Dane wejściowe zawierają łącznie 858 rekordów, z wyszczególnieniem 249 obserwacji poruszania się pieszo, 341 rekordów przemieszczania się rowerem oraz 268 rekordów opisujących jazdę autem. Na potrzeby budowy modelu klasyfikacji, dane podzielono na zbiór uczący oraz testowy w stosunku 80% – 20%. W przypadku klasyfikacji trzech klas, konieczne jest określenie klasy referencyjnej (bazowej) klasyfikacji, w stosunku do której określone jest prawdopodobieństwo przynależności obserwacji do poszczególnych klas. Na potrzeby tego zadania, za klasę referencyjną przyjęto przemieszczanie się rowerem.

W oparciu o analizy parametrów opisanych w podrozdziale 3.3.2, w modelu uwzględniono następujące zmienne

- a) prędkość v_{GPS} ,
- b) procent zmiany azymutu w oknie czasowym $F(\Delta a_n)$,
- c) liczbę widocznych przez odbiornik satelitów N_{GPS} .

W zbiorze treningowym znalazło się 199 rekordów definiujących ruch pieszy, 273 obserwacji określających przemieszczanie się rowerem oraz 214 rekordów opisujących jazdę autem, natomiast w zbiorze testowym, kolejno 50, 68 oraz 54 rekordów określających ruch pieszy, jazdę rowerem oraz przemieszczanie się autem. Dla ułatwienia zapisu przyjmijmy następujące oznaczenia.

$$\begin{cases} \textbf{1} & \text{gdy poruszam się w terenie zabudowanym \textbf{pieszo}} \\ \textbf{2} & \text{gdy poruszam się w terenie zabudowanym \textbf{rowerem}} \\ \textbf{3} & \text{gdy poruszam się w terenie zabudowanym \textbf{samochodem}.} \end{cases}$$

Poniżej przedstawiono podsumowanie wyestymowanych zmiennych dla dwóch modeli zawierających prawdopodobieństwa przynależności do określonych klas, przy założeniu klasy referencyjnej (bazowej).

$\frac{p(1)}{p(2)}$	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	statystyka z	$p(\hat{\beta}_i \neq 0)$
<i>const</i>	-2.328	0.965	-2.412	0.016*
N_{GPS}	0.351	0.094	3.743	0.0001816***
v_{GPS}	-0.856	0.106	-8.039	9.105e-16***
$F(\Delta a_n)$	0.164	0.007	2.428	0.015*

Tabela 4.10: Podsumowanie zmiennych modelu regresji względem sposobu przemieszczania się osoby w terenie zabudowanym.

$$\hat{y}_1 = \ln \frac{p(1)}{p(2)} \approx -2.47 + 0.39 \cdot N_{GPS} - 0.98 \cdot v_{GPS} + 0.02 \cdot F(\Delta a_n) \quad (4.14)$$

$\frac{p(3)}{p(2)}$	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	statystyka z	$p(\hat{\beta}_i \neq 0)$
<i>const</i>	-8.574	1.143	-7.503	6.241e-14***
N_{GPS}	0.420	0.100	4.197	2.701e-05***
v_{GPS}	0.579	0.060	9.589	< 2.2e-16***
$F(\Delta a_n)$	0.046	0.008	6.079	1.208e-09***

Tabela 4.11: Podsumowanie zmiennych modelu regresji względem sposobu przemieszczania się osoby w terenie zabudowanym

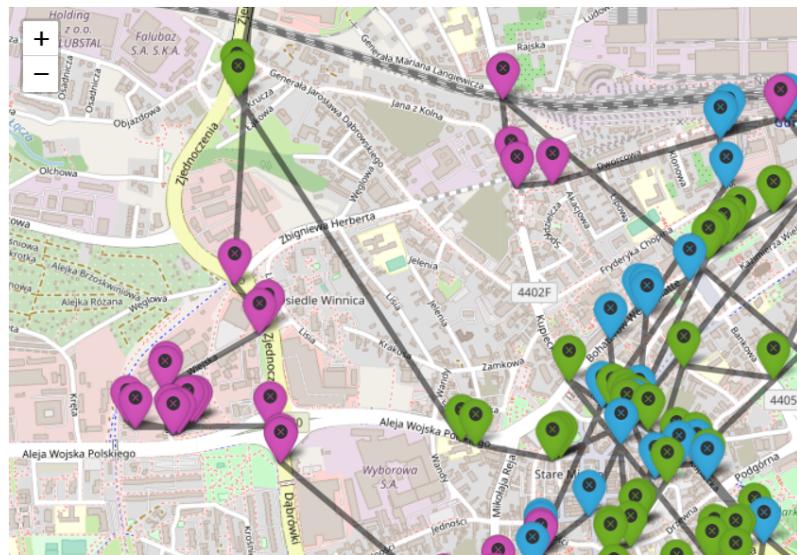
$$\hat{y}_2 = \ln \frac{p(3)}{p(2)} \approx -8.72 + 0.43 \cdot N_{GPS} + 0.56 \cdot v_{GPS} + 0.05 \cdot F(\Delta a_n) \quad (4.15)$$

Na podstawie modeli określonych wzorami 4.14 i 4.15 wyznaczono macierz przynależności obserwacji do poszczególnych klas. Błąd klasyfikacji wynosi $\delta_w \approx 26\%$, natomiast dokładność modelu wynosi $ACC \approx 74\%$.

	klasa 1	klasa 2	klasa 3
klasa 1	41	7	2
klasa 2	8	54	6
klasa 3	10	11	33

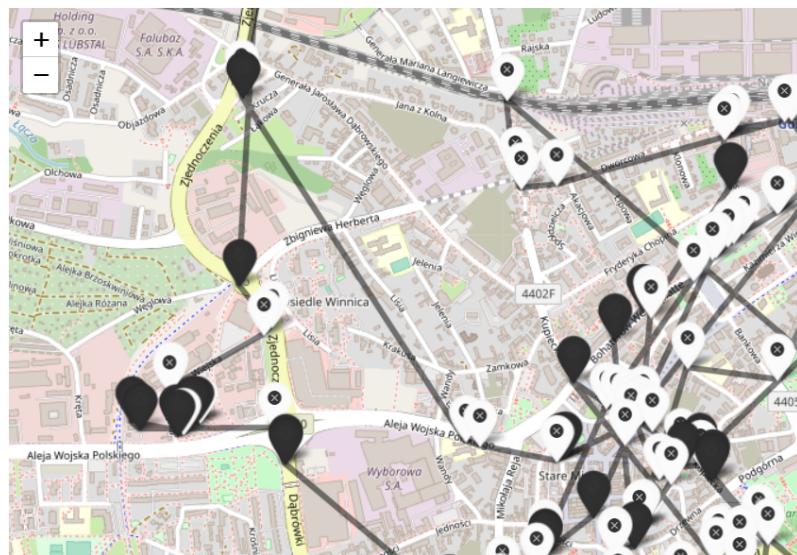
Tabela 4.12: Macierz rozkładu klas dla klasyfikacji względem sposobu przemieszczania się osoby w terenie zabudowanym.

Rysunek 4.16 przedstawia fragment zbioru danych testowych użytych do niniejszego algorytmu klasyfikacji. Znaczniki koloru fioletowego oznaczają przemieszczanie się samochodem osoby z lokalizatorem GPS, koloru zielonego poruszanie się rowerem, natomiast koloru niebieskiego przemieszczanie się pieszo.



Rysunek 4.16: Fragment zbioru danych testowych przed zastosowaniem algorytmu klasyfikacji sposobu poruszania się w terenie zabudowanym, źródło: *opracowanie własne*.

Rysunek 4.17 przedstawia wynik algorytmu klasyfikacji sposobu przemieszczania się osoby w terenie zabudowanym. Błędna klasyfikacja dla obserwacji pojawia się tym częściej, im lokalizator namierza swoją pozycję w sposób chaotyczny, nieregularny. Wynikać to może ze specyfiki obszaru przebywania osoby z lokalizatorem w terenie ograniczonym wysokimi budynkami ograniczającymi widoczność nieba lub niewielkimi, porównywalnymi do siebie prędkościami w przypadku przemieszczania się użytkownika z lokalizatorem rowerem i pieszo.



Rysunek 4.17: Fragment wyniku algorytmu klasyfikacji sposobu przemieszczania się w terenie zabudowanym, źródło: *opracowanie własne*.

4.3 Podsumowanie

W pracy zaproponowano wykorzystanie dwóch modeli regresji logistycznej w celu rozpoznania stanu poruszania się osoby z lokalizatorem GPS. Pierwszy z modeli, oparty na klasyfikacji binarnej, dał bardzo dobre rezultaty. Dokładność klasyfikacji I stopnia (jadę/nie jadę) osiągnęła wynik 88%, a klasyfikacji II stopnia miejsca przemieszczania się autem (teren zabudowany/teren niezabudowany) aż 93%. Najmniej korzystnie wypadła natomiast klasyfikacja II stopnia sposobu przemieszczania się pieszo lub rowerem. Jej dokładność wyniosła około 77%. Drugi z modeli, oparty na przypadku klasyfikacji obserwacji do trzech klas (zarówno dla klasyfikacji rodzącej drogi przemieszczania się autem, jak i klasyfikacji sposobu przemieszczania się osoby), dał dokładność rzędu około 74%. Z analitycznego punktu widzenia, zadowalający wynik dokładności algorytmu, rzędu 80-90%, dał model 1. Pomimo, że oba modele klasyfikacji zostały oparte o kombinację tych samych statystyk danych pomiarowych, takich jak v_{GPS} , N_{GPS} , Δa_n czy $F(\Delta a_n)$, warto zaznaczyć, że wpływ na wynik klasyfikacji mogło mieć zróżnicowanie liczby obserwacji, należących do poszczególnych klas w zestawie danych wejściowych. Nieproporcjonalność ta niekiedy jest przyczyną otrzymywania z ludnie optymistycznego wyniku klasyfikacji.

Podjęcie próby rozwiązywania problemu dotyczącego rozpoznawania stanów poruszania się osoby z lokalizatorem było możliwe dzięki udostępnionemu, na potrzeby testów i analiz, odbiornikowi GPS typu *HP-500* przez zielonogórską firmę *Hertz Systems Ltd*. Na koniec warto dodać, że możliwość uzyskania lepszego wyniku klasyfikacji metodą regresji, być może byłaby możliwa przy wykorzystaniu większego zbioru danych wejściowych oraz dostępie do większej ilości informacji rejestrowanych przez odbiornik GPS, na przykład zapisów depesz nawigacyjnych lub wielkości współczynników geometrycznych (*HDOP*, *VDOP*), opisujących dokładność położenia użytkownika na Ziemi.

Bibliografia

- [1] Stanisław Osowski, *Metody i narzędzia eksploracji danych*, BTC 2013.
- [2] Charu C. Aggarwal, *Data Mining*, Springer 2015.
- [3] Daniel T. Larose *Data Mining Methods and Models*, Wiley 2006.
- [4] Max Bramer *Principles of Data Mining*, Springer 2007.
- [5] Tadeusz Morzy, *Eksploracja danych - Metody i algorytmy*, PWN 2013.
- [6] Przemysław Biecek, *Przewodnik po pakiecie R*, GiS 2014.
- [7] Piotr Kaniewski, *System nawigacji satelitarnej GPS*, Elektronika Praktyczna 2006.
- [8] Ewelina Kamrowska, *Wykrywanie anomalii w pozycjonowaniu lokalizatora GPS*, Praca Inżynierska 2018.
- [9] Jacek Koronacki, Jan Ćwik, *Statystyczne systemy uczące się*, Exit, wydanie II.
- [10] Joanna Giemza, Katarzyna Zwierzchowska, *Wprowadzenie do modelu regresji logistycznej wraz z przykładem zastosowania w pakiecie statystycznym R do danych o pacjentach po przeszczepie nerki*, Praca Licencjacka 2011.
- [11] Jerzy Stefanowski, *Maszynowe uczenie się*, Wykład Instytutu Informatyki Politechniki Poznańskiej 2009.
- [12] Podręcznik użytkownika PQStat 1.6.8, *PQStat Software* 2019, <http://manuals.pqstat.pl/>.
- [13] Przewodnik po nawigacji, *Metody pozycjonowania GPS*, <https://technologiagps.org.pl/pozycjonowanie.html>.
- [14] R-core, dokumentacja funkcji glm, *Fitting Generalized Linear Models*, <https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/glm>.
- [15] Brian Ripley, dokumentacja funkcji multinom, *Fit Multinomial Log-Linear Models*, <https://www.rdocumentation.org/packages/nnet/versions/7.3-12/topics/multinom>.
- [16] Robert Hijmans, dokumentacja funkcji predict, *Spatial Model Predictions*, <https://www.rdocumentation.org/packages/raster/versions/2.9-5/topics/predict>.
- [17] Achim Zeileis, dokumentacja funkcji coefest, *Inference For Estimated Coefficients*, <https://www.rdocumentation.org/packages/lmtest/versions/0.9-37/topics/coefest>.