American Visa Application Checker

Implementation Document

**Data Science Project**

**Institute of Technology, Carlow**

**By:** Jordan Lillis, Ian Mac Mathuna

**Student No:** C00204110, C00206527

**Lecturer:** Greg Doyle

**Date:** 19/02/2019

Contents

[Business Understanding 1](#_Toc2088003)

[Business Objective 1](#_Toc2088004)

[Assess Situation 1](#_Toc2088005)

[Data Mining Goals 1](#_Toc2088006)

[Project Plan 1](#_Toc2088007)

[Data Understanding 2](#_Toc2088008)

[Collect Data 2](#_Toc2088009)

[Describe Data 2](#_Toc2088010)

[Data Preparation 3](#_Toc2088011)

[Clean Data 3](#_Toc2088012)

[Format Data 3](#_Toc2088013)

[Analysis of Visas 4](#_Toc2088014)

[Data 4](#_Toc2088015)

[Naive Bayes (Economic Sector VS Case Status) Results 4](#_Toc2088016)

[Naive Bayes (Country of Citizenship vs Case Status) Results 6](#_Toc2088017)

[Decision Tree 8](#_Toc2088018)

[Assessment 8](#_Toc2088019)

[Recommendations 9](#_Toc2088020)

# Business Understanding

## Business Objective

The objective of this project is to allow people, who plan on applying for an American Visa, to gauge whether or not their application will be successful. Our project aims to only take up to five minutes to give a result with over 80% accuracy.

## Assess Situation

Currently there is no way to check whether a Visa will be accepted or denied without actually going through the process of filling out a full application. This takes a lot of time and also costs a fee.

## Data Mining Goals

* Predict the probability of a Visa application being accepted
* Find what factors are most influential when applying for a Visa. Eg. Country of citizenship.

## Project Plan

We will clean our dataset which we downloaded from Kaggle.com by taking out null values, duplicates and misspellings. We will then import our data into RapidMiner year by year and try to find some correlation between some columns such as the applicant’s economic sector, and the outcome of their application. We will then use this correlation to predict whether a future Visa will be accepted or not.

# Data Understanding

## Collect Data

A dataset of American Visa applications from 2012-2016 was collected from Kaggle.com. This data has approximately 300,000 rows and 120 columns and had an original size of approximately 284mb.

## Describe Data

The primary Columns that we chose to use out of the 120 are as follows:

* **Country\_of\_citizenship**
  + This column tells us the country where the applicant is currently a citizen.
* **Economic Sector**
  + This column tells us what economic sector the applicant is applying to work in when in America.
* **Case\_status**
  + This column tells us whether an application was certified or denied.
* **Application\_type**
  + This is what type of application it was, either Mailed in or Online
* **Job Title**
  + This is the column where the applicants had their jobs entered, these varied from person to person as they were more than likely just what the applicant typed/wrote instead of chosen from a selection.
* **naics\_us\_title**
  + This is more or less the same as job title, however these are much more organised as the applicants had to choose from the already existing Naics titles.

Some of the other columns that were contained in the dataset but we chose not to use would be:

* **Unit of Pay**
  + Whether they were paid weekly, monthly etc.
* **Class of Admission**
  + kind of application it is.
* **Employer Address ( State, Street etc.)**
* **Wage**
* **Case Number**
* **Decision Date**

# Data Preparation

## Clean Data

In the original data there were two country of citizenship columns with some data in one and the rest in the other. To overcome this we decided to merge both columns into a single column.

Some columns also had no data present so we decided to delete these columns which had no data in order to make the reading of our CSV file more efficient.

## Format Data

All of the data was formatted as csv files.

# Analysis of Visas

## Data

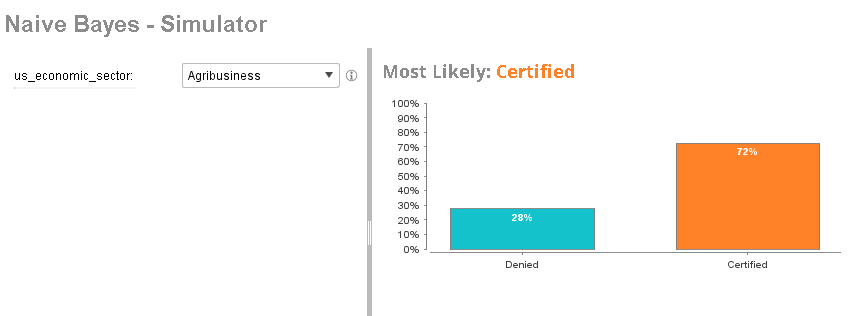
We first used the economic sector column of our data to try find correlation between this column and the outcome of the Visa application. We had some basic predictions before running a Naive Bayes on this in RapidMiner. For example, we had predicted that economic sectors such as Agribusiness and Construction would have a less successful outcome and that sectors such as IT and Energy would have a better success rate.

Our predictions proved to be somewhat correct as you will see from our RapidMiner results below.

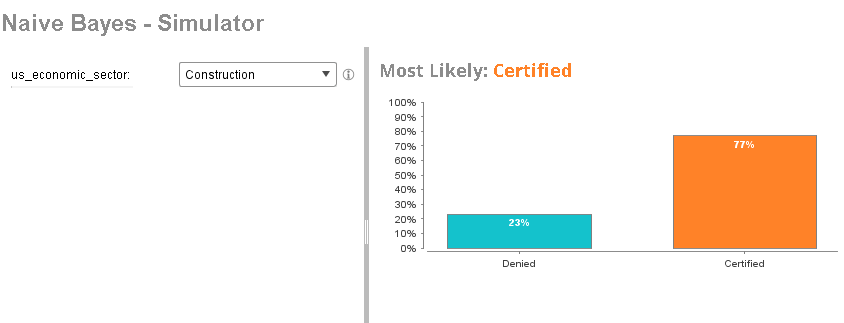
The country of citizenship column was another column we decided to use as a predictor for a Visas outcome. Our initial predictions were that third world countries would have a lower success rate and that first world, well-educated countries would have a higher success rate. Due to the vast differences in the number of people applying from each country, our results did not prove to be as accurate as we expected. For example, almost a third of all applications were from India while some countries may have only had a single application. This may have led to the data being slightly skewed.

## Naive Bayes (Economic Sector VS Case Status) Results

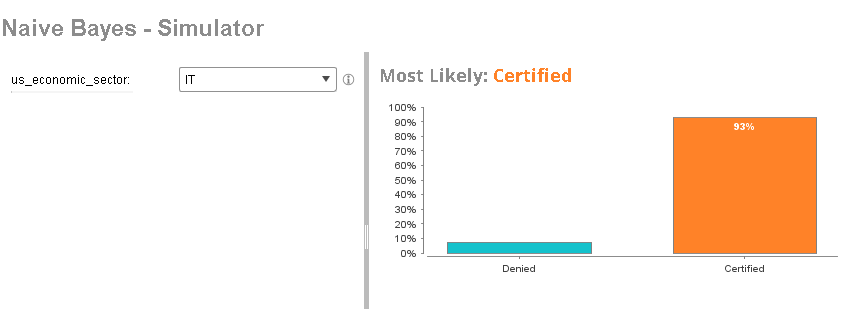
Agribusiness Results



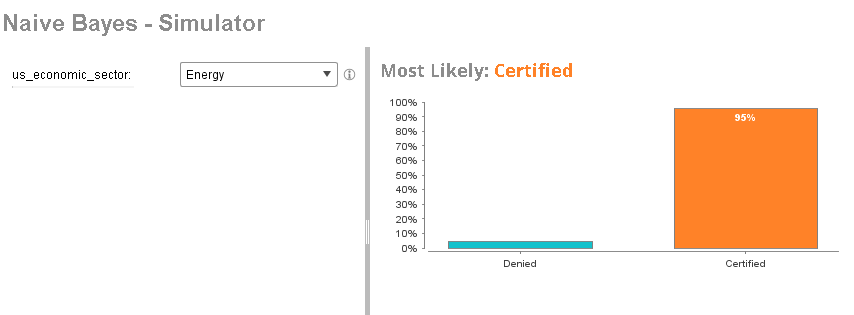
Construction Results



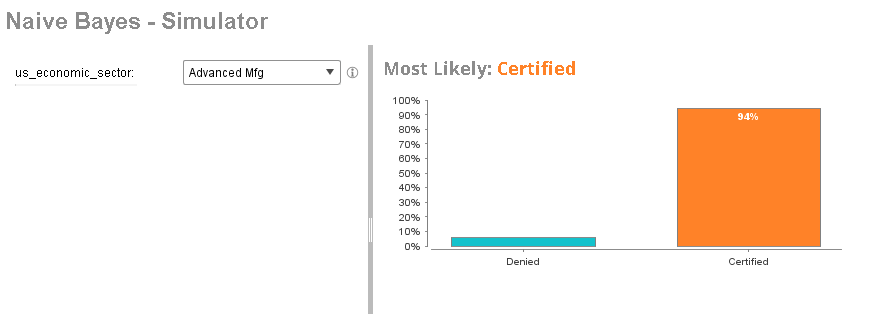
IT Results



Energy results

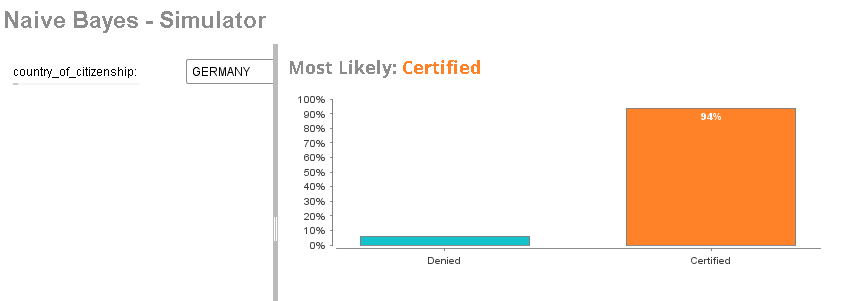


Advanced Manufacturing Results

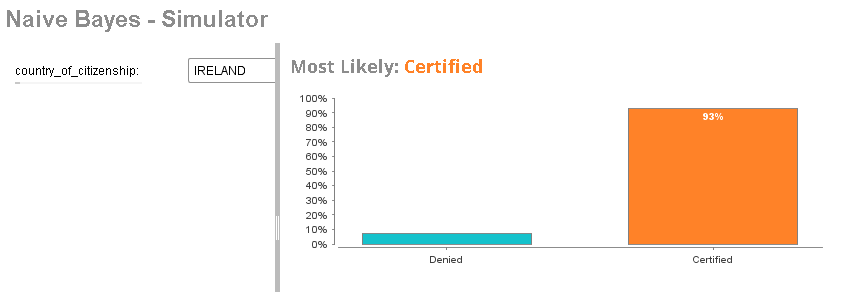


## Naive Bayes (Country of Citizenship vs Case Status) Results

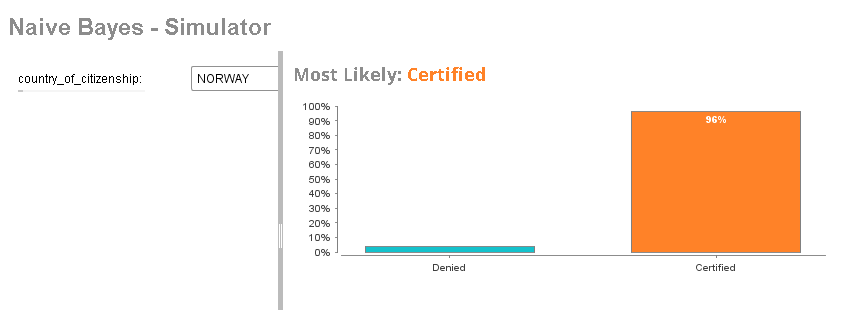
Germany results



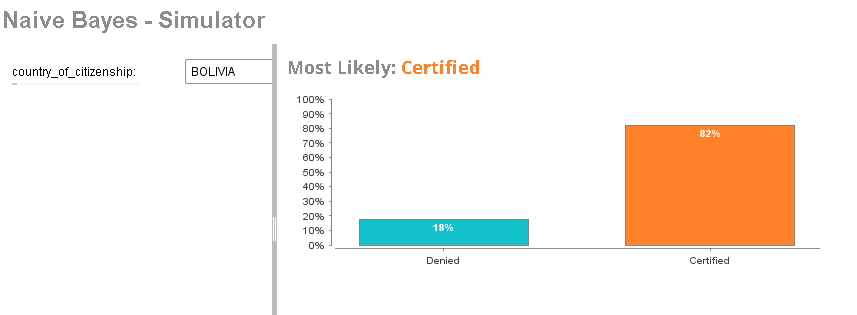
Ireland results



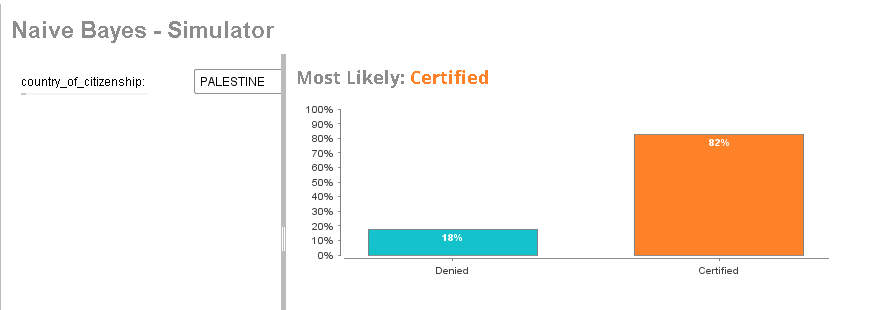
Norway results



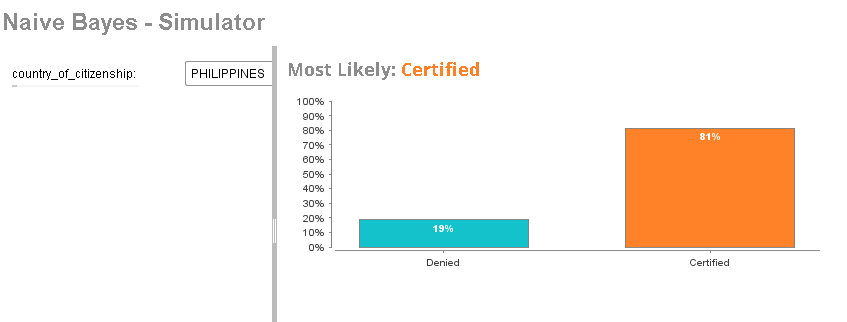
Bolivia results



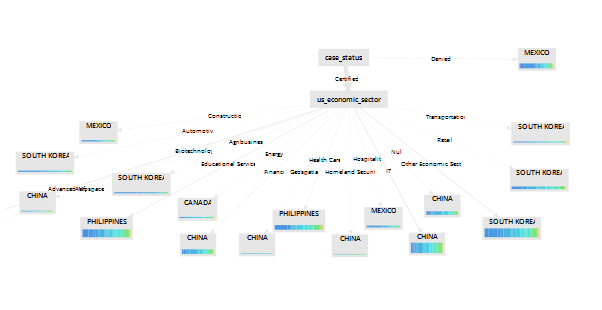
Palestine results



Phillipines results



## Decision Tree



Due to the vast amounts of applications from India, Decision Trees including those applications weren’t ideal.

As such this is a decision tree which used the economic sector, the case status and the country of citizenship from the applications not from India. It shows which Countries were dominant in each sector in terms of certified applicants, and it shows the most dominant Country of Citizenship in terms of denials

## Assessment

From these results we can see that more modern economic sectors which have more jobs available are the most likely to get an application accepted. Older, nonskilled economic sectors such as agribusiness and construction have some of the lowest success rates due to anybody being able to fill these positions. This is what we would have expected before running RapidMiner.

In relation to the country of citizenship vs case status results, we can see that first world countries which are wealthier are also the most likely countries to have a Visa accepted. Smaller, third-world countries such as Bolivia, Phillipines and Palestine as seen above are less likely to have Visa’s accepted. These results may be down to the difference in education in these countries as poorer countries are less likely to have qualified third level education citizens.

## Recommendations

Our recommendation based on these results to people who are applying for an American Visa are as follows:

1. The more skilled you are in terms of job sector, the more likely your Visa will be accepted.
2. If you work in a more modern job sector this will also increase your chances of having your Visa accepted.
3. If you work in a low skilled job there is a much greater chance that your Visa will not be accepted.
4. If you are a citizen of a first world country you are more likely to have a Visa application accepted.
5. If you are a citizen of a third world country your Visa is less likely to be accepted.