

# Raport: Titanic – model predykcyjny

M. Wiśniewska

Projekt wykonany w ramach przedmiotu Podstawy  
Uczenia Maszynowego na Polsko-Japońskiej Akademii  
Technik Komputerowych w Warszawie

## Spis treści

<b>OPIS PROBLEMU .....</b>	<b>3</b>
<b>DANE .....</b>	<b>3</b>
<b>SPOSÓB ROZWIĄZANIA PROBLEMU .....</b>	<b>6</b>
<b>DYSKUSJA WYNIKÓW I EWALUACJA MODELU .....</b>	<b>12</b>
QUICK MODELING.....	12
INTERPRETABLE MODELING .....	24
CHOOSE ALGORITHMS .....	25
<b>PODSUMOWANIE .....</b>	<b>28</b>
<b>SPIS RYSUNKÓW .....</b>	<b>29</b>

# Opis problemu

Poniższy raport dotyczy modelu predykcyjnego w oparciu o rzeczywisty zbiór danych, który zawiera listę pasażerów statku Titanic, który zatonął 15 kwietnia 1912 roku.

Przeprowadzono analizę danych w celu sprawdzenia prawdopodobieństwa przeżycia pasażerów na podstawie informacji o nich i ich podróży w platformie Dataiku.

Przedstawiony problem wydaje się szczególnie interesujący ze względu na połączenie nowoczesnego podejścia w kontekście edukacyjnym i zainteresowanie jakie budzi sama katastrofa statku. Dzięki analizie jesteśmy w stanie poznać cechy pasażerów i zrozumieć czynniki jakie wpływają na przeżywalność w podobnych wydarzeniach.

## Dane

Źródłem danych wykorzystanych podczas analizy jest lista 1309 pasażerów w pliku csv („titanic.csv”). Plik zawiera 10 zmiennych objaśniających:

- Pclass – klasa | (1 = pierwsza; 2 = druga; 3 = trzecia),
- Age – wiek pasażera,
- Sex – płeć pasażera,
- Sibsp – liczba małżonków lub rodzeństwa na pokładzie,
- Parch – liczba rodziców lub dzieci na pokładzie,
- Ticket – numer biletu,
- Fare – opłata za bilet,
- Cabin – kabina,
- Embarked – port startowy (C = Cherbourg; Q = Queenstown; S = Southampton)

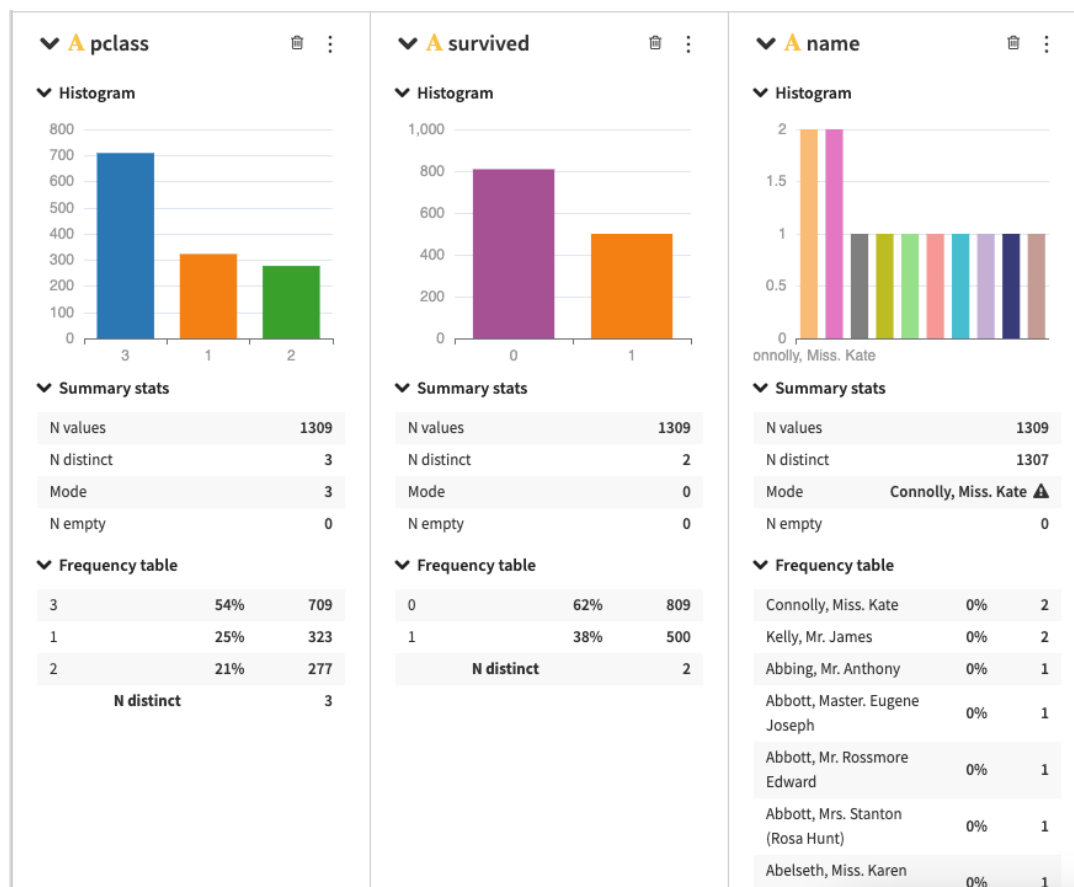
```
df = pd.read_csv('titanic.csv')
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 14 columns):
#   Column      Non-Null Count  Dtype  
---  -
0    pclass      1309 non-null   int64  
1    survived    1309 non-null   int64  
2    name        1309 non-null   object  
3    sex         1309 non-null   object  
4    age         1046 non-null   float64 
5    sibsp       1309 non-null   int64  
6    parch       1309 non-null   int64  
7    ticket      1309 non-null   object  
8    fare        1308 non-null   float64 
9    cabin       295 non-null    object  
10   embarked    1307 non-null   object  
11   boat        486 non-null    object  
12   body        121 non-null    float64 
13   home_dest    745 non-null    object  
dtypes: float64(3), int64(4), object(7)
memory usage: 143.3+ KB
```

Rysunek 1. Informacje o niezmodyfikowanym pliku "titanic.csv" (Google Colab)

Przedstawione dane pozwolą na przygotowanie modelu predykcyjnego przeżywalności katastrofy.

Poniżej przedstawiono analizę i przegląd danych:

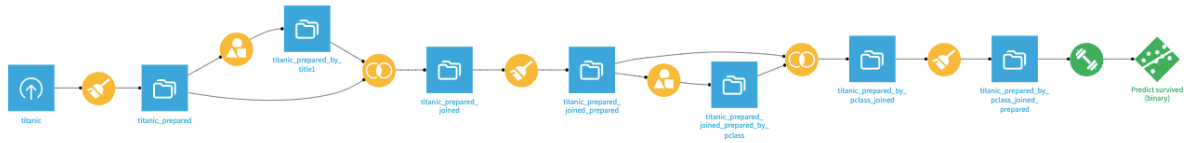


Rysunek 2. Univariate analysis 1



# Sposób rozwiązania problemu

Dane zostały uzupełnione i przygotowane do kolejnych etapów modelowania w systemie Dataiku.



Rysunek 5. Flow (Dataiku)

Plik titanic.csv został pobrany i załadowany do Dataiku wraz z usunięciem nieistotnych kolumn (boat, body, cabin i home\_dest) ze względu na brak wystarczającej liczby wartości.

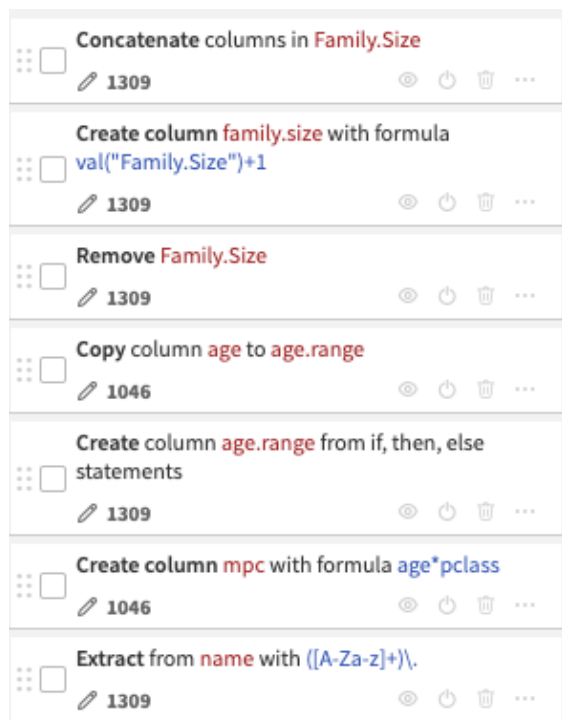
Pozostałe kolumny były całkowicie kompletne w dane o pasażerach lub wymagały uzupełnienia.

Przeprowadzono grupowanie po tytule, aby uzyskać średnią wieku dla danego tytułu. Tytuł został wyekstraktowany dzięki wykorzystaniu wyrażenia regularnego (regular expression).

title1	age_avg	count
string Text	double Decimal	bigint Integer
Capt	70.0	1
Col	54.0	4
Countess	33.0	1
Don	40.0	1
Dona	39.0	1
Dr	43.57142857142857	8
Jonkheer	38.0	1
Lady	48.0	1
Major	48.5	2
Master	5.482703773584906	61
Miss	21.774206666666668	260
Mlle	24.0	2
Mme	24.0	1
Mr	32.25215146299484	757
Mrs	36.99411764705882	197
Ms	28.0	2
Rev	41.25	8
Sir	49.0	1

Rysunek 6. Tabela wynikowa po grupowania po tytule (Dataiku)

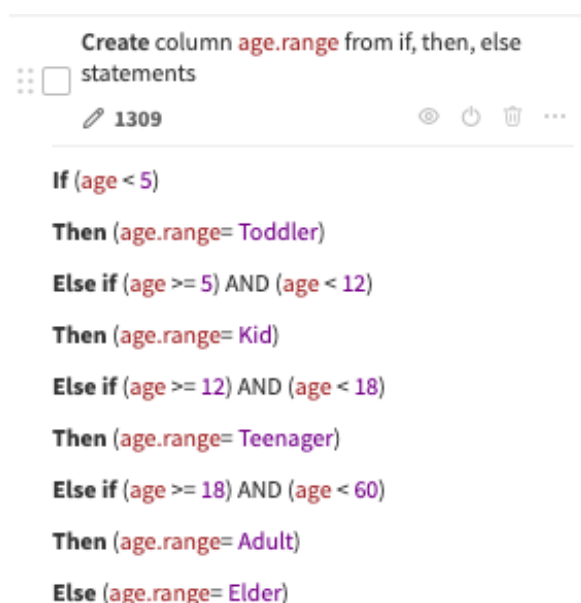
Dzięki grupowaniu można połączyć tabelę wynikową i tabelę ze wszystkimi danymi pasażerów. Użycie „Join recipe” łączy tabele i pozwala na uzupełnienie pustych komórek w kolumnie „age” o wiek średni po tytule (age\_avg).



Rysunek 7. Kroki tworzenia dodatkowych zmiennych (Dataiku)

Utworzono następujące kolumny:

- family.size – dodano do siebie wartości z kolumn „Parch” i „SibSp” oraz dodano 1, aby uzyskać wielkość rodziny,
- age.range – zmienna kategoryczna, określająca kategorię wiekową pasażerów (Baby, Kid, Teenager, Adult, Elder) wg poniższych warunków w Dataiku:




Rysunek 8. Warunki wykorzystane przy budowie zmiennej age.range w Dataiku



- mpc – zmienna powstała po pomnożeniu kolumn age oraz pclass, aby móc pokazać szanse na przeżycie (osoba starsza w klasie o wyższym numerze będzie miała wyższy wskaźnik niż osoba młoda w klasie o niskim numerze).

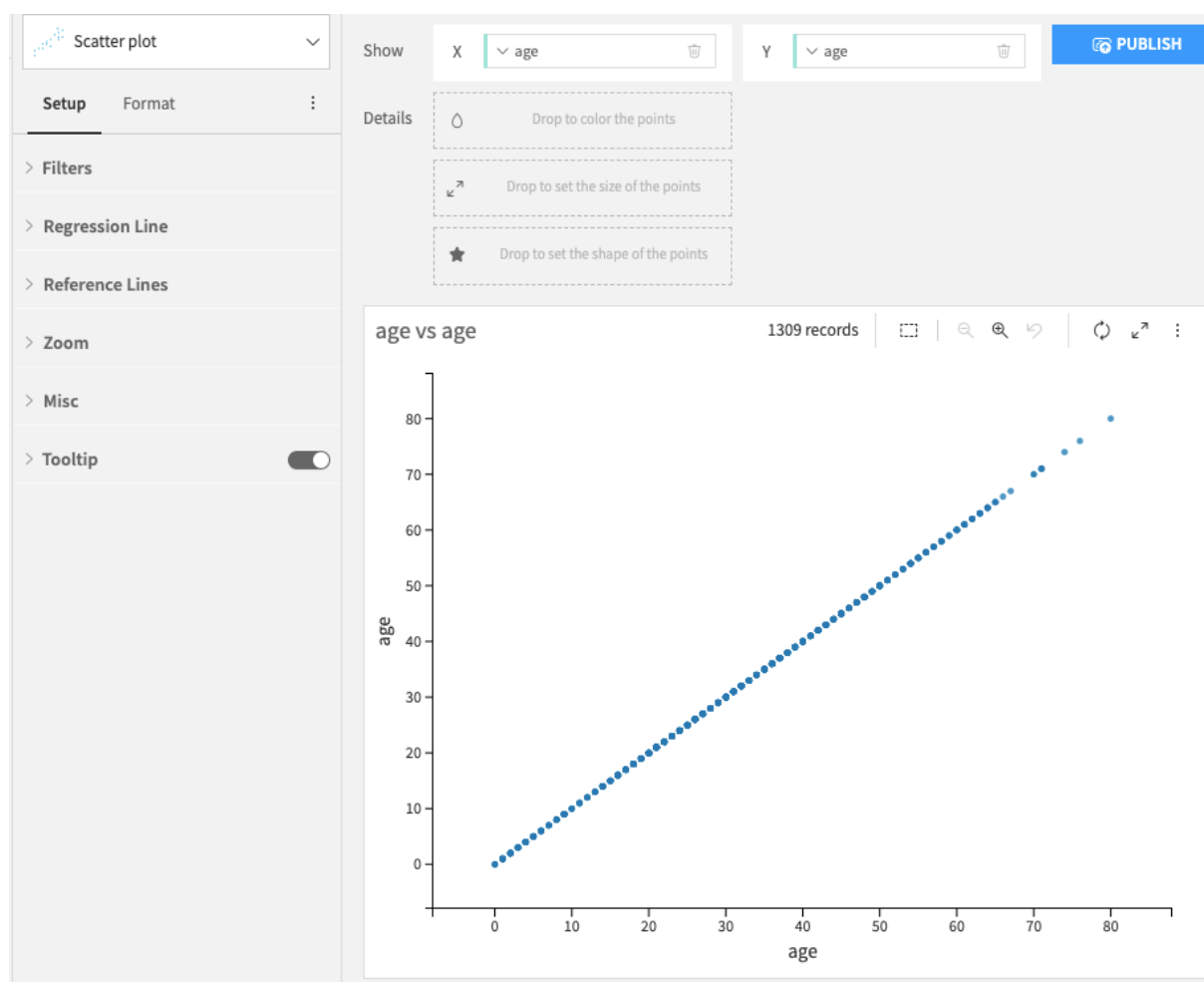
Kolejną kolumną z niekompletnymi jest „fare” z cenami biletów. W celu uzupełnienia danych przeprowadzono kolejne grupowanie. Dodano „Group recipe” wykorzystując kolumnę „pclass”, aby obliczyć średnią cenę po klasie.

pclass	fare_avg	count
string	double	bigint
Text 	Decimal	Integer
1	87.50934984520129	323
2	21.179277978339353	277
3	13.304138418079068	709

Rysunek 9. Tabela wynikowa grupowania po "pclass"

Podobnie jak przy uzupełnianiu wieku, także połączono powstałą tabelę z tabelą główną i wprowadzono średnią w pola puste. Dodatkowo zaokrąglono ceny biletów w kolumnie „fare”, aby otrzymać kwoty z dwoma miejscami po przecinku.

Następnie przeanalizowano wartości odstające. Pierwszą modyfikowaną kolumną jest „age”. Na poniższym wykresie widać punktowym widać wartości odstające powyżej wieku 67 (Rysunek 10).



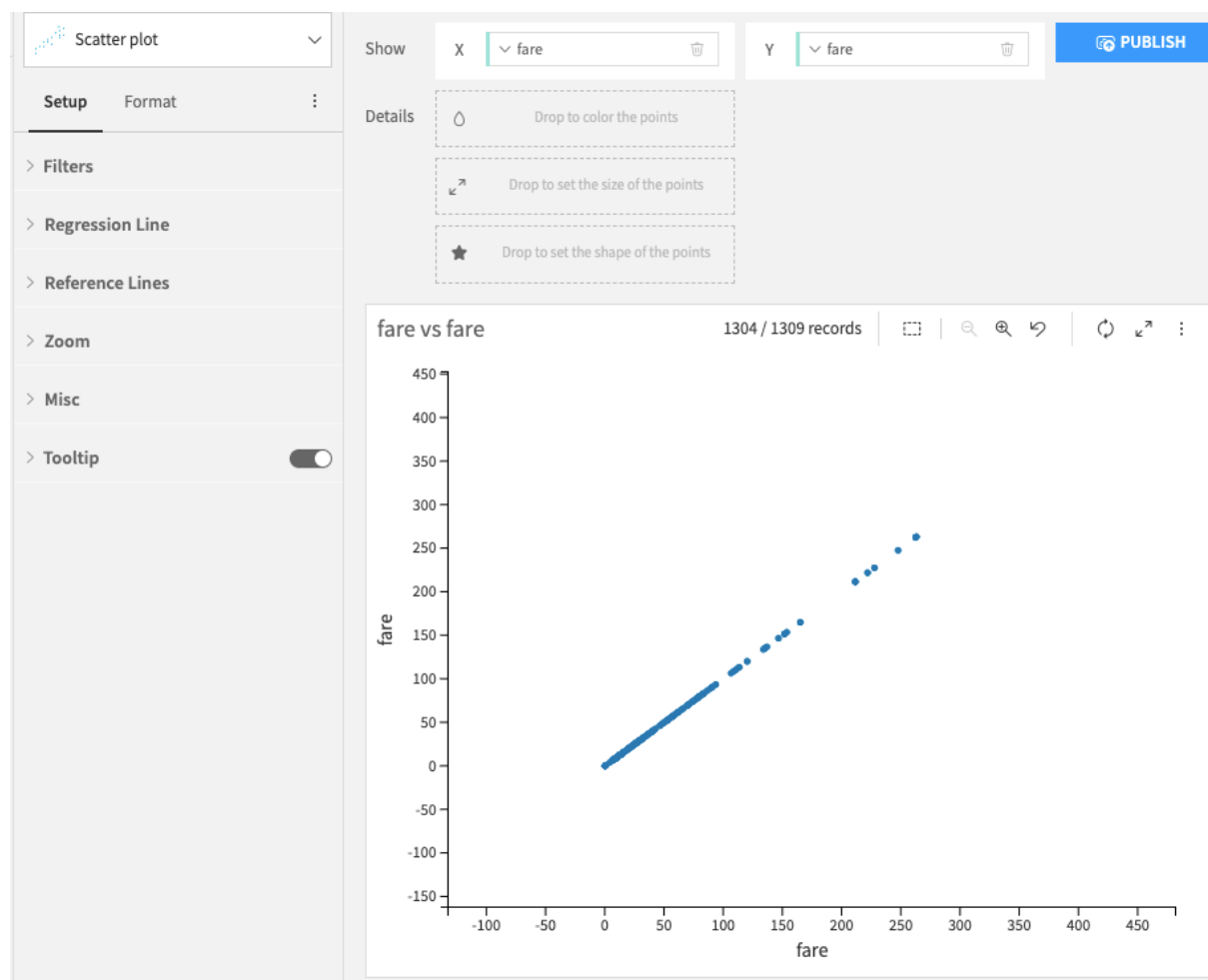
Rysunek 10. Wykres punktowy dla zmiennej "age" (Dataiku)

Dane powyżej o wartościach powyżej 67 zostały zamienione na wartość graniczną. Nie użyto średniej, ponieważ jest ona znacząco niższa niż wartości odstające (Rysunek 11).

age_avg	age
double	double
Decimal	Decimal
70.0	70.0
32.25215146299484	71.0
32.25215146299484	80.0
32.25215146299484	71.0
32.25215146299484	67.0
36.99411764705882	76.0
32.25215146299484	70.0
32.25215146299484	71.0
32.25215146299484	74.0

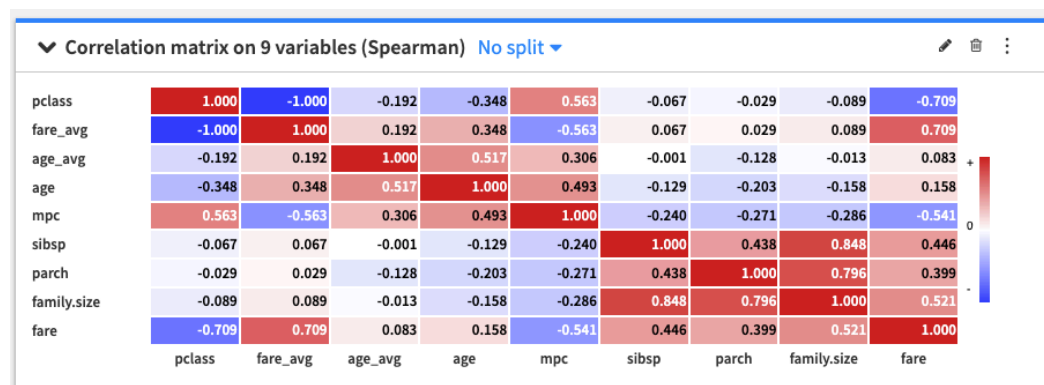
Rysunek 11. Porównanie średnich wieku do wartości odstających

Kolejną kolumną poddaną modyfikacji jest opłata za bilet. Ponownie przygotowano wykres punktowy (Rysunek 12). Wartości odstające zostały zamienione na wartość graniczną.



Rysunek 12. Wykres punktowy zmiennej "fare" (Dataiku).

Analiza korelacji (Rysunek 10.) ukazuje bardzo silne korelacje z powiązаныmi ze sobą zmiennymi jak sibsp i family.size lub mpc i fare. Inne zmienne mają umiarkowaną lub słabą korelację.

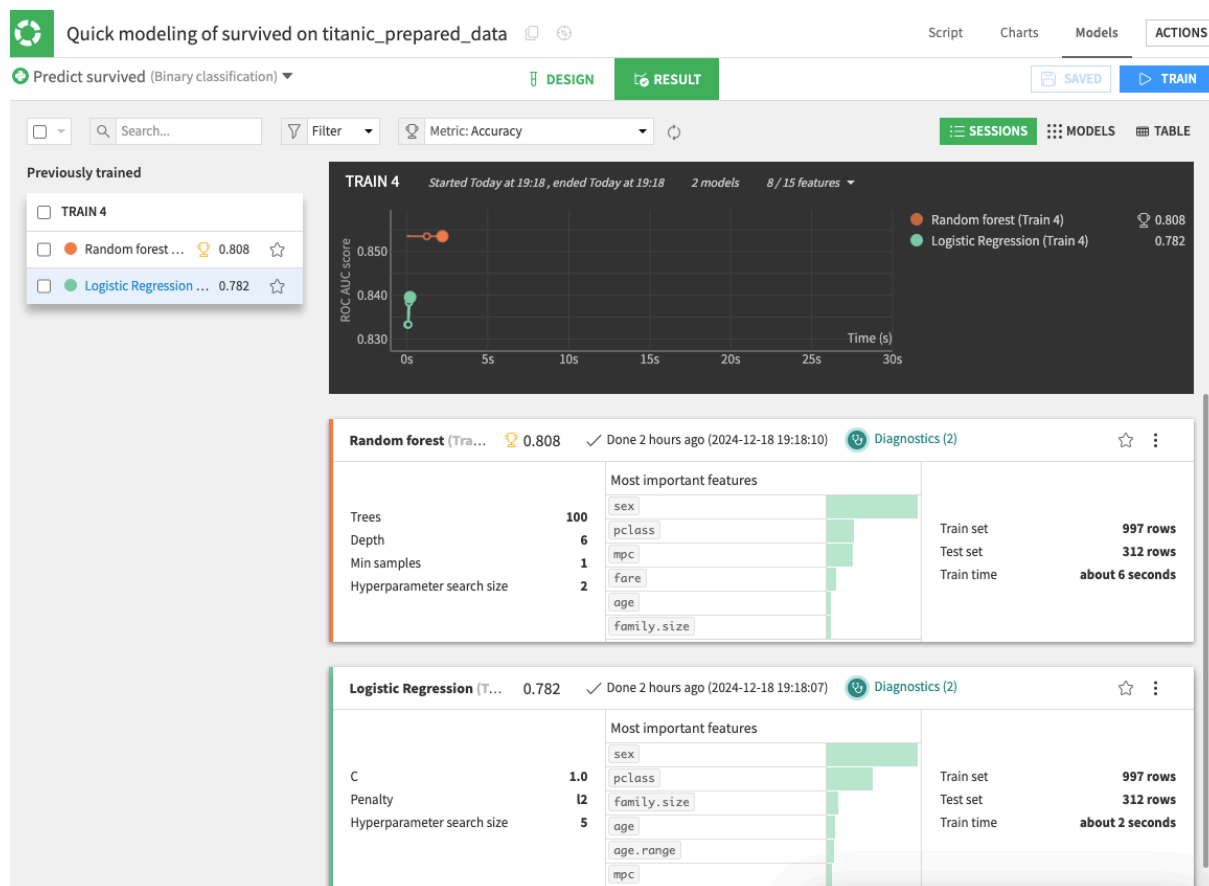


Rysunek 13. Macierz korelacji (Dataiku)

# Dyskusja wyników i ewaluacja modelu

## Quick modeling

Po wyczyszczeniu i modyfikacji danych, stworzono wstępny model dzięki funkcji „Quick modeling”. Model predykcyjny przewiduje prawdopodobieństwo przeżycia pasażerów dzięki kolumnie „survived”, której typ został zmieniony na boolean, ponieważ zamieszczone w niej dane są binarne oraz mogą przyjmować wartości prawda i fałsz.



Rysunek 14. Wyniki modelowania (Dataiku)

Wybrano model „Random forest classification”. Osiąga on poziom trafności „accuracy” rzędu 80,8%.

## ◆ Model

Model ID	A-TITANIC_S25650-6niAC7Bx-hlxsQMpi-s10-pp1-m1 <a href="#">📄</a>
Model type	Two-class classification
Target	survived
Classes	false true
Backend	Python (in memory)
Algorithm	Random forest classification
Trained on	2024/12/18 19:18
Columns	15
Train set rows	997
Test set rows	312
Calibration method	No calibration
Code Env	DSS builtin env
Python version	3.9.20

Rysunek 15. Podsumowanie informacji o modelu (Dataiku)

Do przeprowadzenia modelowania wybrano następujące atrybuty:

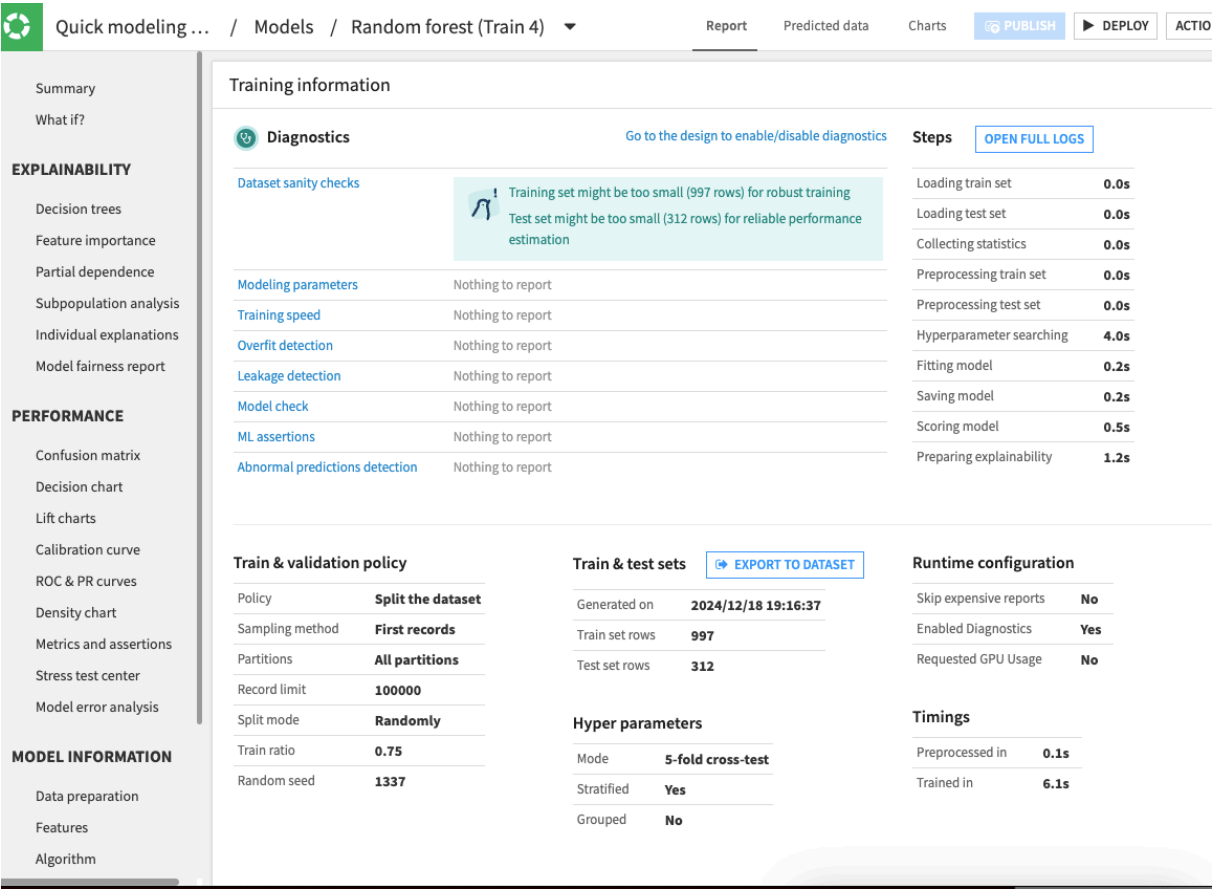
- family.size (typ numeryczny)
- mpc (typ numeryczny)
- age.range
- fare (typ numeryczny)
- sex
- pclass
- age (typ numeryczny)

Wszystkie typy numeryczne zostały poddane transformacji „Min-Max rescaling” (Rysunek 16).

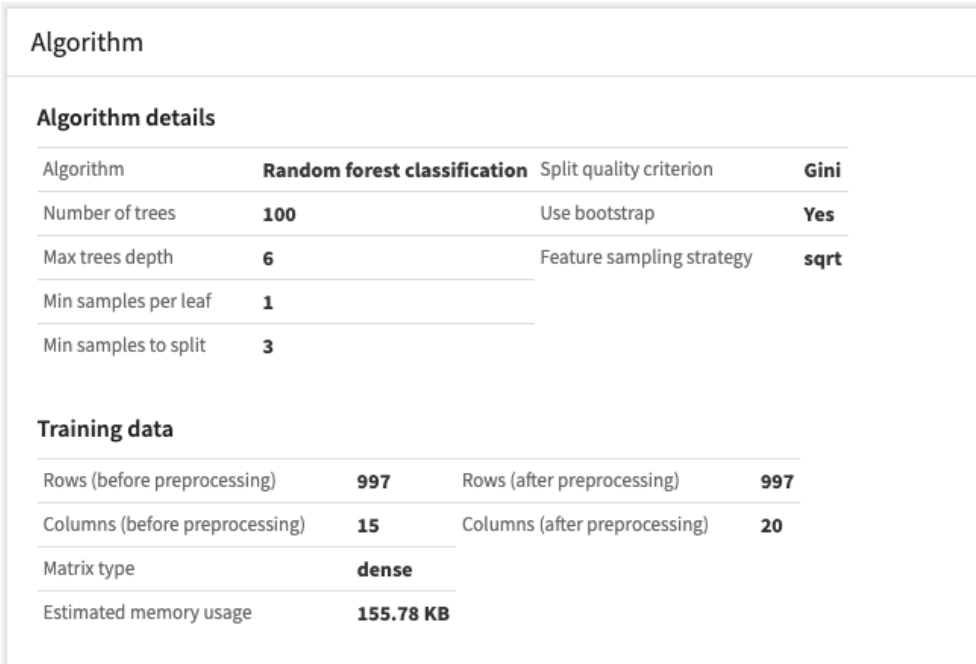
Feature Name	Role	Type	Summary
▼ family.size	➔ Input	# Numeric	Min-max rescaling
▼ mpc	➔ Input	# Numeric	Min-max rescaling
▼ age.range	➔ Input	A Category	Dummy encoding
▼ fare	➔ Input	# Numeric	Min-max rescaling
▼ sex	➔ Input	A Category	Dummy encoding
✖ title1	✖ Rejected	A Category	
✖ age_avg	✖ Rejected	# Numeric	
▼ survived	🎯 Target	A Category	
▼ pclass	➔ Input	A Category	Dummy encoding
✖ sibsp	✖ Rejected	# Numeric	
✖ name	✖ Rejected	I Text	
✖ fare_avg	✖ Rejected	# Numeric	
✖ embarked	✖ Rejected	A Category	
▼ age	➔ Input	# Numeric	Min-max rescaling
✖ parch	✖ Rejected	# Numeric	

Rysunek 16. Lista zmiennych z typem i podsumowaniem (Dataiku)

Poniżej przedstawiono diagnostykę modelu (Rysunek 17) oraz szczegóły algorytmu (Rysunek 18):



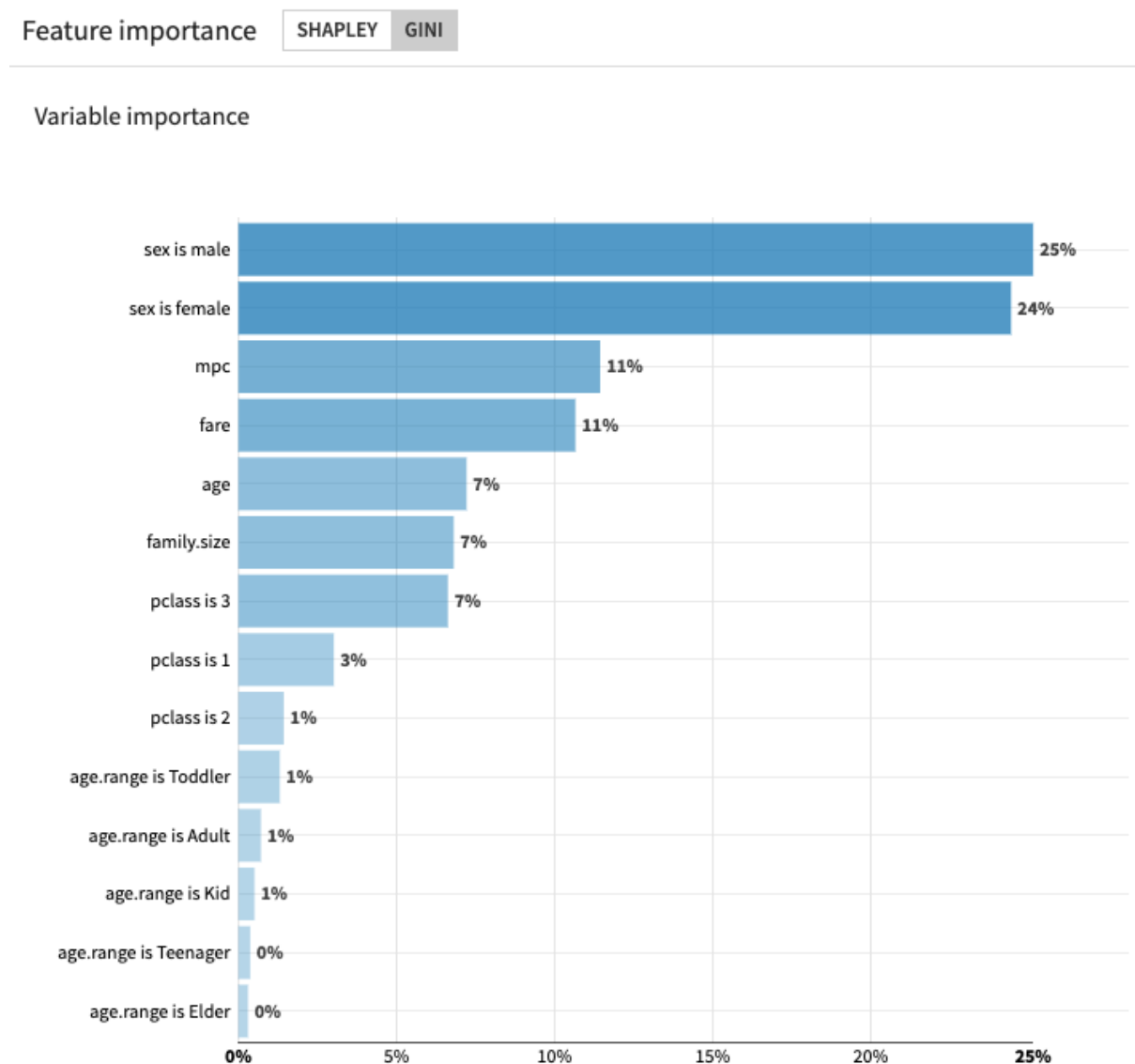
Rysunek 17. Informacje treningowe (Dataiku)



Rysunek 18. Szczegóły algorytmu i danych testowych (Dataiku)

Po analizie danych można zweryfikować, które atrybuty mają największy wpływ na przeżywalność oraz sformułować wnioski.

Dzięki sekcji „Explainability” i podsekcji „Feature importance” sprawdzono, jak zachowuje się model oraz jakie relacje zachodzą między danymi wejściowymi a przewidywanymi wyjściowymi. Największy wpływ na predykcję modelu ma płeć, pclass, mpc oraz fare (Rysunek 19 oraz Rysunek 20).



Rysunek 19. Ważność funkcji (Dataiku)

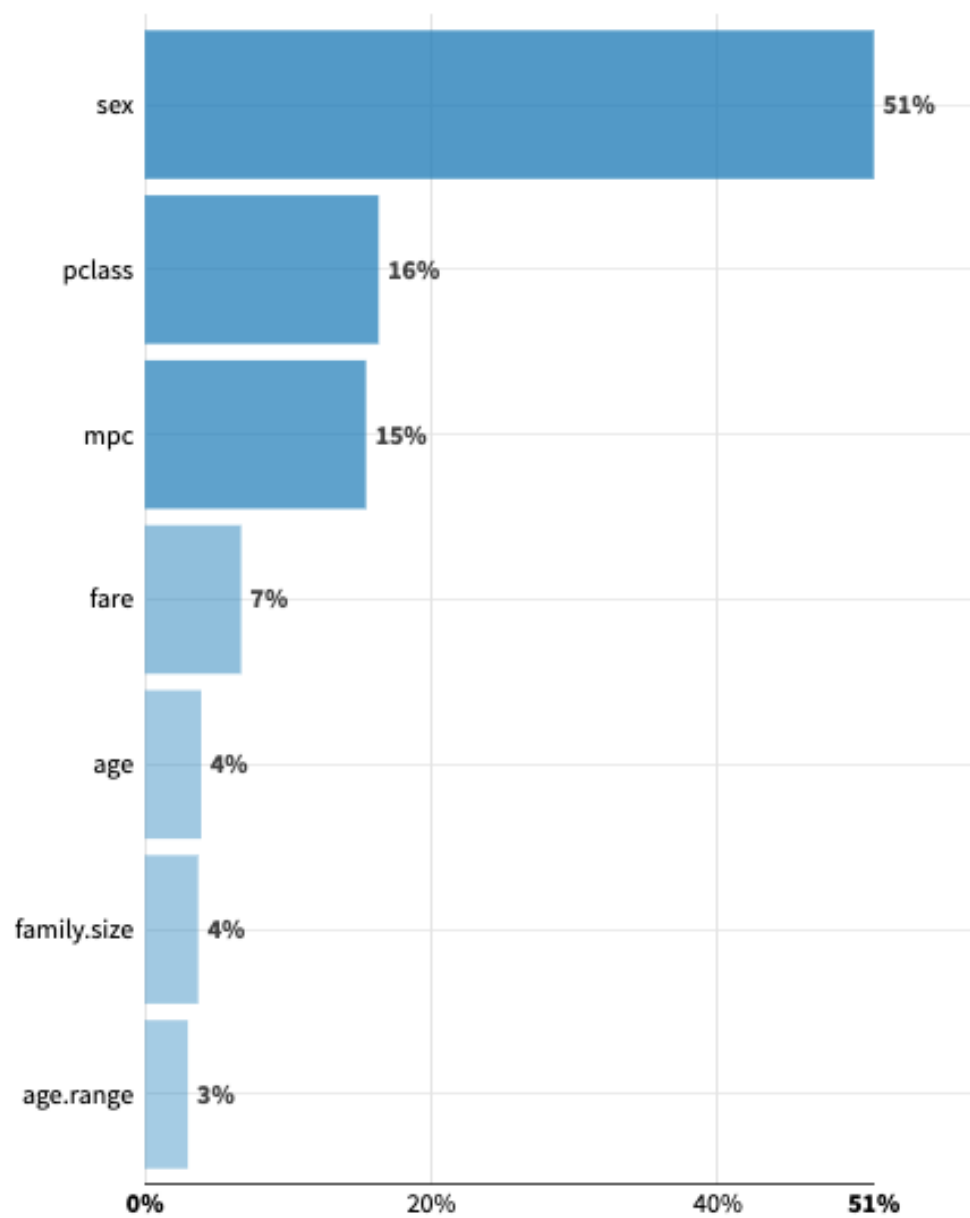


## Feature importance

SHAPLEY

GINI

### Absolute feature importance



Rysunek 20. Ważność atrybutów (Dataiku)

Dalej przeanalizowano podsekcję „Partial dependence”, która pozwala zrozumieć wpływ pojedynczych atrybutów na model:

- „mpc” – im wyższy wskaźnik, tym niższe szanse na przeżycie (Rysunek 21),

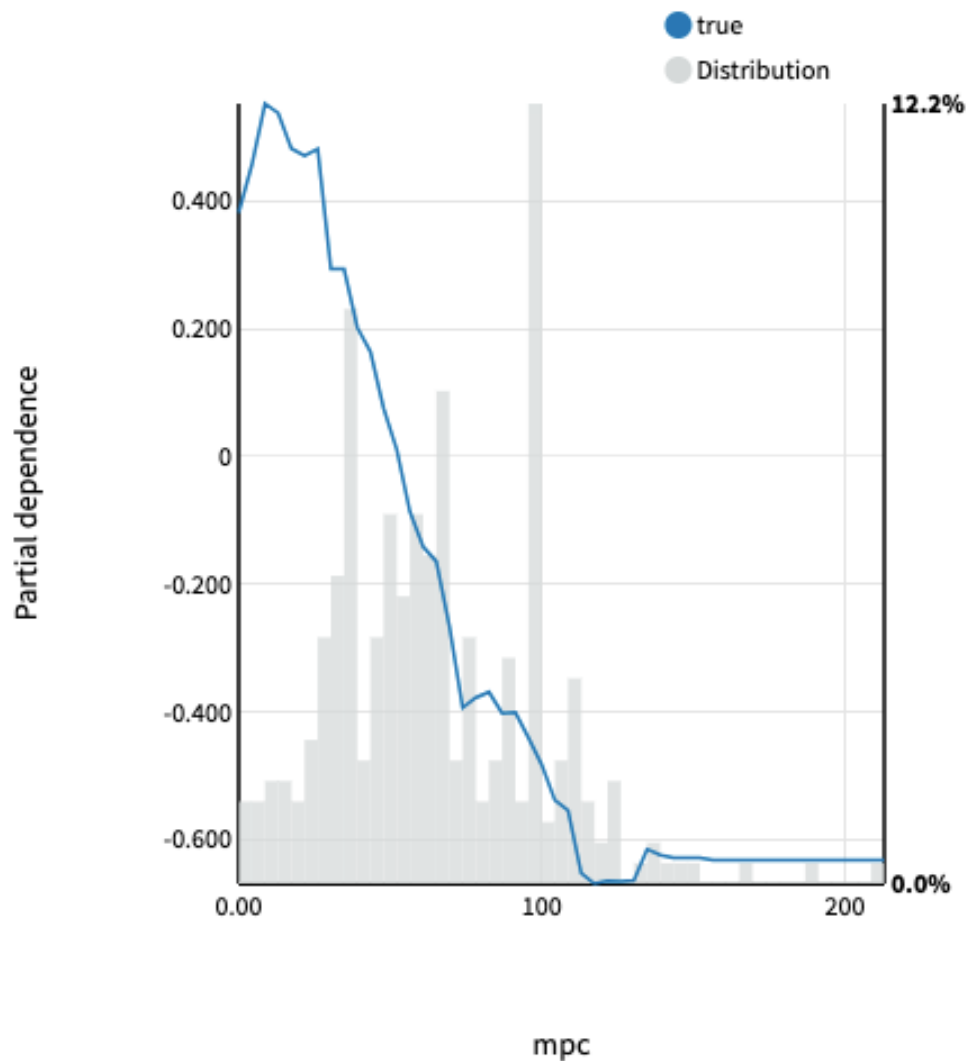
## Partial dependence

Select your variable

# mpc

COMPUTE

50 bins for **mpc**, computed on 312 rows (the full test set)



Rysunek 21. Partial dependence (Dataiku)

- „age” – wraz z większym wiekiem maleje prawdopodobieństwo przeżycia (Rysunek 22),

## Partial dependence

Select your variable

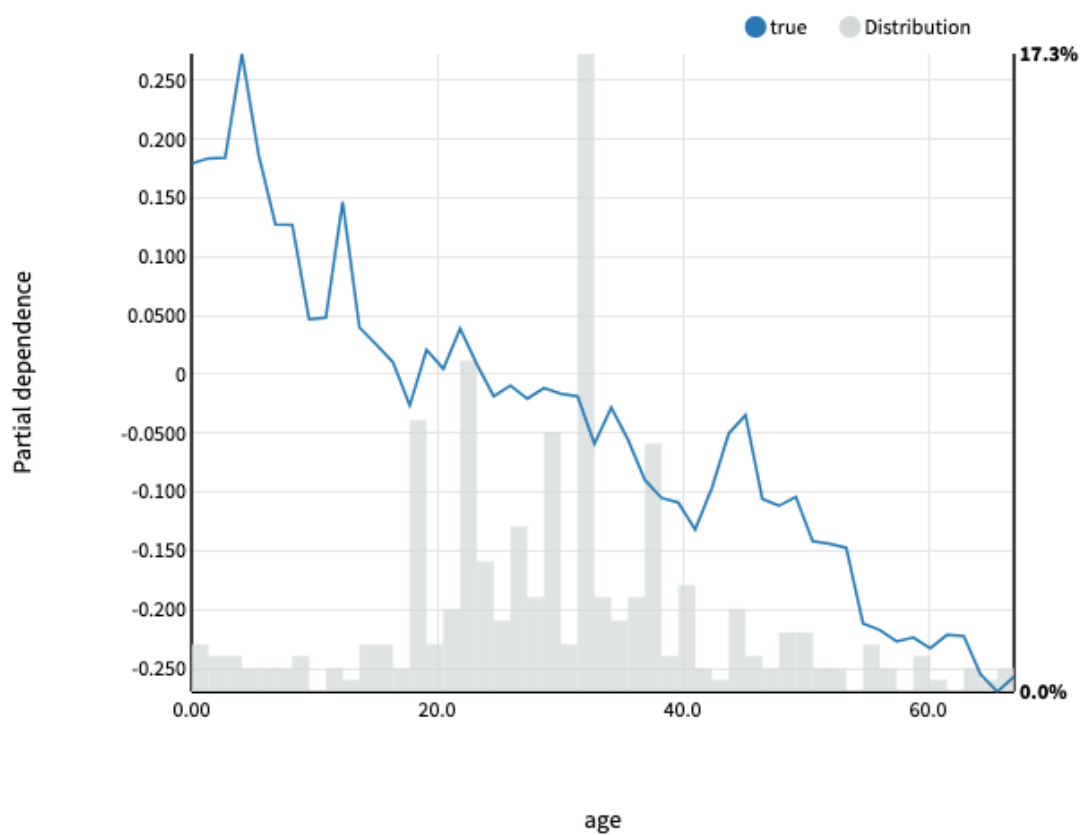
# age

COMPUTE

COMPUTE ALL

EXPOR

50 bins for **age**, computed on 312 rows (the full test set)



Rysunek 22. Partial dependence "age" (Dataiku)

- „fare” – szanse na przeżycie rosną wraz z ceną za bilet (Rysunek 23),

## Partial dependence

Select your variable

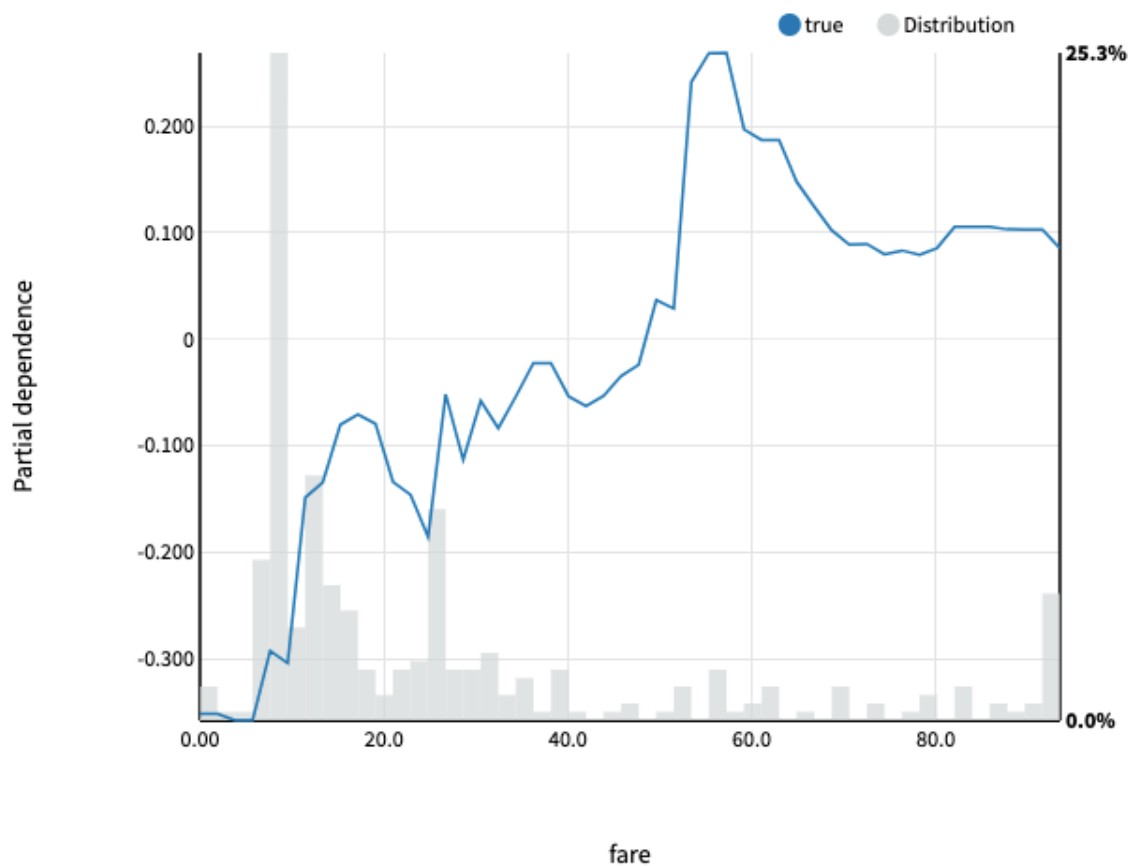
# fare

COMPUTE

COMPUTE ALL

EXPOR

50 bins for **fare**, computed on 312 rows (the full test set)



Rysunek 23. Partial dependence "fare" (Dataiku)

- „pclass” – klasa 3 (najmniej luksusowa) ma zdecydowanie mniejsze szanse na przeżycie niż klasa 1 (Rysunek 24),

## Partial dependence

Select your variable

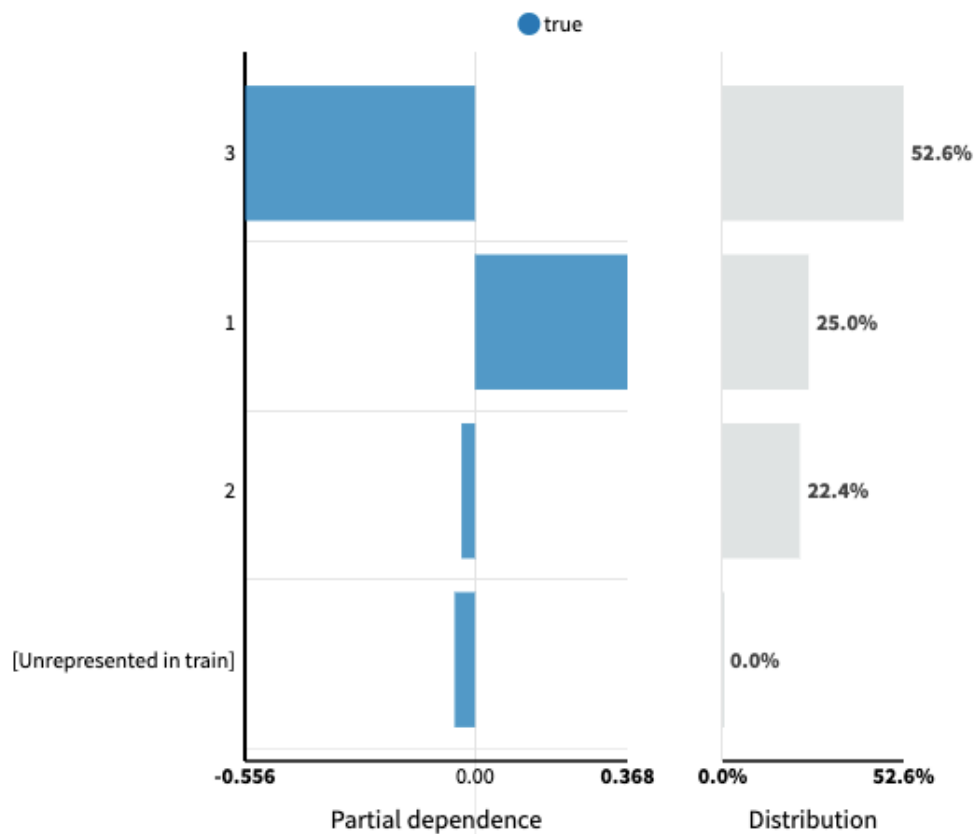
A pclass

COMPUTE

COMPUTE ALL

EXPOR

3 most frequent modalities of **pclass**, computed on 312 rows (the full test set)



Rysunek 24. Partial dependence "pclass" (Dataiku)

- „sex” – kobiety mają większe szanse na przeżycie niż mężczyźni (Rysunek 25),

## Partial dependence

Select your variable

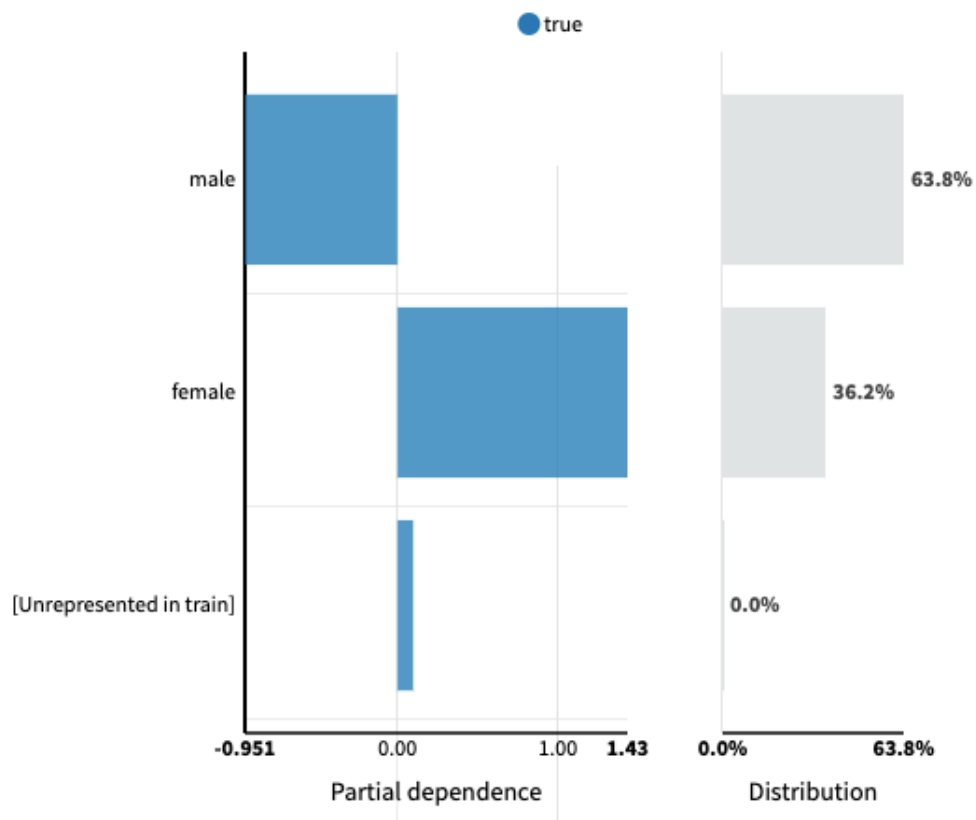
A sex

COMPUTE

COMPUTE ALL

EXPOR


2 most frequent modalities of **sex**, computed on 312 rows (the full test set)



Rysunek 25. Partial dependence "sex" (Dataiku)

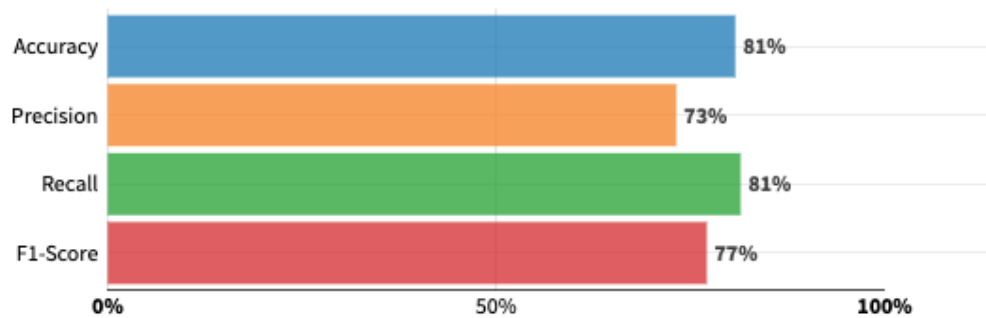
Poniższa macierz (Rysunek 26) porównuje rzeczywiste wartości zmiennej docelowej z wartościami przewidywanymi.

## Confusion matrix

Threshold (cut-off) 0 

Display: Record count ▼

	Predicted true	Predicted false	Total
Actually true	101	23	124
Actually false	37	151	188
Total	138	174	312



## Cost matrix

If model predicts true	and value is true	the gain is	1	×	101	=	101.00
	but value is false	the gain is	-0,3	×	37	=	-11.10
Model predicts false	and value is false	the gain is	0	×	151	=	0.00
	but value is true	the gain is	0	×	23	=	0.00
Average gain per record			0.29	×	312	=	89.90

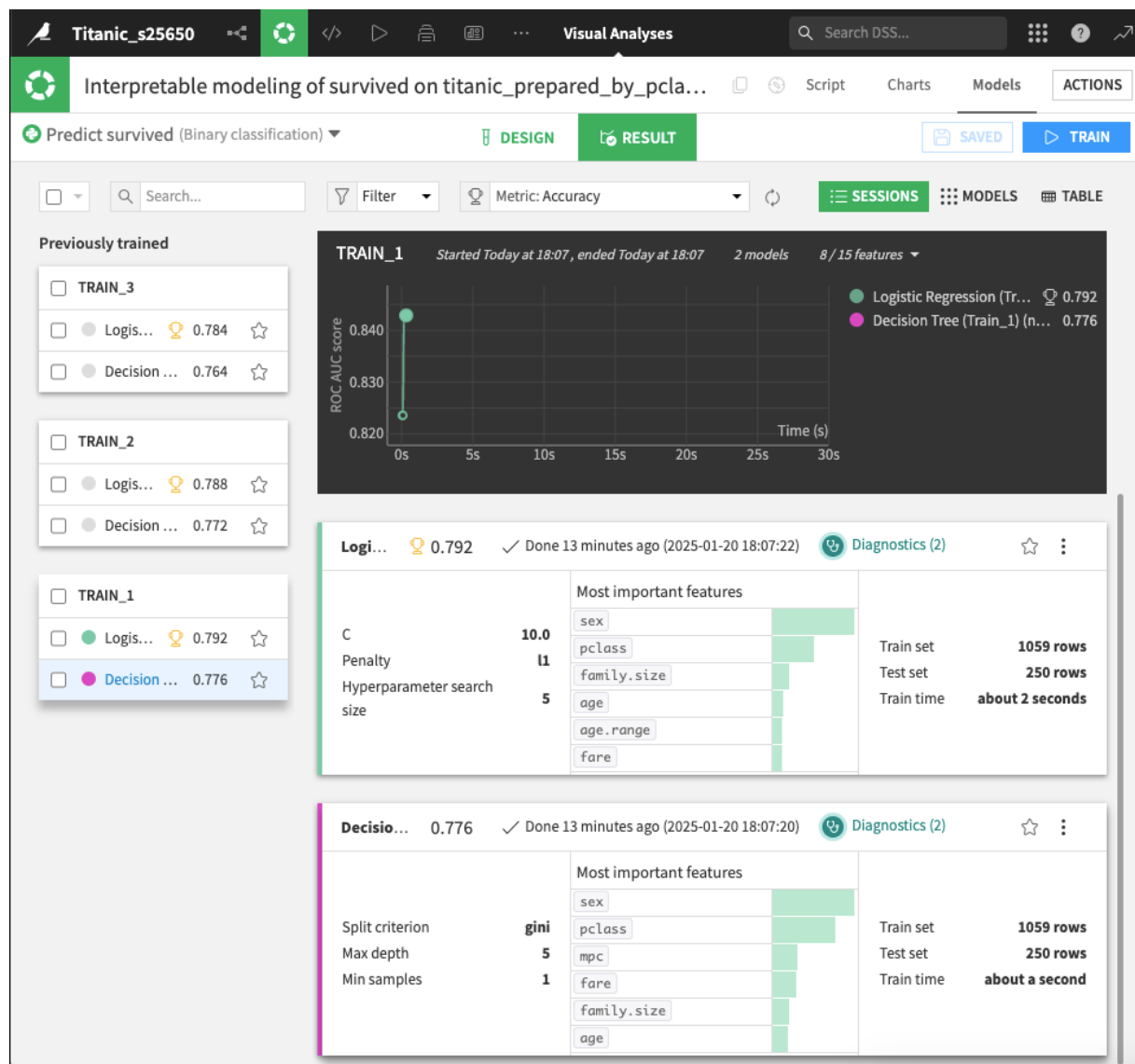
Rysunek 26. Confusion matrix (Dataiku)

## Interpretable modeling

Przetestowano także domyślne modele w zakładce Interpretable modeling (Logistic Regression i Decision Tree) z tym samym zestawem zmiennych jak w przykładzie powyżej oraz z niewielkimi modyfikacjami:

- Train\_1 – atrybuty z przykładu wyżej
- Train\_2 – usunięcie age z zestawu
- Train\_3 – powrót do age i usunięcie age\_range

Wyniki nie są satysfakcjonujące, więc dalej testowano modele.

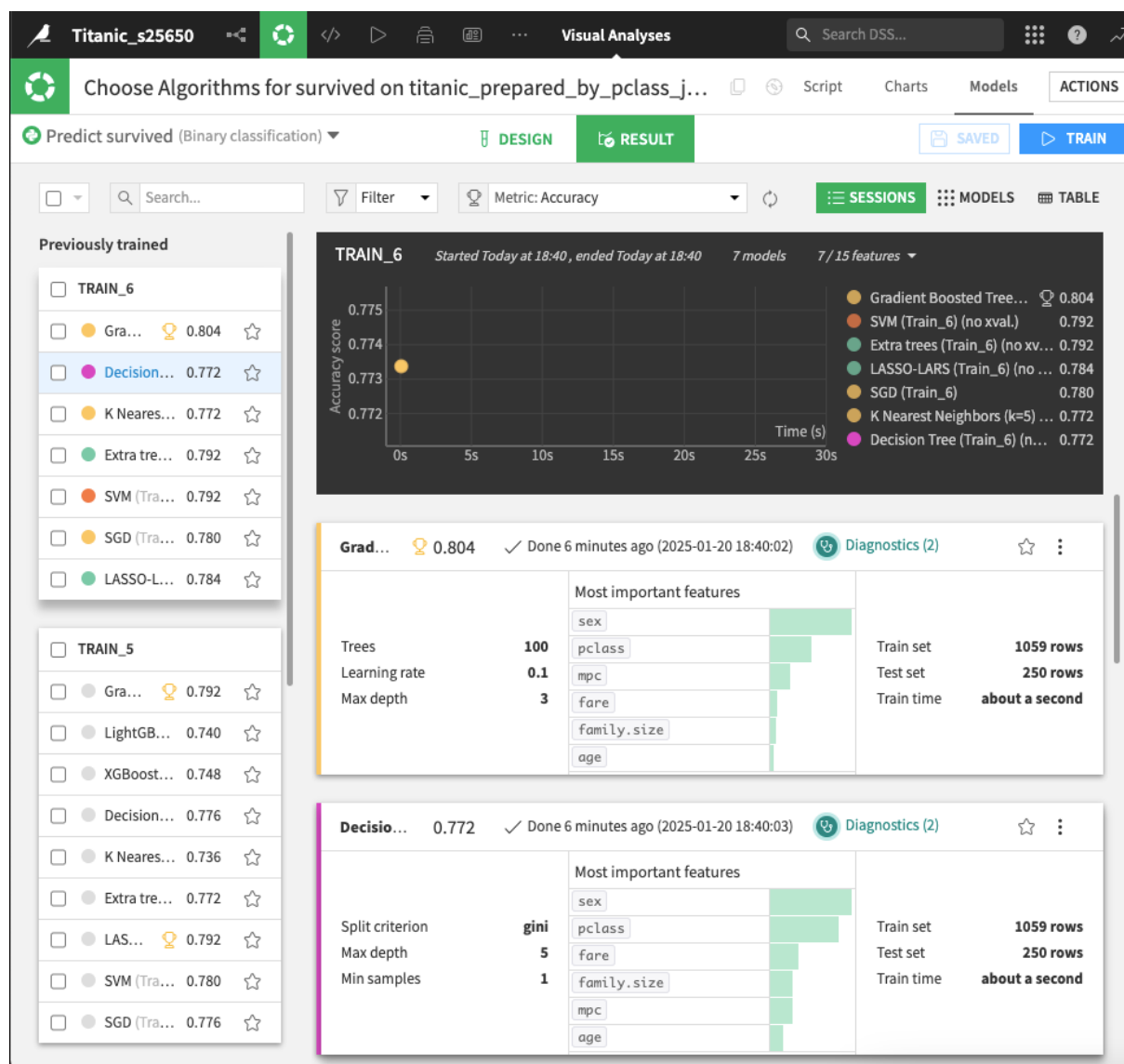


Rysunek 27. Interpretable modeling podsumowanie (Dataiku)



## Choose Algorithms

Dalej wykorzystano zakładkę, w której można dobrać więcej algorytmów, czyli „Choose Algorithms”. Najlepszy wynik okazał się bardzo zbliżony do Random Forest z „Quick Modeling”. Najlepiej wypadł Gradient Boosted Tree osiągając 80,4% przy atrybutach z Rysunku 28.



Rysunek 28. Wybrane algorytmy podsumowanie (Dataiku)

## Features

### Input features

Q

Search...

1-15 of 15

<

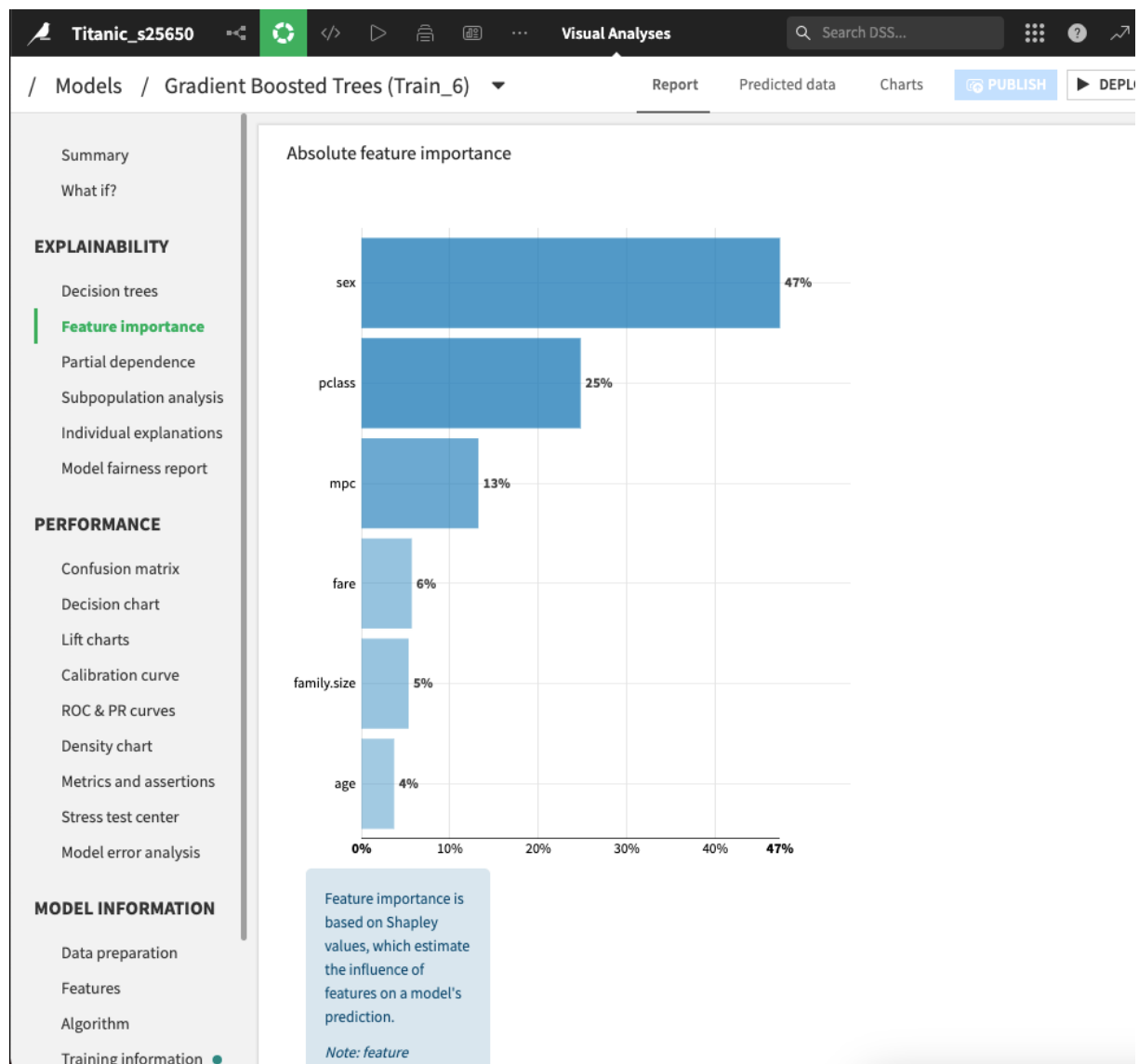
1

>

Feature Name	Role	Type	Summary
▼ family.size	➔ Input	# Numeric	Min-max rescaling
▼ mpc	➔ Input	# Numeric	Min-max rescaling
✖ -age.range	✖ Rejected	A Category	
▼ fare	➔ Input	# Numeric	Min-max rescaling
▼ sex	➔ Input	A Category	Dummy encoding
✖ -title1	✖ Rejected	A Category	
✖ -age_avg	✖ Rejected	# Numeric	
▼ survived	🎯 Target	A Category	
▼ pclass	➔ Input	A Category	Dummy encoding
✖ -sibsp	✖ Rejected	# Numeric	
✖ -name	✖ Rejected	I Text	
✖ -fare_avg	✖ Rejected	# Numeric	
✖ -embarked	✖ Rejected	A Category	
▼ age	➔ Input	# Numeric	Min-max rescaling
✖ -parch	✖ Rejected	# Numeric	

Rysunek 29. Features (Dataiku)

Podobnie jak w modelu z zakładki „Quick modeling”, najważniejsze atrybuty to sex, pclass i mpc. Jest to potwierdzenie, że te atrybuty mają znaczący wpływ na działanie wielu modeli i mogą stanowić podstawę do analizy.



Rysunek 30. Absolute feature importance (Dataiku)

## Podsumowanie

Na podstawie danych z pliku „titanic.csv” przeprowadzono analizę danych pasażerów statku Titanic i przewidziano prawdopodobieństwo przeżycia na podstawie dostępnych informacji oraz przy wykorzystaniu Dataiku DSS.

Wybrano model Random forest (las losowy) o poziomie trafności 80,8%, co oznacza, że dokładność jest na wysokim poziomie, a otrzymane wyniki wskazują na dobre rozróżnienie pasażerów i jasne określenie jakie cechy posiada osoba, która przeżyła lub nie.

Dzięki analizie atrybutów można zauważyć, że najważniejsze zmienne to płeć, klasa oraz mpc, czyli zmienna stworzona na podstawie wieku i klasy, z czego największy wpływ na przewidywania ma płeć. Kobiety, osoby młodsze oraz pasażerowie podróżujący w pierwszej klasie mieli największe szanse na przeżycie katastrofy.

# Spis rysunków

Rysunek 1. Informacje o niezmodyfikowanym pliku "titanic.csv" (Googe Colab).....	4
Rysunek 2. Univariate analysis 1 .....	4
Rysunek 3. Univariate analysis 2 .....	5
Rysunek 4. Univariate analysis 3 .....	5
Rysunek 5. Flow (Dataiku).....	6
Rysunek 6. Tabela wynikowa po grupowania po tytule (Dataiku).....	7
Rysunek 7. Kroki tworzenia dodatkowych zmiennych (Dataiku).....	8
Rysunek 8. Warunki wykorzystane przy budowie zmiennej age.range w Dataiku.....	8
Rysunek 9. Tabela wynikowa grupowania po "pclass" .....	9
Rysunek 10. Wykres punktowy dla zmiennej "age" (Dataiku).....	10
Rysunek 11. Porównanie średnich wieku do wartości odstających .....	10
Rysunek 12. Wykres punktowy zmiennej "fare" (Dataiku). .....	11
Rysunek 13. Macierz korelacji (Dataiku).....	11
Rysunek 14. Wyniki modelowania (Dataiku).....	12
Rysunek 15. Podsumowanie informacji o modelu (Dataiku) .....	13
Rysunek 16. Lista zmiennych z typem i podsumowaniem (Dataiku) .....	14
Rysunek 17. Informacje treningowe (Dataiku) .....	15
Rysunek 18. Szczegóły algorytmu i danych testowych (Dataiku) .....	15
Rysunek 19. Ważność funkcji (Dataiku).....	16
Rysunek 20. Ważność atrybutów (Dataiku) .....	17
Rysunek 21. Partial dependence (Dataiku) .....	18
Rysunek 22. Partial dependencie "age" (Dataiku).....	19
Rysunek 23. Partial dependence "fare" (Dataiku) .....	20
Rysunek 24. Partial dependence "pclass" (Dataiku) .....	21
Rysunek 25. Partial dependece "sex" (Dataiku).....	22
Rysunek 26. Confusion matrix (Dataiku) .....	23
Rysunek 27. Interpretable modeling podsumowanie (Dataiku) .....	24
Rysunek 28. Wybrane algorytmy podsumowanie (Dataiku) .....	25
Rysunek 29. Features (Dataiku) .....	26
Rysunek 30. Absolute feature importance (Dataiku).....	27