PSZT - Uczenie Maszynowe

Stawczyk Przemysław 293153, Piotr Zmyślony 268833

Contents

1 Opis Preprocesingu i Modelowania

1.1 Analiza Zbioru danych

1.1.1 Brakujące dane

Zaczęliśmy od analizy brakujących danych w wierszach. Jak widać w poniższych wynikach w większości zbiorów około połowa wierszy ma brakujące pola.

	rok 1	*rok 2*	*rok 3*	*rok 4*	*rok 5*
długość	7027	10173	10503	9792	5910
pełne wiersze	3194	4088	4885	4769	3031
brakujące dane	3833	6085	5618	5023	2879

Następnie przeprowadziliśmy analizę rozkładu brakujących danych w kolumnach i wierszach korzystając z biblioteki pythona *missingno* [fig 1-5]

Figure 1: rok 1

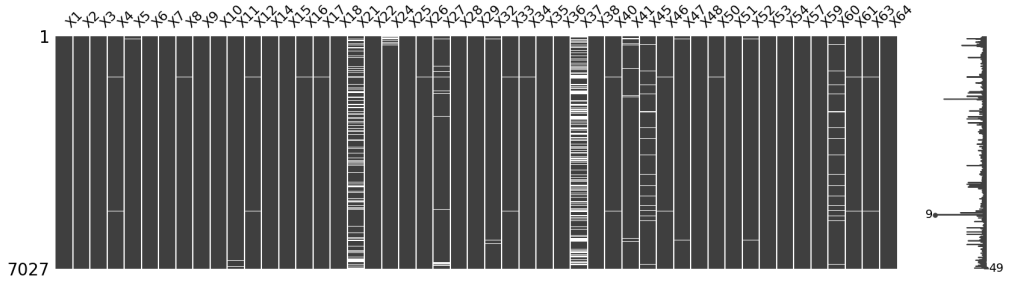


Figure 2: rok 2

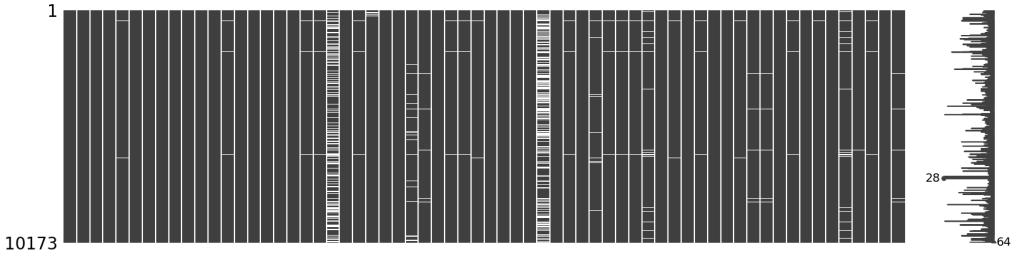
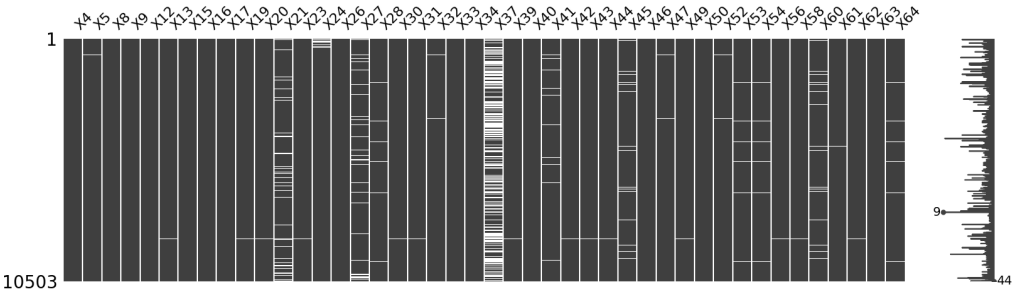


Figure 3: rok 3



Jak widać większość brakujących danych jest w kolumnie *X37*. Kolumna *X21* ma brakujące w niektórych ale nie wszystkich latach.

Figure 4: rok 4

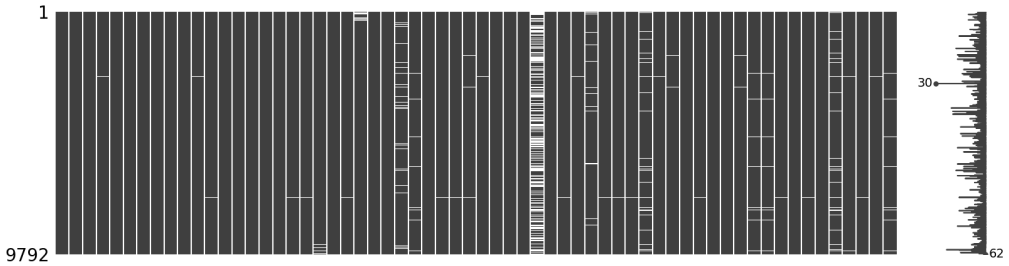
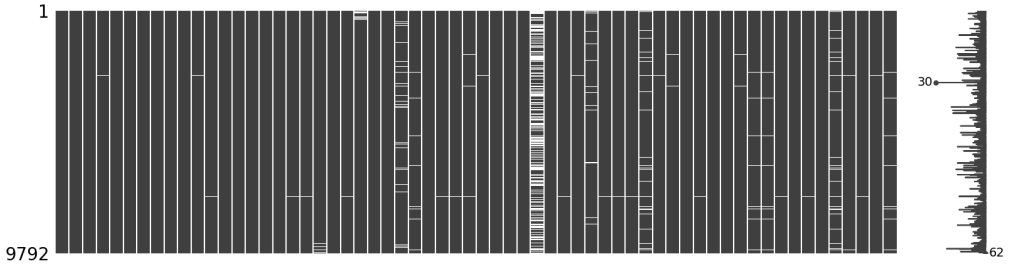


Figure 5: rok 5



Trudno nam było ocenić jaki charakter mają braki w tych danych, czy są skorelowane w wartościami w innych kolumnach czy zupełnie losowe. Wierszy z brakującymi danymi jest około połowy lub więcej. Aby nie odrzucać tak dużej liczby krotek zdecydowaliśmy się interpolować brakujące dane.
W tym celu wybraliśmy 4 metody:

1. Wstawianie średniej w danej kolumnie *(Jako punkt odniesienia)*
2. K najbliższych krotek
3. Spodziewanej Maksymalizacji *(Expected Maximalisation)*
4. Algorytm MICE

1.1.2 Zbalansowanie danych

Dokonaliśmy analizy ile z poszczególnych rekordów należy do klas klasyfikacyjnych

Czy zbankrutowano:	*rok 1*	*rok 2*	*rok 3*	*rok 4*	*rok 5*
Tak	6756	9773	10008	9277	5500
Nie	271	400	495	515	410
procent większości	3.857 %	3.932 %	4.713 %	5.259 %	6.937 %

Dane w zbiorach są mocno niezbalansowane dlatego zdecydowaliśmy się na interpolację korzystając z metody *SMOTE (Synthetic Minority Over Sampling Technique)*

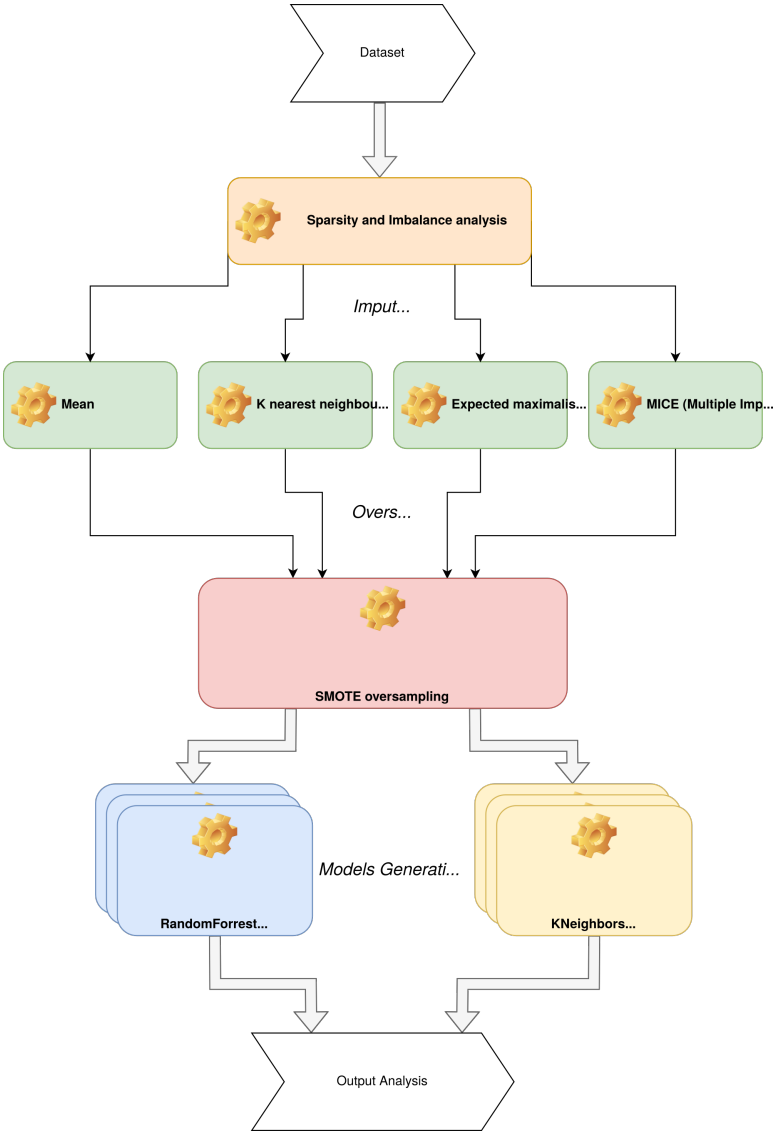
1.2 Przepływ Danych

Po powyższej analizie zdecydowaliśmy o następującym przepływie oryginalnych danych do konstrukcji modeli.

Walidacji modeli planujemy dokonać korzystając K-krotnej walidacji krzyżowej.

1.2.1 Wizualizacja Przepływu Danych

Figure 6: Przepływ Danych



2 Modele

Zgodnie z poleceniem wykorzystaliśmy algorytmy tworzenia modeli :

- Las Losowy *[RF - Random Forrest]*
- K Najbliższych sąsiadów *[KNN - K Nearest Neighbors]*

Implementacje wymienionych algorytmów pochodzą z biblioteki *sklearn*.

2.1 Parametry Modeli

3 Wyniki Eksperymentu

3.1 Wykresy

3.2 Interpretacja