Analysis of U.S. Permanent Visa Decisions

Kiana Alikhademi, Mahdi Kouretchian, Diandra Prioleau

Introduction

Problem Statement:

- Using different set of variables in the data set, how well can a classification model predict a visa decision(case status)?
- Are clustering techniques able to group applicants based on case status (denied, certified, certified-expired, withdrawn) in a meaningful way?

Data Description:

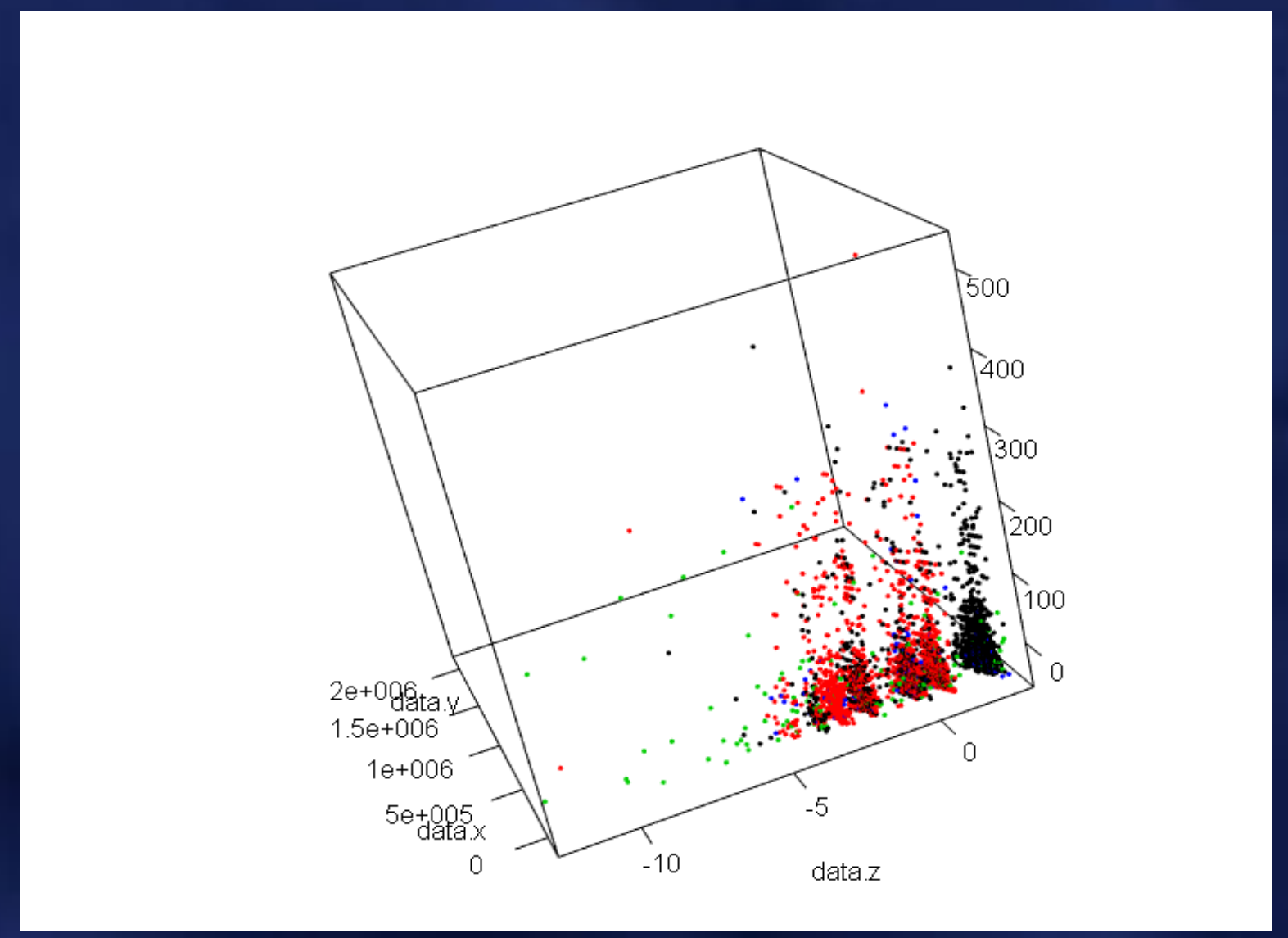
The data set includes **374,362** observations and **154** variables.
The target variable is case status. It is a categorical variable with 4 levels:

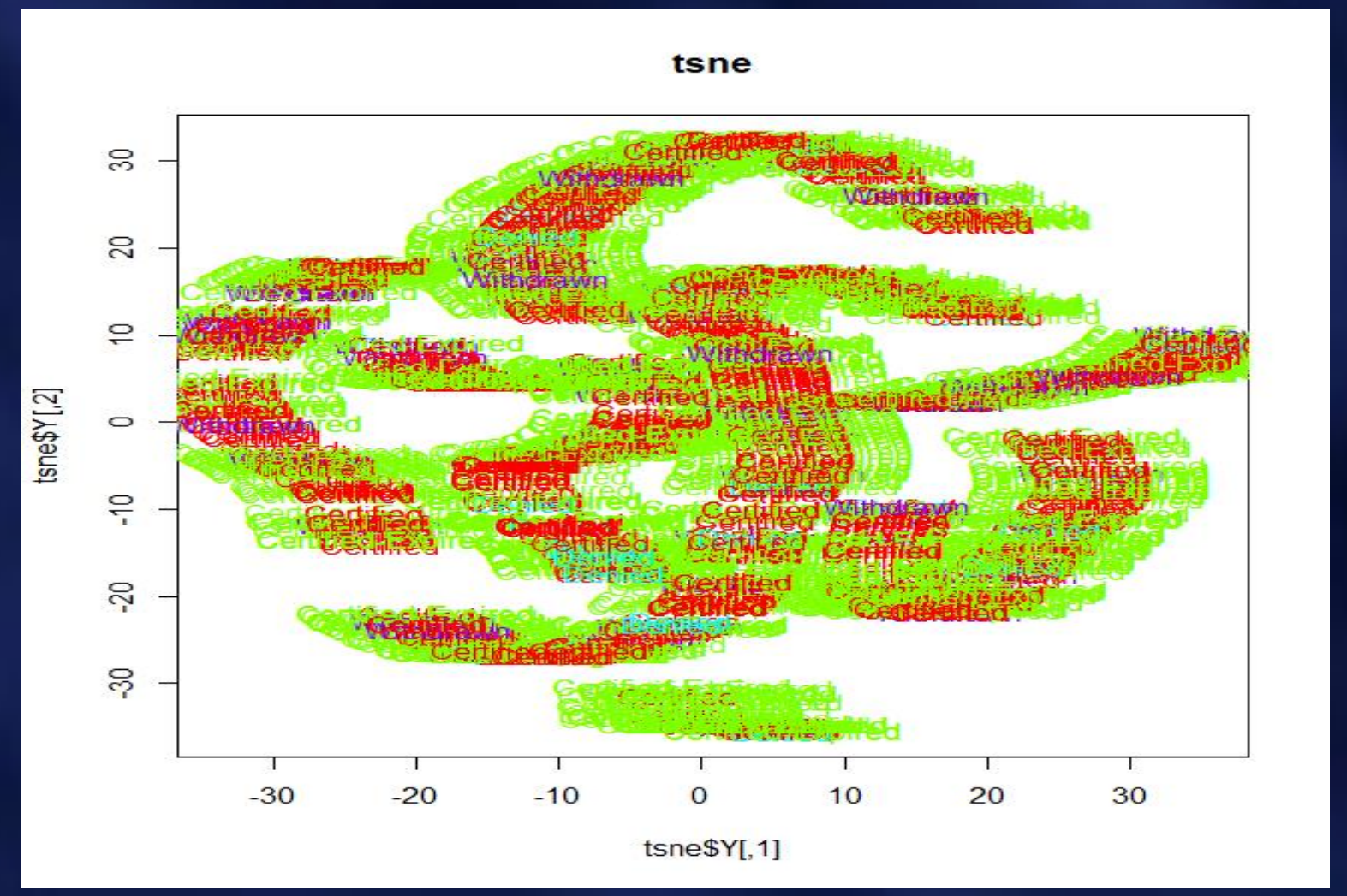
- Certified
- Certified-Expired
- Denied
- Withdrawn

The data was gathered by the U.S. Department of Labor from 2012 through 2016.

Preliminary Data Anlysis

Preliminary data analysis was conducted using **Multi Dimensional Scaling (MDS)** and **TSNE** applied on sample of data, which shows that the dataset has strong non-convex and nonlinear elements .





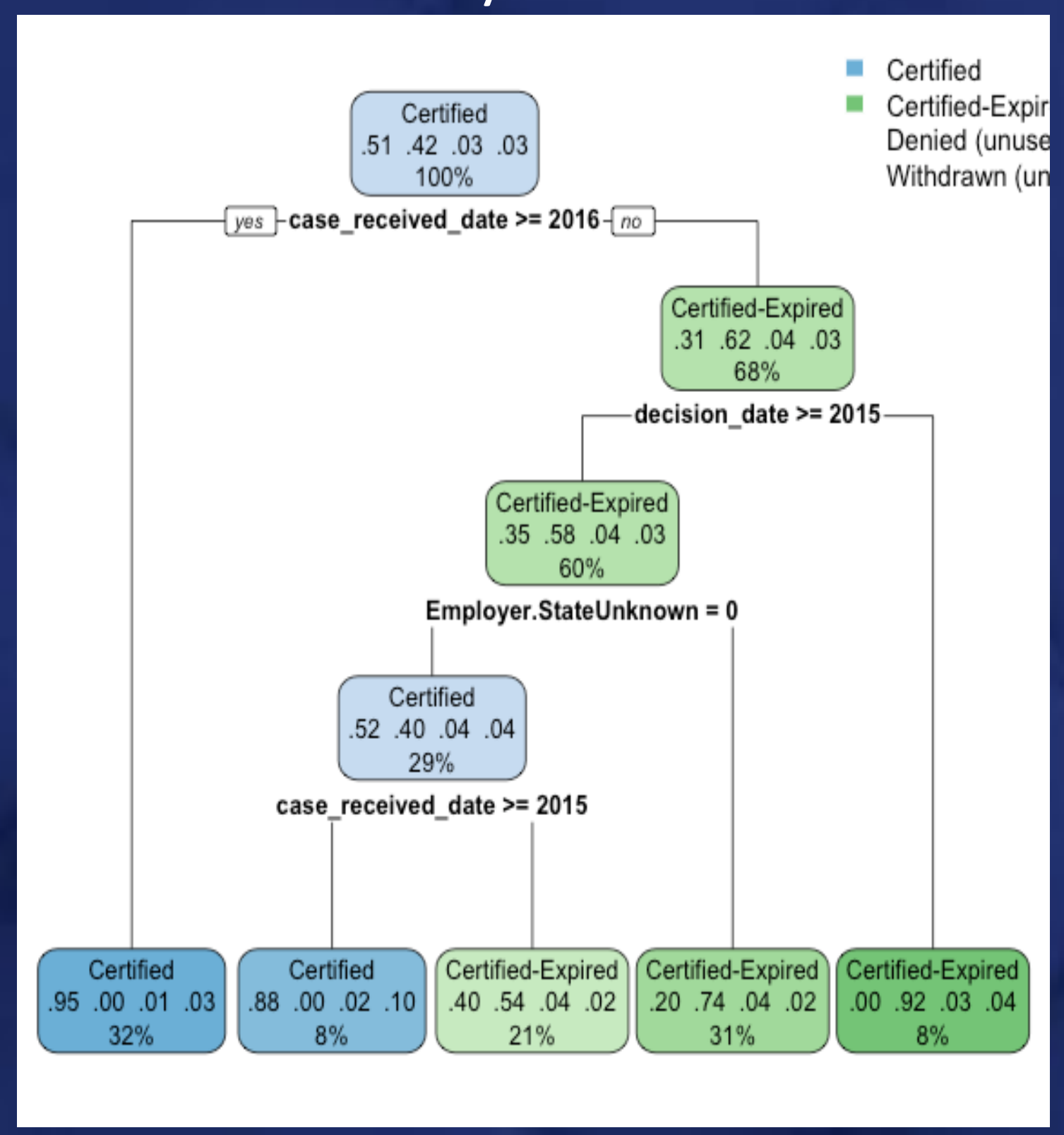
Methodology

Pre-processing*:

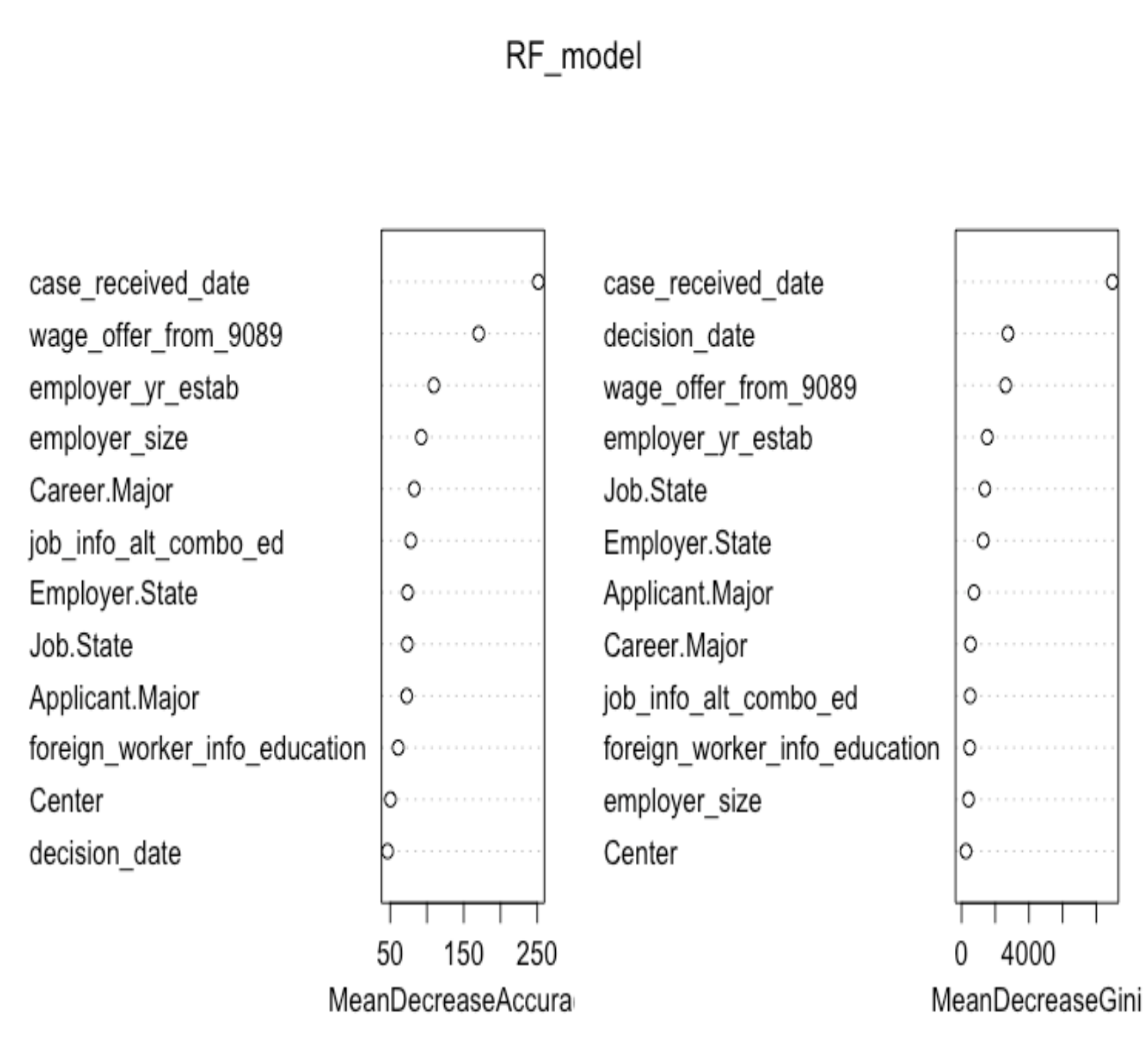
- Sub-selected predictor variables that best addressed the research question(s) to use for training and testing the classification models and for clustering
- Handle missing values for sub-selected predictor variables
- Mapping of certain variables that had too many categories within them
- Created new predictor variable that indicated which U.S. visa center applications were processed based on which state the employer was located
- Resulted in reduction of data set from **374,362** observations to **69,552** observations

Applied Methods*:

- Classification: Train data dimension is **48,686** observations by **13** variables.
  - Decision Tree
  - Random Forest (number of trees were set to 1000)
  - Naive Bayes



Certified
Certified-Expir
Denied (unuse
Withdrawn (un
Certified .51 .42 .03 .03 100%
yes
case_received_date >= 2016
no
Certified-Expired .31 .62 .04 .03 68%
decision_date >= 2015
Certified-Expired .35 .58 .04 .03 60%
Employer.StateUnknown = 0
Certified .52 .40 .04 .04 29%
case_received_date >= 2015
Certified .95 .00 .01 .03 32%
Certified .88 .00 .02 .10 8%
Certified-Expired .40 .54 .04 .02 21%
Certified-Expired .20 .74 .04 .02 31%
Certified-Expired .00 .92 .03 .04 8%



- **Clustering*:**
  - Partition Methods (PAM)
  - Hierarchical Clustering (Complete Linkage)
  - Divisive Method (Diana)

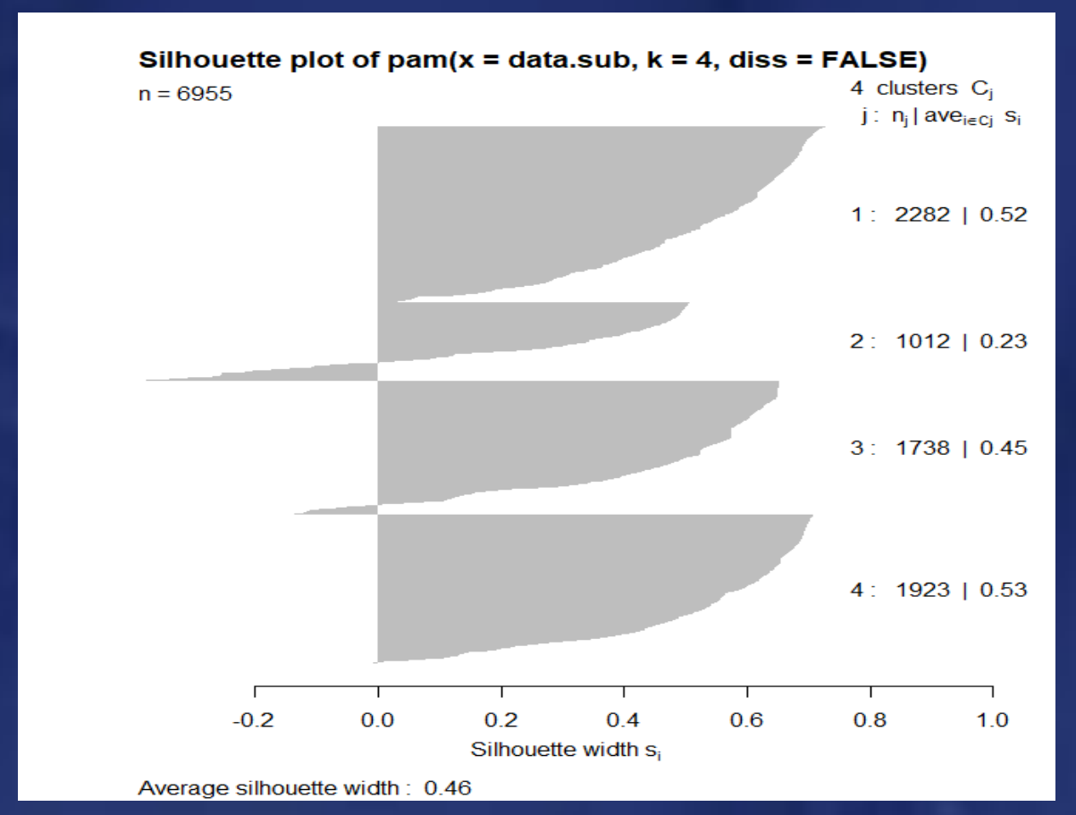
*For more detailed information, please refer to our report.

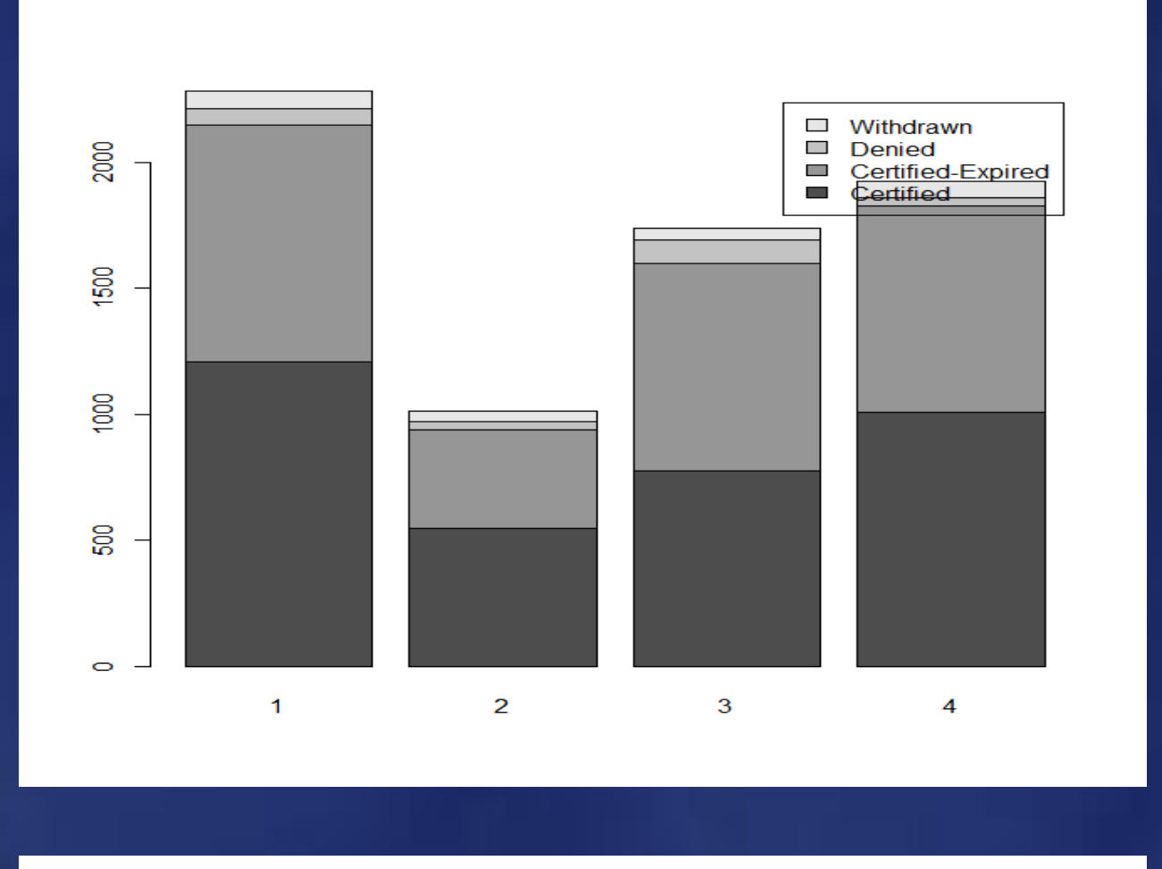
Limitations

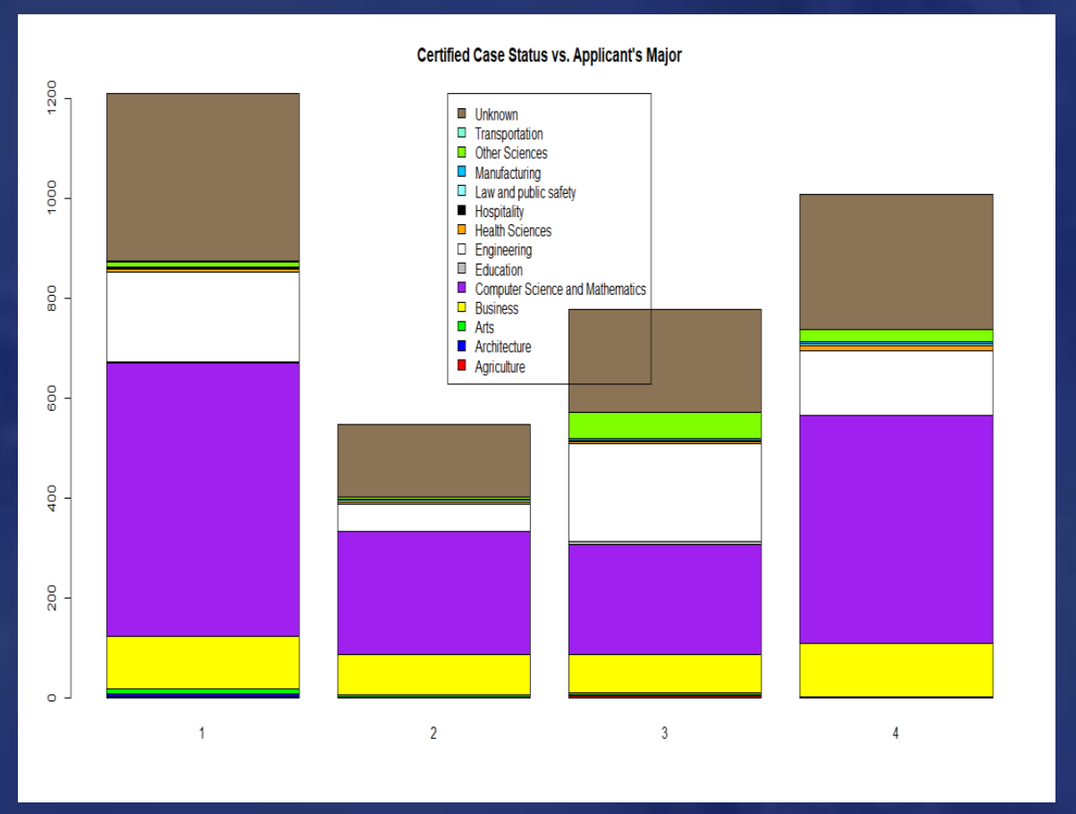
- Multiple missing values in the dataset
  - Resulted in loss of instances for training and testing
- Mapping of variables
  - Inability to acquire proper information about certain variables
- Imbalance in the predictor variable “case status”

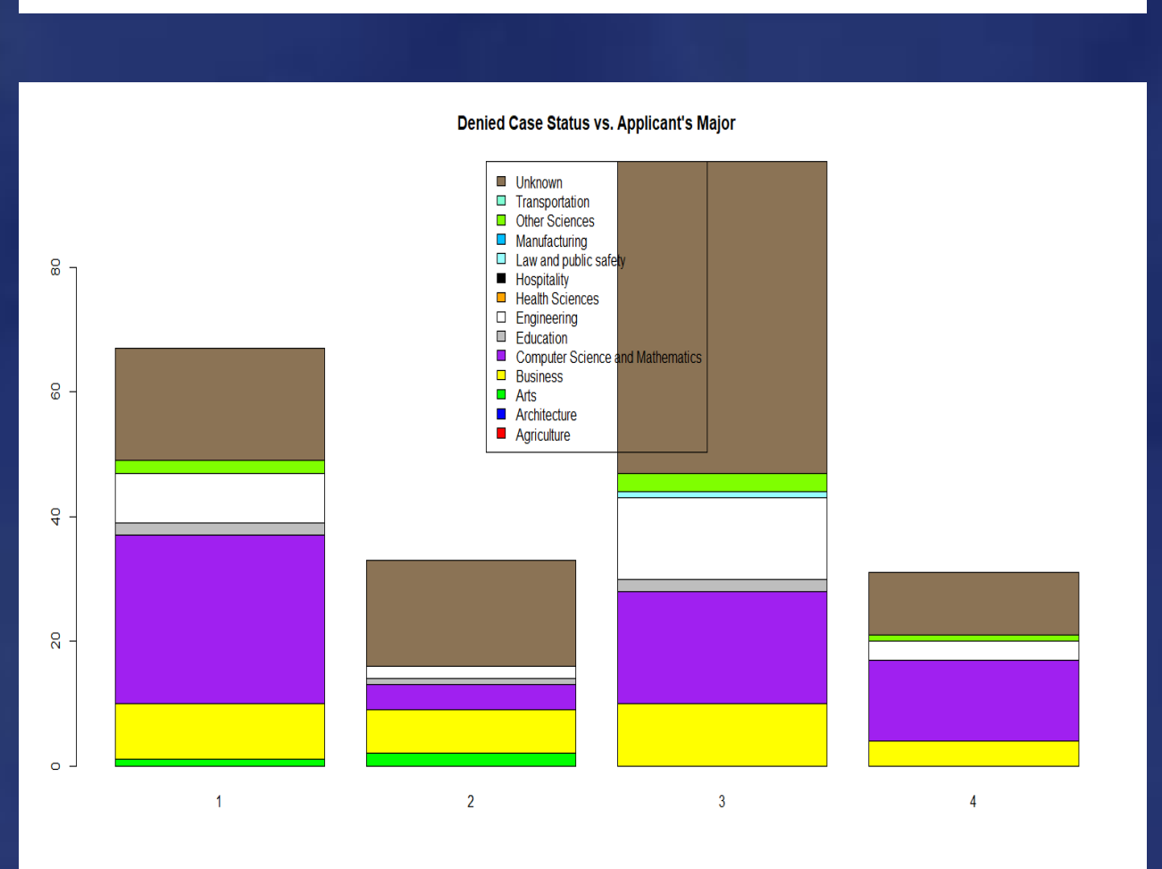
Results

- **Classification*:**
  - Decision Tree:
    - Accuracy: **79.32**
  - Naive Bayes:
    - Accuracy: **59.41**
  - Random Forest:
    - Accuracy: **82.5**
- **Clustering*:**









*For more information about the metrics please refer to our report.

Conclusion

- U.S. visa decisions can be predicted using machine learning classification models
  - Although achieved accuracy of 82%, ML techniques may not be viable option for predict visa decisions, which affect livelihood of people, since it does not achieve near perfect accuracy
  - Also we observed that variables which do not show the qualifications of foreign worker such as case received date and employer size is considered as important
- Clustering did not show extreme differences between clusters
  - Showed that individuals in computer science, mathematics, and engineering majors likely to have application certified
  - Although small, individuals in education major were more likely to be denied than certified