

## Лабораторная работа №2

Студент: Лунгу Н. А.

Группа: М8О-304Б

### Постановка задачи:

Необходимо реализовать алгоритмы машинного обучения. Применить данные алгоритмы на наборы данных, подготовленных в первой лабораторной работе. Провести анализ полученных моделей, вычислить метрики классификатора. Произвести тюнинг параметров в случае необходимости. Сравнить полученные результаты с моделями реализованными в scikit-learn. Аналогично построить метрики классификации. Показать, что полученные модели не переобучились. Также необходимо сделать выводы о применимости данных моделей к вашей задаче.

- 1) ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ
- 2) KNN
- 3) SVM
- 4) ДЕРЕВО РЕШЕНИЙ
- 5) RANDOM FOREST

Я реализовал три алгоритма на платформе azure machine learning:

- 1)линейная регрессия
- 2)мультиклассовая логическая регрессия
- 3)дерево решений

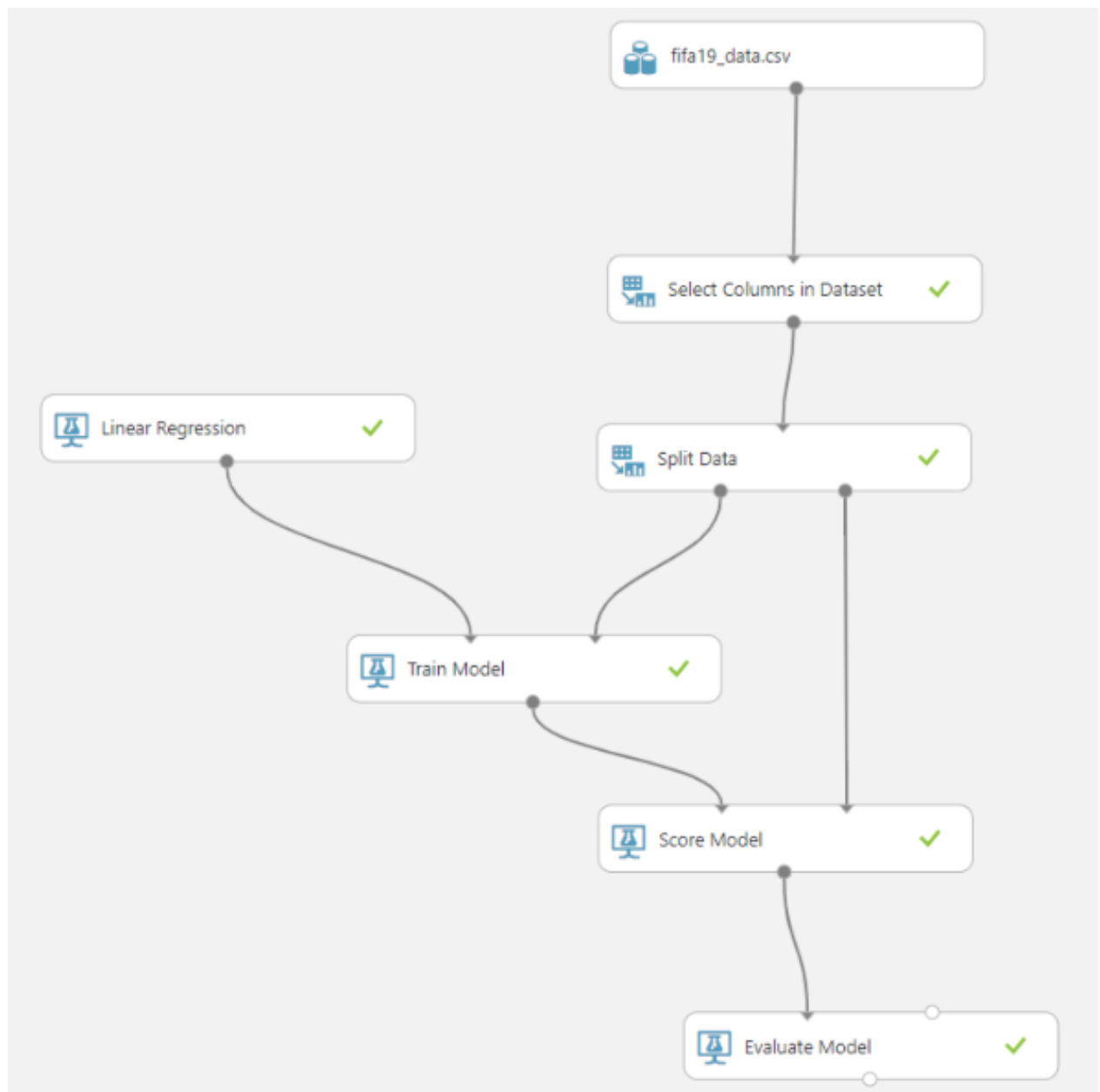
## **Задача 1**

Предсказание рейтинга футболиста на основе возраста, национальности, ударной ноги, способности “слабой ноги”, навыков обращения с мячом.








Мною был использован алгоритм линейной регрессии, так как он является одним из самых простых методов прогнозирования.

Из датасета я выделил 80% данных для обучающей выборки.

## **Модель эксперимента**



Полученные оценки

Age	Nationality	Overall	Preferred Foot	Weak Foot	Skill Moves	Scored Labels
						
33	England	67	Right	3	3	71.289095
19	Germany	63	Right	2	2	61.233413
25	Germany	64	Right	4	2	66.272451
29	China PR	62	Right	3	1	58.843164
23	England	61	Right	3	2	62.008422
21	Congo	57	Right	3	2	61.573488

Здесь можно видеть рейтинг, который был в датасете, и предсказанный рейтинг, полученный в ходе расчёта.

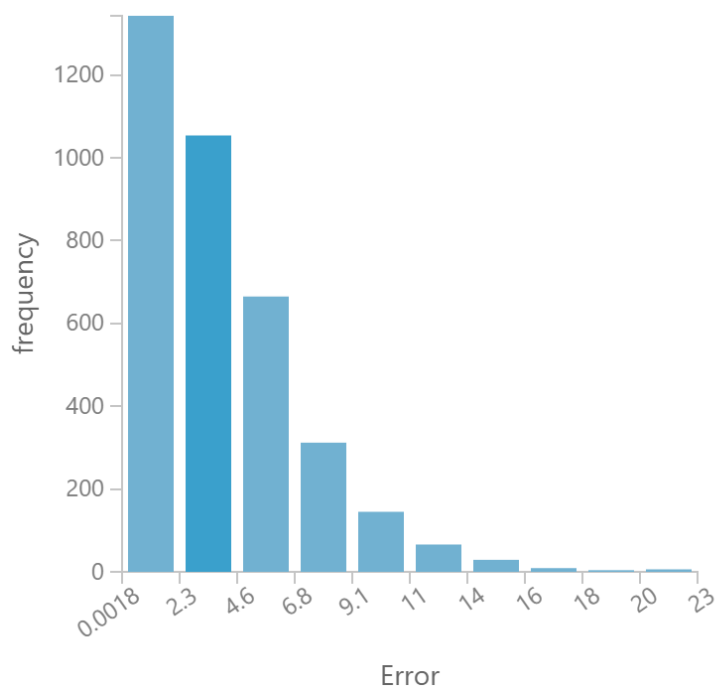
Метрики

Metrics

Mean Absolute Error	3.913915
Root Mean Squared Error	5.04513
Relative Absolute Error	0.724007
Relative Squared Error	0.540954
Coefficient of Determination	0.459046

Средняя абсолютная ошибка, среднеквадратичная ошибка, относительные ошибки и коэффициент смешанной корреляции (детерминированности).

## Ошибки и их частоты

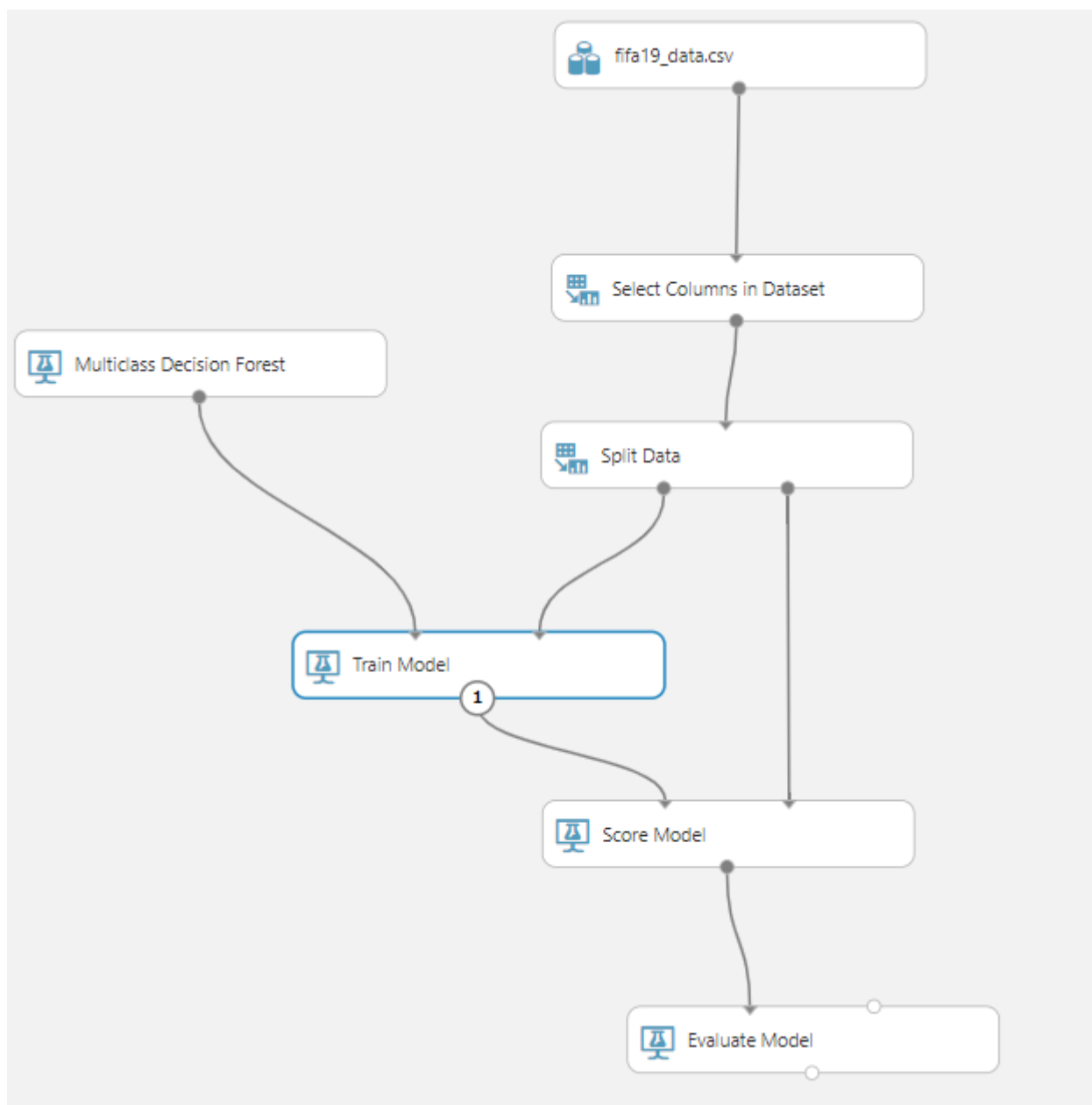


## Задача 2







Мультиклассовая классификация позиции, на которой играет игрок. Из датасета взяты: национальность, возраст, общий рейтинг, «рабочая» нога, «слабая» нога, умение владения мячом, интенсивность передвижения, позиция и большое количество умений игрока с оценкой от 0 до 100.

Для классификации я взял дерево решений, так как оно просто для представления, строится в самом Azure, не требует предварительной подготовки данных и работает со всеми типами. Я выделил 80% датасета для обучения.

## Модель эксперимента

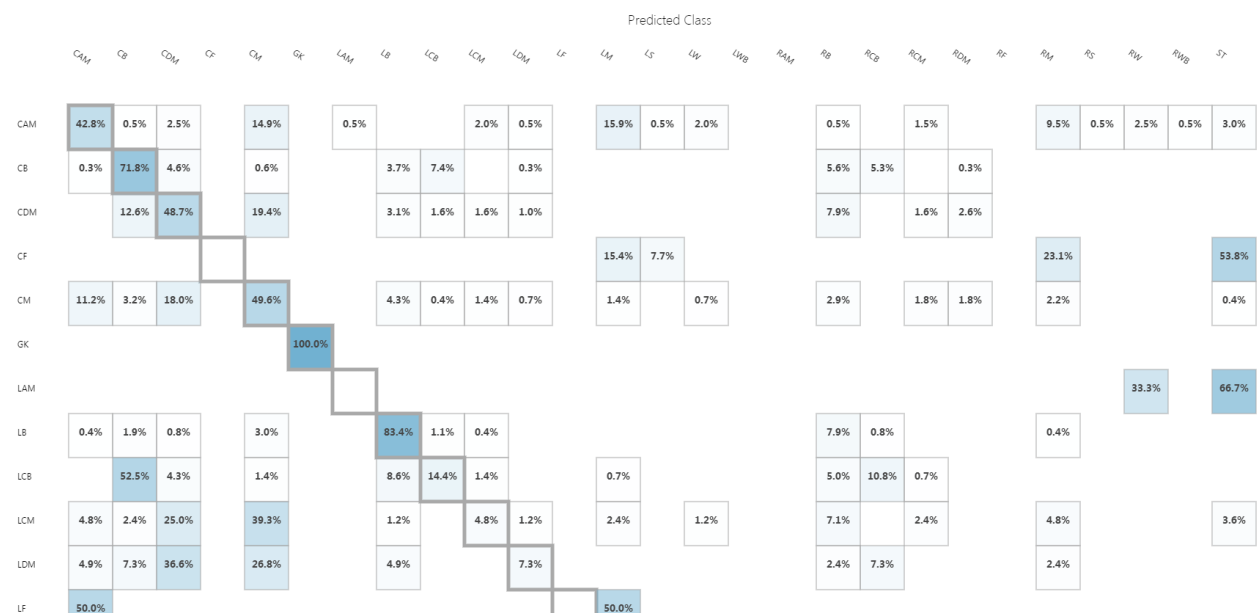


## Полученные оценки

rows	columns					
3641	71					
ties	Scored Probabilities for Class "CDM"	Scored Probabilities for Class "CF"	Scored Probabilities for Class "CM"	Scored Probabilities for Class "GK"	Scored Probabilities for Class "LAM"	Scored Probabilities for Class "LB"
						
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	1	0	0

В результате мы имеем оценённые вероятности каждой позиции и позицию, которая наиболее вероятна.

## Матрица ошибок



## Метрики

### Metrics

---

Overall accuracy	0.51799
Average accuracy	0.965571
Micro-averaged precision	0.51799
Macro-averaged precision	NaN
Micro-averaged recall	0.51799
Macro-averaged recall	0.233337

Общая точность, средняя точность, микро- и макро- усреднённая точность и полнота.

## Задача 3

Классификация национальности игрока. Из датасета взял: возраст, национальность, общий рейтинг, клуб, «сильную» ногу, «слабую» ногу, умение владения мячом, тип телосложения, позицию, возраст, вес, всевозможные умения, классифицированные от 0 до 100. Выделил 80% данных на обучение.

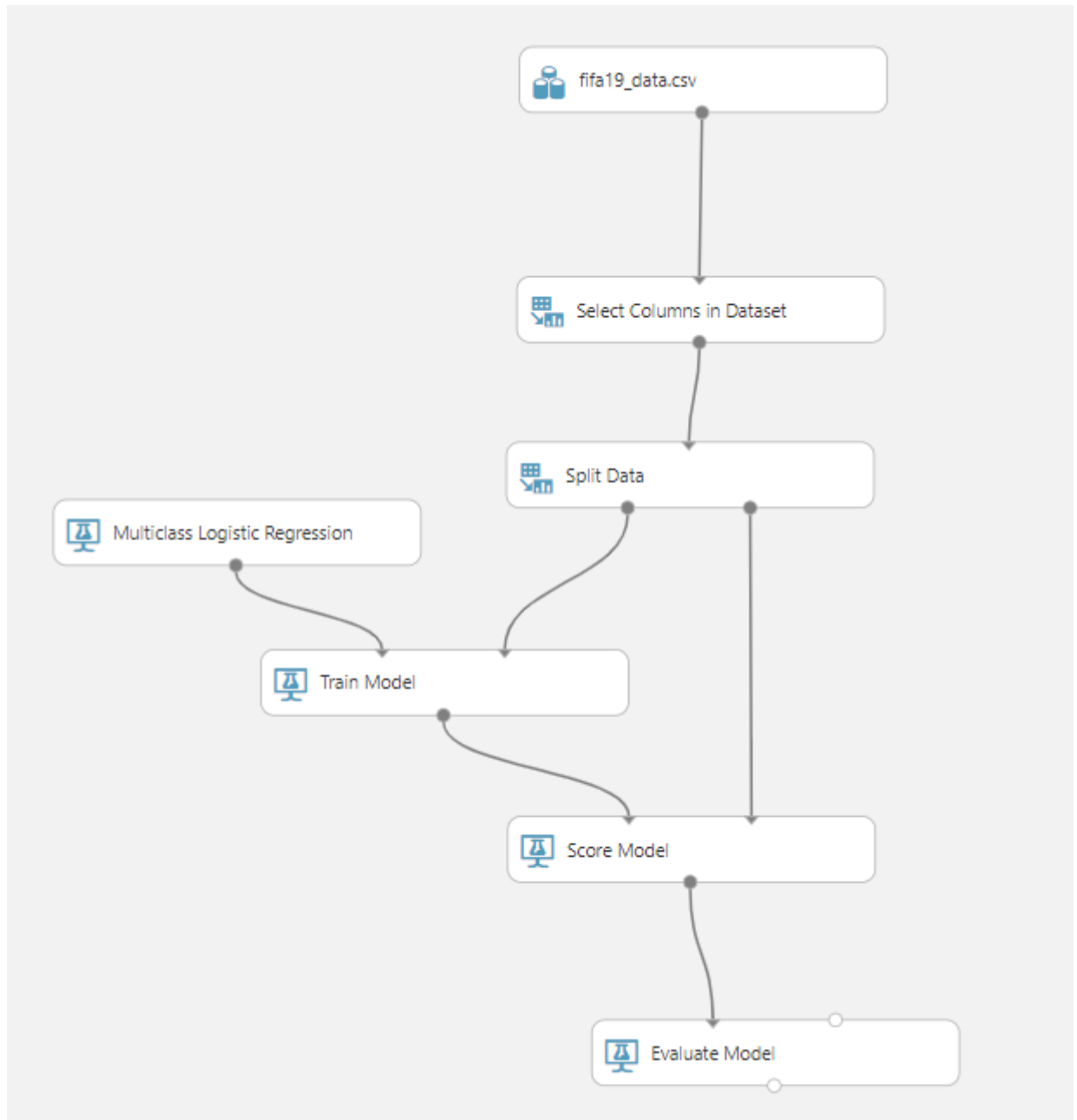
Для классификации выбрал мультиклассовую логистическую регрессию, поскольку она хорошо подходит для решения подобных задач:

1. В какой стране фирма будет располагать офисом, учитывая характеристики фирмы и различных стран-кандидатов?
2. Какой тип крови у человека, учитывая результаты различных диагностических тестов?

В данном случае я хочу понять национальность игрока в зависимости от различных параметров, описанных выше.



## Модель эксперимента



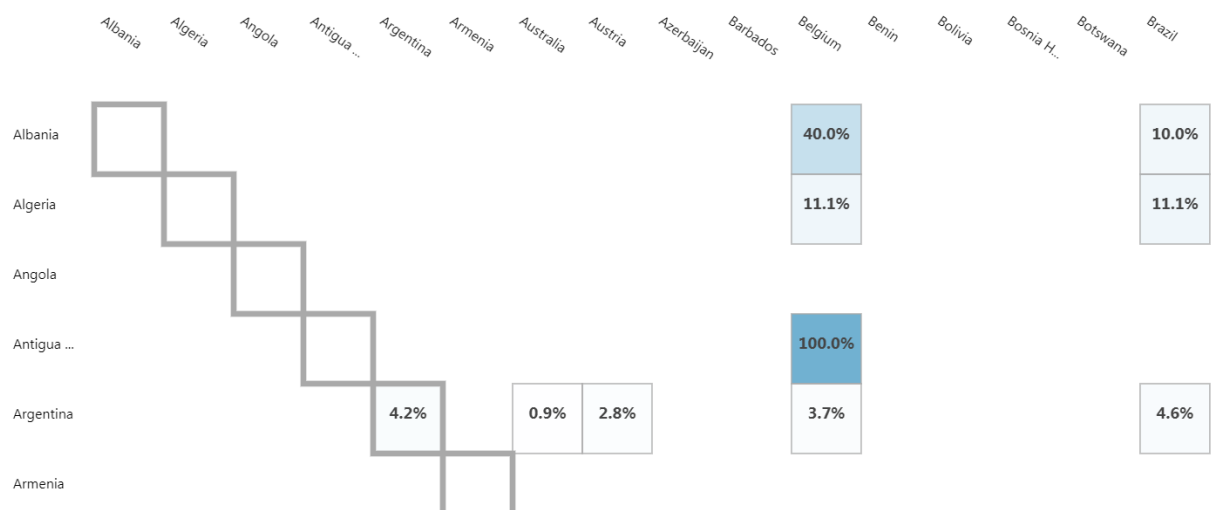
## Полученные оценки

rows	columns
3641	206

Scored Probabilities for Class "Congo"	Scored Probabilities for Class "Ecuador"	Scored Probabilities for Class "Egypt"	Scored Probabilities for Class "El Salvador"	Scored Probabilities for Class "England"	Scored Probabilities for Class "Equatorial Guinea"	Scored Probabilities for Class "Eritrea"
0.000000	0.002624	0.001228	0.000679	0.453425	0.000339	0.000000
0.000000	0.001641	0.001387	0.000288	0.050189	0.000419	0.000000
0.000000	0.000715	0.000624	0.000122	0.0149	0.000178	0.000000

В данной таблице получилось очень много столбцов из-за количества стран, но если внимательно на неё посмотреть, то видно, что у правильной национальности вероятность гораздо больше, чем у альтернативных. У игрока в первой строке национальность – Англичанин, и вероятность, как видно на картинке, очень велика.

## Матрица ошибок



## Метрики

### Metrics

---

Overall accuracy	0.03955
Average accuracy	0.988358
Micro-averaged precision	0.03955
Macro-averaged precision	NaN
Micro-averaged recall	0.03955
Macro-averaged recall	NaN

## Вывод

Эта лабораторная работа оказалось довольно полезной, так как благодаря ей можно было довольно быстро и просто ознакомиться с основными алгоритмами машинного обучения. Также для написания самих алгоритмов хорошо бы посмотреть, как они работают на примере готовых данных и какие результаты выдают, чтобы можно было уже сверять со своими и вносить необходимые правки. Также я познакомился с Azure ML – платформой, которая быстро и легко, без настройки и развёртывания на локальной машине, позволяет выполнять задачи машинного обучения в облаке.