Wojciech Żelasko

Nr indeksu: 406288

Case Study – Data Mining

Global_Education (education_data)

Dane:

Wylosowane dane przeznaczone do wykorzystania w projekcie dotyczą globalnego poziomu edukacji. Po wstępnej analizie można śmiało stwierdzić, że zawierają wiele ciekawych zmiennych, a w szczególności interesujące dane na temat poziomu dostępu do edukacji, poziomu rozwoju gospodarczego oraz czynników na to wpływających.

Kolumny w zbiorze danych zawierają następujące informacje:

- Countries and areas – kolumna zawiera informacje na temat nazw krajów lub obszarów na ziemi
- Latitude – kolumna zawiera dane informujące o szerokości geograficznej
- Longitude – kolumna zawiera dane informujące o długości geograficznej
- OOSR_Pre0Primary_Age_Male oraz Female – (Out-of-school) – wskaźnik osób spoza szkoły dla chłopców i dziewcząt w wieku przedszkolnym
- OOSR_Primary_Age_Male oraz Female - (Out-of-school) – wskaźnik osób spoza szkoły dla chłopców i dziewcząt w wieku podstawowo szkolnym
- OOSR_Lower_Secondary_Age_Male oraz Female - (Out-of-school) – wskaźnik osób spoza szkoły dla chłopców i dziewcząt w wieku ponadpodstawowym niższym szkolnym
- OOSR_Upper_Secondary_Age_Male oraz Female - (Out-of-school) – wskaźnik osób spoza szkoły dla chłopców i dziewcząt w wieku ponadpodstawowym szkolnym
- Completion_Rate_Primary_Male – wskaźnik ukończenia dla chłopców w szkole podstawowej
- Completion_Rate_Primary_Female – wskaźnik ukończenia dla dziewcząt w szkole podstawowej
- Completion_Rate_Lower_Secondary_Male -– wskaźnik ukończenia dla chłopców w szkole ponadpodstawowej niższej
- Completion_Rate_Lower_Secondary_Female – wskaźnik ukończenia dla dziewcząt w szkole ponadpodstawowej niższej
- Completion_Rate_Upper_Secondary_Male – wskaźnik ukończenia dla chłopców w szkole ponadpodstawowej
- Completion_Rate_Upper_Secondary_Female - wskaźnik ukończenia dla dziewcząt w szkole ponadpodstawowej

- Grade_2_3_Proficiency_Reading and Math – wskaźniki procentowe biegłości w czytaniu i matematyce dla klas 2-3
- Primary_End_Proficiency_Reading and Math – wskaźniki procentowe biegłości w czytania i pisania po zakończeniu edukacji
- Lower_Secondary_End_Proficiency_Reading – wskaźniki procentowe biegłości w czytaniu i matematyce na zakończenie edukacji średnio niższej
- Youth_15_24_Literacy_Rate_Male and Female – wskaźniki umiejętności czytania i pisania dla młodzieży w wieku 15-24 lat
- Birth Rate – liczba urodzeń na 1000 mieszkańców
- Gross_Primary_Education_Enrollment – Całkowita liczba uczniów w szkole podstawowej jako procent populacji szkolnej
- Gross_Tertiary_Education_Enrollment – Całkowita liczba studentów na poziomie szkolnictwa wyższego jako procent populacji szkolnej
- Unemployment_Rate – Procent ludności aktywnej zawodowo będącej bez stałego zatrudnienia.

Analizę i przygotowanie danych rozpoczynamy od wyświetlenia podstawowych statystyk dla naszego zbioru. Analizując dane można od razu zauważyć, że wiele rekordów w danych posiada w znacznej ilości wartości zerowe dla kolumn pomiędzy *OOSR_Pre0Primary_Age_Male* a *Youth_15_24_Literacy_Rate_Female*. Wiedząc, że każdy rekord danych odpowiada za inny kraj/region na świecie (*Kolumna Countries and areas* zawierająca nazwy państw oraz kolumny *Latitude* i *Longitude* odpowiadające za współrzędne geograficzne stolic), to na podstawie własnej wiedzy oraz informacji znalezionych w internecie, można bez większego problemu zauważyć, że wiele danych z powyższych kolumn, nie ma kompletnie sensu. Na *Rysunku 1* zostały przedstawione wskaźniki procentowe ilości wartości zerowych względem wszystkich rekordów w poszczególnych kolumnach.

```
> dim(edu_data)
[1] 202  29
> colSums(edu_data == 0, na.rm = TRUE) / nrow(edu_data) * 100
                     Countries.and.areas                                 Latitude                                Longitude
                                0.000000                                 0.000000                                 0.000000
             OOSR_Pre0Primary_Age_Male             OOSR_Pre0Primary_Age_Female                   OOSR_Primary_Age_Male
                               25.742574                                27.227723                                39.603960
             OOSR_Primary_Age_Female           OOSR_Lower_Secondary_Age_Male         OOSR_Lower_Secondary_Age_Female
                               42.079208                                36.633663                                37.128713
         OOSR_Upper_Secondary_Age_Male         OOSR_Upper_Secondary_Age_Female            Completion_Rate_Primary_Male
                               25.247525                                25.247525                                47.029703
       Completion_Rate_Primary_Female     Completion_Rate_Lower_Secondary_Male   Completion_Rate_Lower_Secondary_Female
                               47.029703                                47.029703                                47.029703
  Completion_Rate_Upper_Secondary_Male   Completion_Rate_Upper_Secondary_Female            Grade_2_3_Proficiency_Reading
                               47.029703                                47.029703                                64.851485
            Grade_2_3_Proficiency_Math          Primary_End_Proficiency_Reading             Primary_End_Proficiency_Math
                               70.297030                                78.217822                                75.742574
Lower_Secondary_End_Proficiency_Reading     Lower_Secondary_End_Proficiency_Math           Youth_15_24_Literacy_Rate_Male
                               57.425743                                54.950495                                60.891089
       Youth_15_24_Literacy_Rate_Female                              Birth_Rate       Gross_Primary_Education_Enrollment
                               60.891089                                 6.435644                                 7.425743
    Gross_Tertiary_Education_Enrollment                       Unemployment_Rate
                               9.405941                                12.871287
```

Rysunek 1. Ilość wartości zerowych w poszczególnych kolumnach (całkowita ilość wierszy wynosi 202)

Przykładowo, państwa bardzo słabo rozwinięte pod względem gospodarczym i edukacyjnym, rzędu Afganistanu, ma takie same wartości wskaźników ukończenia szkoły podstawowej przez chłopców oraz dziewcząt, co Kanada (państwo wysoko rozwinięte), i w obu przypadkach wynosi zero. Takich przykładów w danych mamy naprawdę dużo. Na *Rysunku 1* widoczne jest, że wiele kolumn posiada wartości zerowe rzędu osiągającego nawet 75% możliwych danych (kolumna *Primary_End_Proficiency_Math*).

Jest to niemal pewne, że miejsca w danych które zawierały wartości NA, zostały automatycznie zamienione na wartości zerowe podczas generowania danych. Trudność wykorzystania naszego zbioru pogarsza charakter danych, który pozwala porównywać dane pomiędzy poszczególnymi państwami oraz fakt, że w wielu kolumnach wartości zerowe mogą oznaczać bardzo dobre wyniki, gdzie z kolei w innych zerowe wartości oznaczają najgorsze wyniki.

Aby potwierdzić przypuszczenia dotyczące braku sensu analizy większości kolumn została wykonana podwójna klasteryzacja na podstawie algorytmu k-means. Najpierw na wybranych kolumnach dla których im mniejsze wartości oznaczały lepsze rezultaty, a następnie dla kolumn, którym lepsze wyniki dawały wartości wyższe.

```
> dd[,c(1,33,34)]
                  Countries.and.areas cluster_level_up cluster_level_low
1                         Afghanistan            Dobry             Słaby
2                             Albania     Bardzo Dobry      Bardzo Dobry
3                             Algeria     Bardzo Dobry      Bardzo Dobry
4                             Andorra            Słaby      Bardzo Dobry
5                              Angola     Bardzo Słaby      Bardzo Słaby
6                            Anguilla            Słaby      Bardzo Dobry
7               Antigua and Barbuda            Słaby      Bardzo Dobry
8                           Argentina     Bardzo Dobry      Bardzo Dobry
9                             Armenia     Bardzo Słaby      Bardzo Słaby
10                          Australia            Słaby      Bardzo Dobry
11                            Austria            Słaby      Bardzo Dobry
12                         Azerbaijan            Słaby      Bardzo Dobry
13                        The Bahamas            Słaby             Słaby
14                            Bahrain            Dobry      Bardzo Słaby
15                         Bangladesh     Bardzo Dobry             Słaby
16                           Barbados     Bardzo Słaby      Bardzo Dobry
17                            Belarus     Bardzo Dobry      Bardzo Dobry
18                            Belgium            Słaby      Bardzo Dobry
19                             Belize     Bardzo Słaby             Słaby
20                              Benin            Dobry             Słaby
```

Rysunek 2. Wstępne podwójne sklasteryzowanie danych

Na *Rysunku 2* możemy zauważyć, że nadmierne wartości zerowe zaburzyły nasze dane. Im lepszy poziom, tym powinien być wyższy poziom edukacji w poszczególnych państwach. Niestety od razu widoczne jest, że np. kraje nisko rozwinięte typu Bengladesz czy Afganistan, mają lepsze poziomy niż Belgia czy Austria, które są krajami wysoko rozwiniętymi.

Na podstawie powyższej analizy postanowiłem wyodrębnić tylko te kolumny w danych, które posiadają nie więcej niż 25 procent danych z wartościami zerowymi. Wybrałem następujące kolumny:

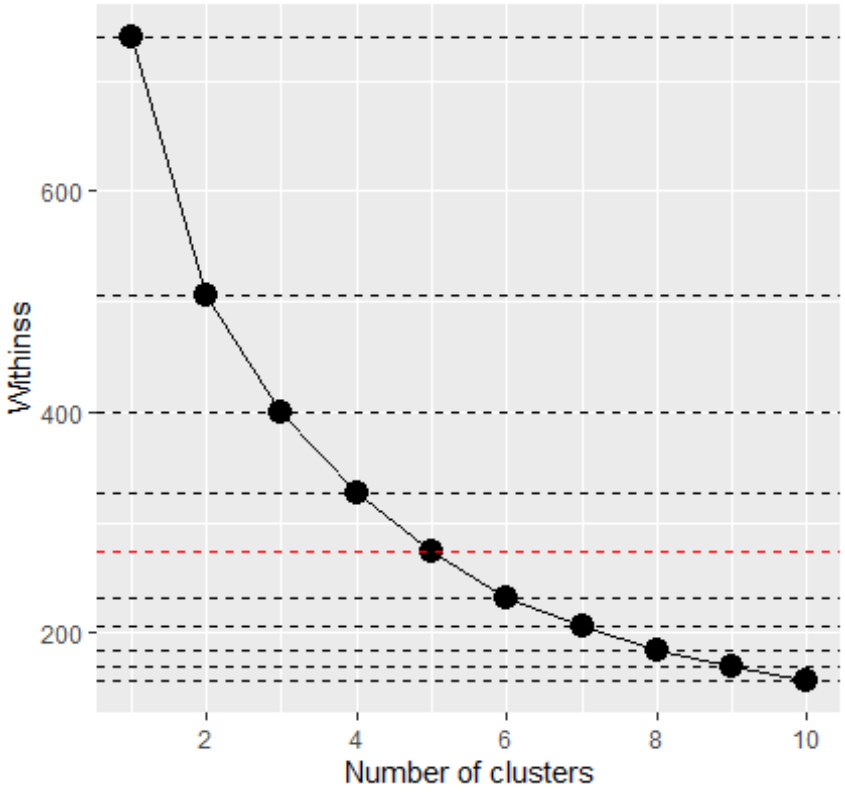
Countries.and.areas , Latitude, Longitude, Birth_Rate, Gross_Primary_Education_Enrollment, Gross_Tertiary_Education_Enrollment, Unemployment_Rate.

```
> dim(selected_edu_data)
[1] 202   7
> colSums(selected_edu_data == 0, na.rm = TRUE) / nrow(selected_edu_data) * 100
             Countries.and.areas                      Latitude                     Longitude
                        0.000000                      0.000000                      0.000000
                      Birth_Rate Gross_Primary_Education_Enrollment
                        6.435644                      7.425743
Gross_Tertiary_Education_Enrollment                 Unemployment_Rate
                        9.405941                     12.871287
```

Rysunek 3. Odsetek wartości zerowych w wyodrębnionych kolumnach.

Na *Rysunku 3* wybrane kolumny zawierają znacznie mniej wartości zerowych. Postanowiłem użyć części powyższych kolumn, w celu wykonania klasteryzacji na 5 poziomów rozwoju gospodarczego. Użyte kolumny to *Birth_Rate, Gross_Primary_Education_Enrollment, Gross_Tertiary_Education_Enrollment* oraz *Unemployment_Rate.*

Najpierw usunąłem wiersze, w których znajdowała się więcej niż jedna wartość zerowa w celu uniknięcia anomalii. Następnie przeprowadziłem normalizacje danych w kolumnach za pomocą funkcji scale(). Dzięki temu wyeliminowałem różnice w dysproporcjach pomiędzy wartościami poszczególnych zmiennych. Aby dobrać odpowiednią liczbę klastrów w funkcji kmeans(), porównałem w pętli wartości „withinss" dla liczby klastrów od 1 do 10. Na *Rysunku 4* zostały zwizualizowane wyniki (czerwona przerywana linia wyznacza optymalną liczbę klastrów do dalszej analizy).



Rysunek 4. Wykres ilości klastrów od wskaźnika Withnss.

Dzięki zastosowanym wyżej operacjom, klasteryzacja przebiegła zadawalająco. Poszczególne poziomy rozwoju gospodarczego zostały przydzielone bardzo sensownie pod względem rzeczywistego stopnia rozwoju gospodarczego poszczególnych państw. *Rysunek 5* przedstawia

fragment danych zawierający dwie kolumny, nazwę państwa oraz przydzielony mu poziom rozwoju gospodarczego (1-najnizszy, 5-najwyższy).

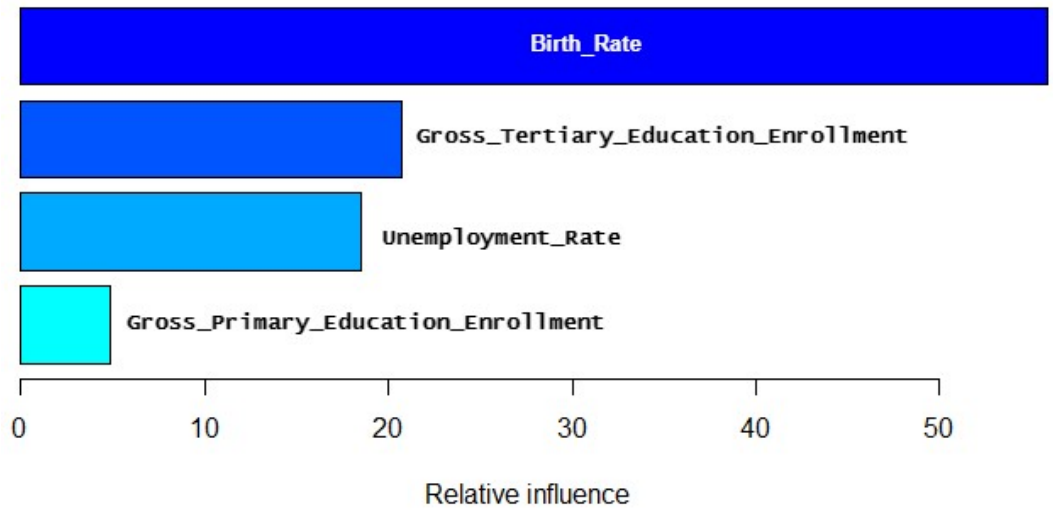
	Countries.and.areas	economic_level
1	Afghanistan	1
2	Albania	5
3	Algeria	2
5	Angola	1
7	Antigua and Barbuda	4
8	Argentina	5
9	Armenia	2
10	Australia	5
11	Austria	5
12	Azerbaijan	4
13	The Bahamas	4
14	Bahrain	4
15	Bangladesh	4

Rysunek 5. Fragment danych z przeprowadzonej klasteryzacji k-means

GBM (Gradient Boosting Machines)

Przed przystąpieniem do modelowania data set został podzielony na zbiór testowy i treningowy. Dane zostały podzielone w stosunku 0.8 do 0.2 (80% zbiór treningowy, 20% zbiór testowy). Naszym głównym celem było zbadanie efektywności modelu GBM do przewidywania stopnia rozwoju gospodarczego państwa.

Parametry modelu zostały ustawione tak, aby w zależności od kolumn *Birth_Rate, Gross_Primary_Education_Enrollment, Gross_Tertiary_Education_Enrollment* oraz *Unemployment_Rate* model uczył się przewidywać *economic_level*.



Rysunek 6. Wykres pokazujący, które zmienne miały jaki wpływ procentowy na wytrenowanie modelu.

Następnie za pomocą funkcji predict() została wykonania testowa predykacja z użyciem wcześniej wytrenowanego modelu. Konieczna okazała się także funkcja factor() w celu skategoryzowania zmiennych do tych samych poziomów, które posiada zmienna *economic_level*. Czynność ta okazała się niezbędna w celu wykonania macierzy pomyłek (*Rysunek 7*).

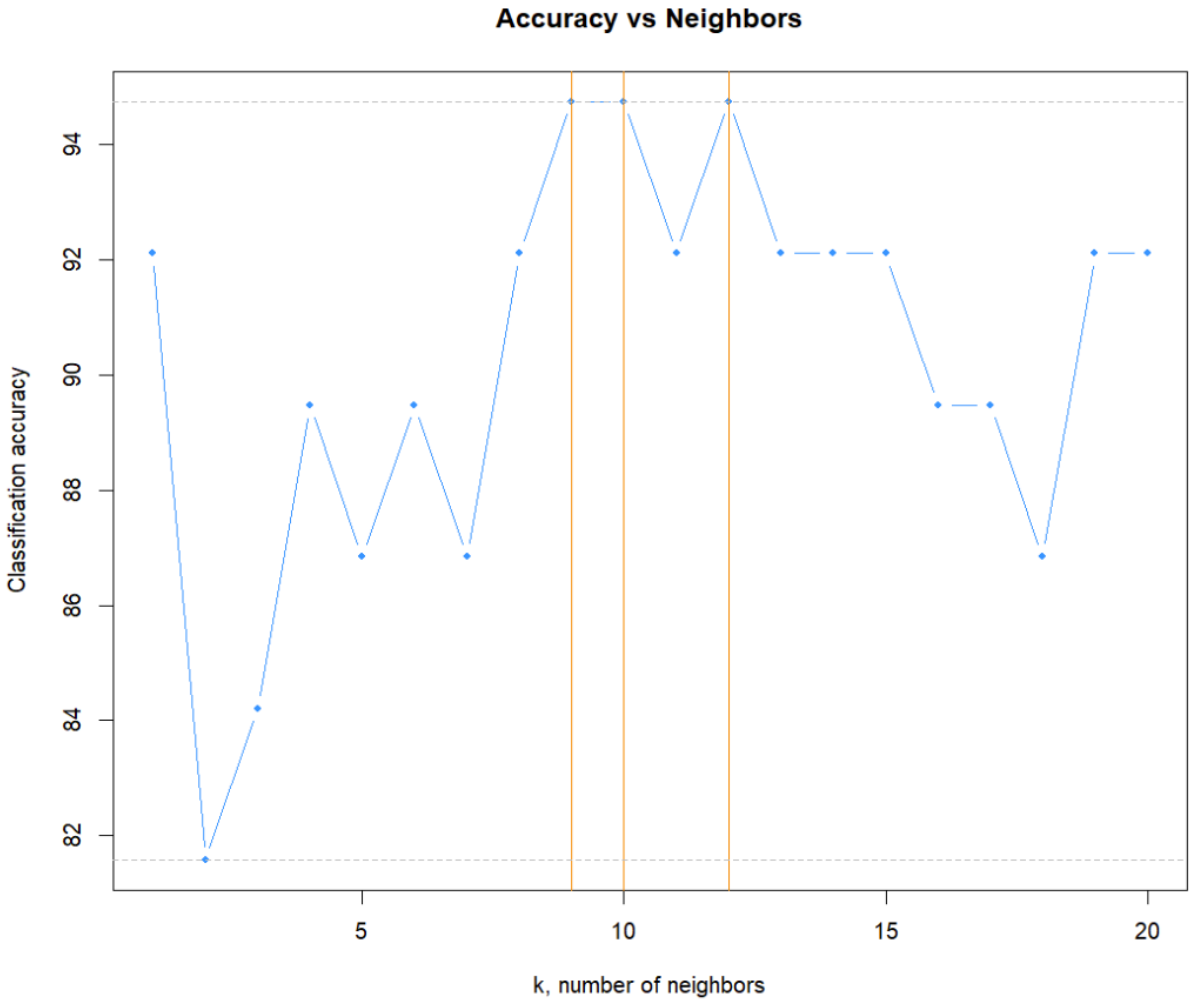
```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5
         1  6  0  0  0  0
         2  2  3  0  1  0
         3  0  1  0  1  1
         4  0  0  1 10  4
         5  0  0  0  0  8

Overall Statistics

               Accuracy : 0.7105
                 95% CI : (0.541, 0.8458)
    No Information Rate : 0.3421
    P-Value [Acc > NIR] : 3.967e-06

                  Kappa : 0.6147

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.7500  0.75000  0.00000   0.8333   0.6154
Specificity            1.0000  0.91176  0.91892   0.8077   1.0000
Pos Pred Value         1.0000  0.50000  0.00000   0.6667   1.0000
Neg Pred Value         0.9375  0.96875  0.97143   0.9130   0.8333
Prevalence             0.2105  0.10526  0.02632   0.3158   0.3421
Detection Rate         0.1579  0.07895  0.00000   0.2632   0.2105
Detection Prevalence   0.1579  0.15789  0.07895   0.3947   0.2105
Balanced Accuracy      0.8750  0.83088  0.45946   0.8205   0.8077
```

Rysunek 7. Macierz pomyłek dla predykcji z wykorzystaniem modelu GBM

Model poprawnie przewidział zdecydowaną większość przykładów klas ze zbioru testowego. Najwięcej pomyłek posiada dla klasy 1 oraz 5. Dokładność wynosiła 0.7105, co oznacza, że 71.05% próbek w zbiorze testowym zostało sklasyfikowane poprawnie. Także wartość Kappa wynosząca 0.6147 wskazuje na dobrą jakość klasyfikacji. Wysokie wartości dla Sensitivity (dokładność), Specificity (specyficznosc) oraz Pos/Neg Pred Value również wskazują na wysoką skuteczność modelu. Zatem biorąc pod uwagę powyższe wnioski, model poprawnie przewiduję na zadawalającym poziomie.

KNN (K-Nearest Neighbors)

Kolejnym zastosowanym modelem okazała się metoda najbliższych sąsiadów. Zostały użyte te same kolumny co w trakcie trenowania modelu wyżej. Do wybrania najlepszej ilości sąsiadów stworzyliśmy pętle, która wywoływała w pętli funkcje knn(), kolejno dla liczby sąsiadów od 1 do 20 i zapisywała wyniki procentowe parametru *accuracy*. Następnie wyniki zostały przedstawione na poniższym wykresie *Rysunek 8.*



Rysunek 8. Wykres wartości procentowych accuracy w zależności od ilości sąsiadów w algorytmie knn.

Została wybrana liczba sąsiadów jako k=9, która została użyta w modelu. Po wytrenowaniu modelu, a następnie sfaktoryzowaniu zmiennej *economic_level* została wykonana macierz

pomyłek (*Rysunek 9*), która wskazuje, że dokładność (accuracy) osiąga prawie 95 procent, co czyni model wysoko zadawalającym.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5
         1  8  0  0  0  0
         2  0  4  0  0  0
         3  0  0  0  0  0
         4  0  0  1 11  0
         5  0  0  0  1 13

Overall Statistics

               Accuracy : 0.9474
                 95% CI : (0.8225, 0.9936)
    No Information Rate : 0.3421
    P-Value [Acc > NIR] : 5.312e-15

                  Kappa : 0.9268

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   1.0000  0.00000   0.9167   1.0000
Specificity            1.0000   1.0000  1.00000   0.9615   0.9600
Pos Pred Value         1.0000   1.0000      NaN   0.9167   0.9286
Neg Pred Value         1.0000   1.0000  0.97368   0.9615   1.0000
Prevalence             0.2105   0.1053  0.02632   0.3158   0.3421
Detection Rate         0.2105   0.1053  0.00000   0.2895   0.3421
Detection Prevalence   0.2105   0.1053  0.00000   0.3158   0.3684
Balanced Accuracy      1.0000   1.0000  0.50000   0.9391   0.9800
```

Rysunek 9. Macierz pomyłek dla modelu KNN

Model poprawnie przewidział prawie wszystkie przykłady klas ze zbioru testowego. Zauważalne zaledwie dwie pomyłki dla klasy 3 oraz 4. Dokładność wynosiła 0.9474, co oznacza, że 94.74% próbek w zbiorze testowym zostało sklasyfikowane poprawnie. Także wartość Kappa wynosząca 0.9268 wskazuje na bardzo wysoką jakość klasyfikacji. Wysokie wartości dla Sensitivity (dokładność), Specificity (specyficzność) oraz Pos/Neg Pred Value również wskazują na wysoką skuteczność modelu. Zatem biorąc pod uwagę powyższe wnioski, model przewiduje na bardzo zadawalającym poziomie.

Na podstawie powyższych wyników obydwu modeli, możemy stwierdzić, że oba modele spisały się bardzo dobrze z odpowiednio dobranymi parametrami. Przewidywały one klasy rozwoju poszczególnych państw na zadawalającym poziomie. Nie wiemy jak mogły by się zachowywać modele wytrenowane na większej ilości danych ale z pewnością zauważalne jest, że model KNN sprawdził się lepiej w porównaniu do modelu GBM.