UNIVERSITY OF
TEXAS
ARLINGTON

PROJECT REPORT:
H1B DISCLOSURE DATASET

Analyzing the case status of an application submitted by the employer to hire non-immigrant workers under the H-1B visa program

INSY 5339 – PRINCIPLES OF BUSINESS DATA MINING

Submitted by: Group 6

- Charmi Narendrakumar Trivedi
- Pearl Firoz Mehta

Guided By:

Dr. Riyaz Sikora

Table of Contents

H1B Disclosure Dataset

1 Data Background

In the Data Mining class, we had the opportunity to analyze data by performing data mining algorithms to a dataset. Our dataset is from Office of Foreign Labor Certification (OFLC). OFLC is a division of the U.S. Department of Labor. The main duty of OFLC is to assist the Secretary of Labor to enforce part of the Immigration and Nationality Act (INA), which requires certain labor conditions exist before employers can hire foreign workers.

H-1B is a visa category in the United States of America under the INA, section 101(a)(15)(H) which allows U.S. employers to employ foreign workers. The first step employer must take to hire a foreign worker is to file the Labor Condition Application. In this project, we will analyze the data from the Labor Condition Application.

1.1 Introduction to H1B Dataset

The H-1B Dataset selected for this project contains data from employer's Labor Condition Application and the case certification determinations processed by the Office of Foreign Labor Certification (OFLC) where the date of the determination was issues on or after October 1, 2016 and on or before June 30, 2017.

The Labor Condition Application (LCA) is a document that a perspective H-1B employer files with U.S. Department of Labor Employment and Training Administration (DOLETA) when it seeks to employ non-immigrant workers at a specific job occupation in an area of intended employment for not more than three years.

1.2 Goal of the Project

Our goal for this project is to predict the case status of an application submitted by the employer to hire non-immigrant workers under the H-1B visa program. Employer can hire non-immigrant workers only after their LCA petition is approved. The approved LCA petition is then submitted as part of the Petition for a Non-immigrant Worker application for work authorizations for H-1B visa status.

We want to uncover insights that can help employers understand the process of getting their LCA approved. We will use WEKA software to run data mining algorithms to understand the relationship between attributes and the target variable.

1.3 Dataset and Attributes

The H-1B dataset from OFLC contained 40 attributes and 528,147 instances. The attributes are in the table below. The attributes highlighted in red were removed during the data cleaning process.

ATTRIBUTES	DESCRIPTION
CASE_NUMBER	Unique identifier assigned to each application submitted for processing to the Chicago National Processing Center.
CASE_SUBMITTED	Date and time the application was submitted.
DECISION_DATE	Date on which the last significant event or decision was recorded by the Chicago National Processing Center.
VISA_CLASS	Indicates the type of temporary application submitted for processing. R = H-1B; A = E-3 Australian; C = H-1B1 Chile; S = H-1B1 Singapore. Also, referred to as “Program” in prior years.
EMPLOYMENT_START_DATE	Beginning date of employment.
EMPLOYMENT_END_DATE	Ending date of employment.
EMPLOYER_NAME	Name of employer submitting labor condition application.
EMPLOYER_ADDRESS	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_CITY	
EMPLOYER_STATE	
EMPLOYER_POSTAL_CODE	
EMPLOYER_COUNTRY	
EMPLOYER_PROVINCE	
EMPLOYER_PHONE	
EMPLOYER_PHONE_EXT	
AGENT_ATTORNEY_NAME	Name of Agent or Attorney filing an H-1B application on behalf of the employer.

AGENT_ATTORNEY_CITY	City information for the Agent or Attorney filing an H-1B application on behalf of the employer.
AGENT_ATTORNEY_STATE	State information for the Agent or Attorney filing an H-1B application on behalf of the employer.
JOB_TITLE	Title of the job.
SOC_CODE	Occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System.
SOC_NAME	Occupational name associated with the SOC_CODE.
NAICS_CODE	Industry code associated with the employer requesting permanent labor condition, as classified by the North American Industrial Classification System (NAICS) .
TOTAL_WORKERS	Total number of foreign workers requested by the Employer(s).
FULL_TIME_POSITION	Y = Full Time Position; N = Part Time Position
PREVAILING_WAGE	Prevailing Wage for the job being requested for temporary labor condition.
PW_UNIT_OF_PAY	Unit of Pay. Valid values include “Daily (DAI),” “Hourly (HR),” “Bi-weekly (BI),” “Weekly (WK),” “Monthly (MTH),” and “Yearly (YR)”.
PW_SOURCE	Variables include "OES", "CBA", "DBA", "SCA" or "Other".
PW_SOURCE_YEAR	Year the Prevailing Wage Source was Issued .
PW_SOURCE_OTHER	If "Other Wage Source", provide the source of wage.
WAGE_RATE_OF_PAY_FROM	Employer’s proposed wage rate.
WAGE_RATE_OF_PAY_TO	Maximum proposed wage rate.
WAGE_UNIT_OF_PAY	Unit of pay. Valid values include “Hour", "Week", "Bi-Weekly", "Month", or "Year".
H-1B_DEPENDENT	Y = Employer is H-1B Dependent; N = Employer is not H-1B Dependent.
WILLFUL_VIOLATOR	Y = Employer has been previously found to be a Willful Violator; N = Employer has not been considered a Willful Violator
WORKSITE_CITY	City information of the foreign worker's intended area of employment.
WORKSITE_COUNTY	County information of the foreign worker's intended area of employment.
WORKSITE_STATE	State information of the foreign worker's intended area of employment.

WORKSITE_POSTAL_CODE	Zip Code information of the foreign worker's intended area of employment.
ORIGINAL_CERT_DATE	Original Certification Date for a Certified_Withdrawn application.
CASE_STATUS*	Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.

Table 1: Dataset and Attributes

1.4 Class Attribute

For the H-1B Dataset our class attribute is ‘CASE_STATUS’. There are 4 categories of Case Status. The values of Case_Status attributes are:

1. Certified
2. Certified_Withdrawn
3. Withdrawn
4. Denied

Certified means the LCA of an employer was approved. Certified Withdrawn means the case was withdrawn after it was certified by OFLC. Withdrawn means the case was withdrawn by the employer. Denied means the case was denied OFLC.

2 Data Cleaning Process

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database to improve the quality of data. Data cleaning refers to identifying incomplete, incorrect, inaccurate, and irrelevant parts of the data and then replacing, modifying, or deleting that data.

Cleaning the dataset is critical to avoiding errors when running data mining algorithms. We must analyze the raw data for missing or invalid values, typographical errors, redundant data and irrelevant attributes. After cleaning the dataset, we can run data mining algorithms and predict our class variable.

2.1 Data Cleaning Tools

We utilized Microsoft Excel to clean, balance and normalize our dataset. Our raw dataset would not open in Weka, so our first step in cleaning the dataset was to remove punctuations to avoid errors. We removed the following punctuations from our dataset: (, @ # $ () “ ‘ : ; / ? ` ~)

After removing punctuations, we further analyzed the dataset and proceed with our next steps to cleaning it.

2.2 Irrelevant Attributes

Irrelevant attributes can mislead the classifier into building an incorrect model for predicting accuracy. Attributes with unique values such as ID can build a model that would be based on that attribute with unique value and predict with maximum accuracy. It would not help us in analyzing the data set. Such an attribute is also called a False Predictor.

The attributes with unique values are Case_Number, Employer_Address, Employer_Phone and Employer_Phone_Ext. Case_Number is a unique identifier that is assigned to each case. Employer address and phone number are unnecessary attributes in predicting the Case_Status. So, we removed unnecessary attributes Case_Number, Employer_Address, Employer_Phone and Employer_Phone_Ext because they have unique values.

For Labor Condition Application, employment time cannot be more than three years. In our dataset there are two variables Employment_Start_Date and Employment_End_Date. We calculated the difference in time between those two attributes and more than 90% of the difference was 3 years for our dataset. That would make them irrelevant attributes in predicting our class variable, so they were removed.

We also had repetitive attributes such as Employer_City, Worksite_City, Worksite_County, Employer_Province, Employer_Postal_Code that were providing the same information. Those attributes were removed as well. We kept Employer_State and Worksite_State since there are 50 states and 5 US territories, we can analyze class variable based on that.

Job_Title, SOC_Code and SOC_Name were all providing the same occupational information. So, we removed Job_Title and SOC_Code and kept SOC_Name.

Our dataset had 11 irrelevant attributes that were removed.

2.3 Attributes with Missing Values

Missing values negatively affect the quality of prediction. Our dataset had many attributes with missing values. Our data comes from the Labor Certification Application. Some of the attributes did not apply to majority of the applicants and were left blank thus creating missing values. We used Microsoft Excel to determine attributes with more than 80% of the values missing and removed those attributes.

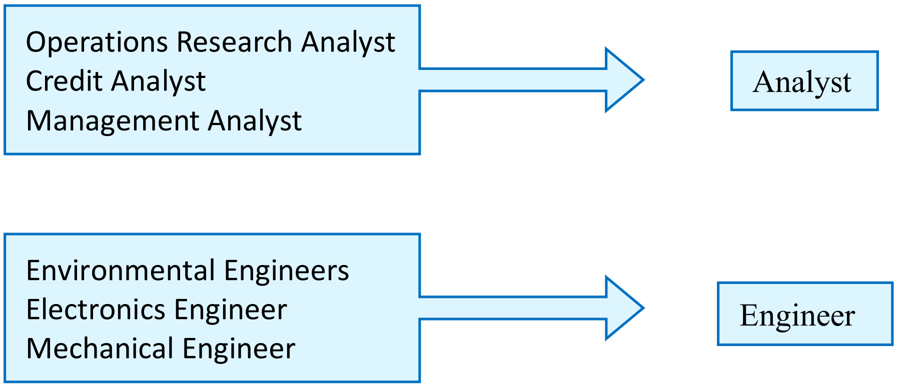
The following attributes had more than 80% of missing values:

- AGENT_ATTORNEY_NAME
- AGENT_ATTORNEY_CITY
- AGENT_ATTORNEY_STATE
- ORIGINAL_CERT_DATE

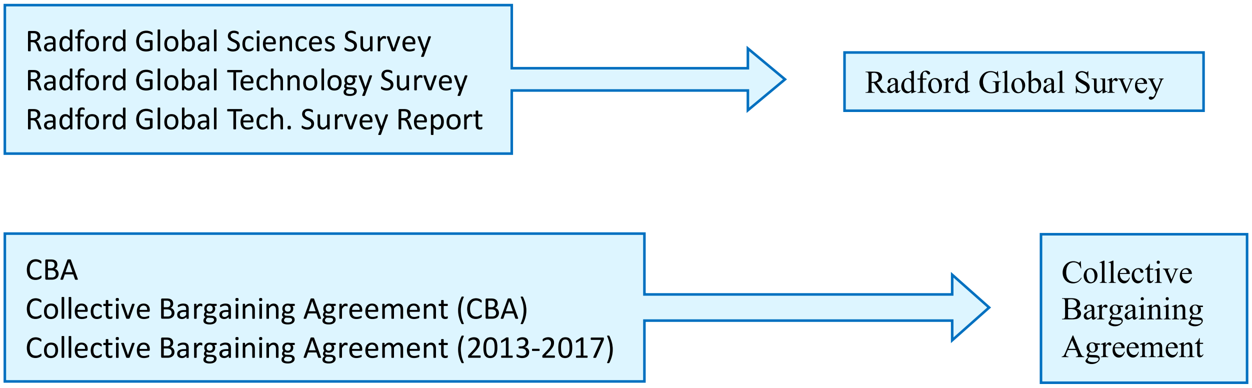
2.4 Data Integration

We had two attributes that had values which were the same but written differently increasing the number of unique values within the attribute and negatively affecting the prediction. Those two attributes are SOC_Name and PW_Source_Other. SOC_Name is the occupational name and PW_Source_Other provides the source of wage. We grouped together values in those two attributes to narrow down SOC_Name from approximately 2100 unique values to 56 and PW_Source_Other from approximately 1200 unique values to 236.

For example: Below are two examples of how we grouped together values in SOC_Name attribute. As you can see below, 3 different analyst and engineering occupations were grouped into Analyst and Engineer.



Below are two examples of how we grouped together values in PW_Source_Other. As you can see below, 3 different values that are essentially the same are grouped into one name.

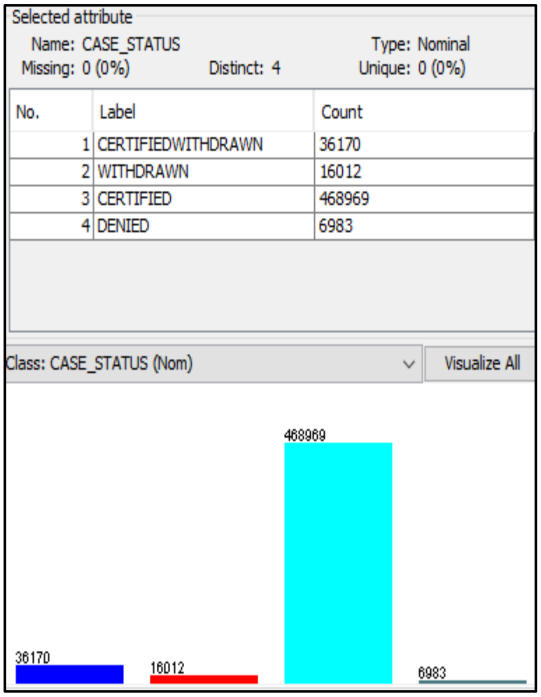

Radford Global Sciences Survey
Radford Global Technology Survey
Radford Global Tech. Survey Report
Radford Global Survey
CBA
Collective Bargaining Agreement (CBA)
Collective Bargaining Agreement (2013-2017)
Collective Bargaining Agreement

2.5 Skewed Data

Skewed dataset can provide good predictions, but those results would be biased in favor of target variable value with most records. In the real world those models would fail to provide accurate results. Our raw dataset had 40 attributes and 528,147 records. After all of the above cleaning steps, our final cleaned dataset had 27 attributes and 528,135 records. The class variable Case_Status was skewed in the final cleaned dataset. 88.79% of values were Certified, 6.85% of values were CertifiedWithdrawn, 3.03% of values were Withdrawn and 1.32% of values were Denied. Most of the values in the class variable were Certified.

We had to balance the Class_Status attribute to reduce the skewness of our dataset. Instead of using 10 different seed values in experiment design, we created 10 different datasets. 1.32% of values in the target variable were denied. That is 6983 records. Therefore, we used random sampling with replacement to create 10 different data sets with 6983 records from Certified, CertifiedWithdrawn and Withdrawn values. Each of the 10 datasets then had 27,932 records and 25% of values were Certified, CertifiedWithdrawn, Withdrawn and Denied. After that process, we had 10 balanced datasets which we used in Experimental Design to predict Case_Status.

Skewed Data



Selected attribute

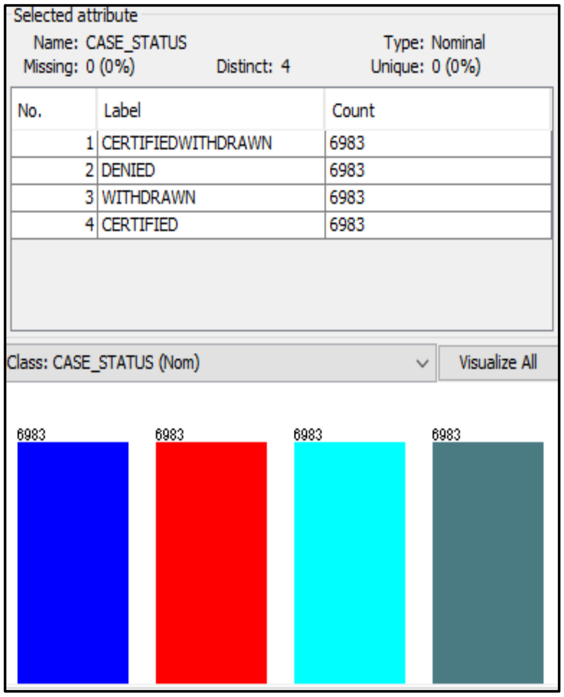
Name: CASE_STATUS　Type: Nominal

Missing: 0 (0%)　Distinct: 4　Unique: 0 (0%)

No.	Label	Count
1	CERTIFIEDWITHDRAWN	36170
2	WITHDRAWN	16012
3	CERTIFIED	468969
4	DENIED	6983

Class: CASE_STATUS (Nom)

Un-skewed Data



Selected attribute

Name: CASE_STATUS　Type: Nominal

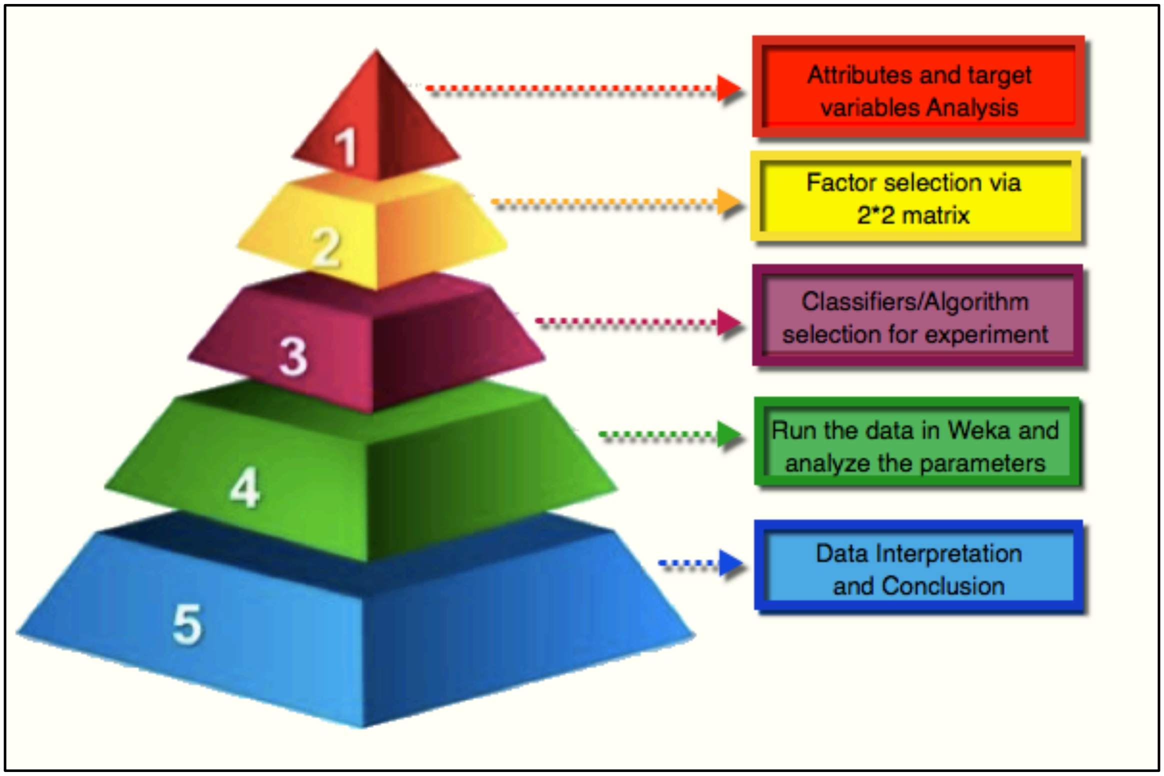
Missing: 0 (0%)　Distinct: 4　Unique: 0 (0%)

No.	Label	Count
1	CERTIFIEDWITHDRAWN	6983
2	DENIED	6983
3	WITHDRAWN	6983
4	CERTIFIED	6983

Class: CASE_STATUS (Nom)

3 Data Analysis Steps

The diagram below shows the sequence of actions we are going to follow for our project.


1
2
3
4
5
Attributes and target variables Analysis
Factor selection via 2*2 matrix
Classifiers/Algorithm selection for experiment
Run the data in Weka and analyze the parameters
Data Interpretation and Conclusion

- **Step 1: Cleaning**

  Analyze the attributes and assign weight to each irrelevant attribute. Weight is allotted based on unique identifiers, redundancy and missing values. We removed the irrelevant attributes.

- **Step 2: Experimental Design Matrix**

  Select two factors from noise, attribute selection, class imbalance, size of training set and discretization to create a 2*2 experimental design matrix.

- **Step 3: Classification Algorithm Selection**

  Set the benchmark by using ZeroR algorithm. Select three algorithms to predict the accuracy for 2*2 experiment design matrix.

- **Step 4: Run Algorithm using WEKA**

  Upload the .csv file in WEKA, convert it to .arff format and run the algorithms to analyze the parameters.

- **Step 5: Conclusion**

  Interpret the results achieved in each algorithm by analyzing the accuracy and variance for each algorithm to predict the class variable.

4 Imbalanced vs Balanced Dataset

The True Positive Rate(TPR) measures the proportion of examples that are positive and are correctly identified as positive.

Using Naïve Bayes algorithm on the original dataset, we received the TPR for each value of the target variable. The TPR for Naïve Bayes algorithm is mentioned in the table below:

Target Variable Value	True Positive Rate (TPR)
Certified Withdrawn	64.3%
Withdrawn	24.9%
Certified	95.8%
Denied	11.7%

From the above table, we can observe that the highest TPR is 95.8% for the target variable value "Certified", and the lowest TPR is 11.7% for the target variable value "Denied". The difference between the highest and the lowest TPR is 84.1%. Because of the high difference between those TPR values, the target attribute "Case_Status" has a high probability of being predicted as "Certified". Therefore, we can conclude that the dataset is imbalanced.

In order to balance the dataset, we used random sampling with replacement technique discussed in section "Skewed Data". After creating ten different datasets for the experiment design, we ran the Naïve Bayes algorithm on one of the ten balanced datasets. The TPR for Naïve Bayes algorithm on balanced dataset is mentioned in the table below:

Target Variable Value	True Positive Rate (TPR)
Certified Withdrawn	87.3%
Withdrawn	50.6%
Certified	60.3%
Denied	69.1%

From the above table, we can observe that the highest TPR is 87.3% for the target variable value “Certified Withdrawn”, and the lowest TPR is 50.6 % for the target variable value “Withdrawn”. The difference between the highest and the lowest TPR is 36.7%. Also, we observe that TPR for target variable value “Certified” has decreased from 95.8% to 60.3 % and TPR for target variable value “Denied” has increased from 11.7% to 69.1%. Thus, we can conclude that we have reasonably balanced dataset.

5 Experiment Design

We used two factors to create the 2*2 experimental design matrix. The two factors are percentage split and noise.

Factor 1 – Percentage Split (66 and 34%, 80% and 20%)

Factor 2 – Noise (0%, 10%)

	Noise (0%)	**Noise (10%)**
Percentage Split (66-34%)	C1	C2
Percentage Split (80-20%)	C3	C4

- C1: Data Set with 66% Training and 34% Test set with no noise.
- C2: Data Set with 66% Training and 34% Test set with 10% noise.
- C3: Data Set with 80% Training and 20% Test set with no noise.
- C4: Data Set with 80% Training and 20% Test set with 10% noise.

Using the above mentioned experimental design matrix, we ran each algorithm once on each of the ten balanced datasets. Therefore, the total number of runs is given by:

Total Number of Experiment Runs = Number of Classifier * Number of Criteria * 10

= 3 * 4 * 10

= 120 runs

Note: Noise was inserted only in the class variable.

6 Algorithms

After selecting Percentage Split and Noise factors for the 2*2 experimental design matrix, we selected our benchmark algorithm Zero R.

Once we had the benchmark accuracy from ZeroR algorithm, we selected three algorithms for our experimental design based on following criteria:

- Higher accuracy prediction
- Three different classifiers that will provide a model with higher stability.

The three algorithms we selected to run on the dataset are:

1. OneR
2. Naiye Bayes
3. J48

ZeroR is used as a benchmark for accuracy measure. Let us now briefly have a look at those algorithms.

6.1 ZeroR

ZeroR is a simple classification algorithm that predicts the majority class. Since it predicts the majority class, we used the accuracy provided by ZeroR as our benchmark. This means that the other algorithms we selected must provide higher accuracy rate compared to the accuracy provided by ZeroR algorithm.

6.2 OneR

OneR is simple classification algorithm creates one rule that predicts an attribute based on the lowest error rate.

We selected OneR algorithm to find the most important attribute in our dataset based on the rule generated by it. For our dataset, OneR algorithm selected attribute “Employer_Name” for each of the 10 bins.

6.3 Naïve Bayes

Naïve Bayes algorithm is based on the Bayes theorem that takes a probabilistic approach to selecting the attributes for the model. This approach produces a highly stable model with less variance. Naïve Bayes algorithm assumes that all attributes are independent of each other.

6.4 J48

J48 algorithm uses the Decision Tree Learning (DTL) process to find and optimize the most efficient attribute which increases the prediction accuracy of the model. J48 is a well-known algorithm in the data mining field because of its capability to build models with high accuracy.

7 Experimental Results

Before we ran the algorithms, we had ten balanced datasets, 2*2 experimental design matrix and four selected algorithms that are mentioned below:

- ZeroR (Benchmark)
- OneR
- Naïve Bayes
- J48

Next, we ran each of the algorithm on each file for 4 times, once for each category, C1, C2, C3 and finally C4. The table below displays the accuracy that we received in each category. We calculated the average, variance and standard deviation of the accuracy for each category.

7.1 ZeroR

In case of ZeroR, we found the benchmark by calculating the average of the averages of 4 categories. **Our benchmark is of 24.52%.** Here the highest accuracy is given by C2.

Trials	C1	C2	C3	C4
File 1	24.5972	25.05	24.1855	25.0806
File 2	24.6709	24.1023	24.1139	23.5231
File 3	24.6499	24.6604	24.6688	24.2034
File 4	24.85	24.9974	24.3287	24.2571
File 5	24.5446	24.7657	24.6867	24.7404
File 6	24.6078	24.4393	24.4719	24.3287
File 7	24.6499	24.7236	24.6151	24.4361
File 8	24.6815	24.6288	24.2213	24.0602
File 9	24.7131	24.85	24.7404	24.7225
File 10	24.4709	24.3551	24.4361	24.0064
Average	**24.6435**	**24.6572**	**24.4468**	**24.3358**
Variance	**0.007975845**	**0.241443005**	**0.03140018**	**0.57695282**
Std Dev.	**0.1015180**	**0.2926614**	**0.2275483**	**0.4405476**

7.2 OneR

In case of OneR, we received higher accuracy for C1 and C3 compared to C2 and C4. Among C1 and C3, the highest accuracy was given by C3 of 47.72%.

Trials	C1	C2	C3	C4
File 1	43.4242	40.0021	44.6473	41.7114
File 2	46.8043	43.4769	48.908	45.3455
File 3	46.4673	42.5608	47.0104	42.9825
File 4	50.258	45.9935	51.1457	46.903
File 5	46.4884	42.5082	47.1536	43.3763
File 6	49.2682	45.2459	50	45.435
File 7	47.3729	43.3084	47.8697	44.0745
File 8	49.6157	45.8039	50.9488	46.8851
File 9	42.5503	39.4967	43.8417	40.5299
File 10	43.6559	40.4128	45.7035	42.0337
Average	**46.5905**	**42.88092**	**47.72287**	**43.92769**
Variance	**7.268497813**	**5.604065666**	**6.462883467**	**4.793777417**
Std Dev.	**2.69601517**	**2.36729079**	**2.54222018**	**2.18946967**

7.3 Naïve Bayes

In case of Naïve Bayes too, we received higher accuracy for C1 and C3 compared to C2 and C4. Among C1 and C3, the highest accuracy was given by C3 of 66.97%.

Trials	C1	C2	C3	C4
File 1	66.7263	60.6507	67.7408	61.5646
File 2	66.3683	60.5349	67.2216	61.2424
File 3	66.7895	60.2927	66.5414	60.1683
File 4	66.1683	59.861	66.8636	60.2041
File 5	66.3578	60.198	66.2728	60.3831
File 6	65.7997	60.1664	66.7383	60.759
File 7	66.9685	61.0614	67.6155	61.5288
File 8	66.9264	60.6297	66.7383	60.4368
File 9	66.3789	60.1453	66.2728	60.7948
File 10	66.9264	61.0719	67.7587	62.1554
Average	**66.541**	**60.461**	**66.9763**	**60.92373**
Variance	**0.150445092**	**0.160837233**	**0.330034497**	**0.450785331**
Std Dev.	**0.38787252**	**0.4010451**	**0.5744869**	**0.6714054**

7.4 J48

In case of J48, while we received higher accuracy for C1 and C3 compared to C2 and C4 but among C1 and C3, the highest accuracy this time was given by C1 instead of 47.72%.

Trials	C1	C2	C3	C4
File 1	68.6427	59.1871	68.8865	59.7207
File 2	68.8112	60.6086	69.0118	60.759
File 3	68.7586	60.0821	68.4748	60.7411
File 4	68.4005	60.914	68.5643	60.9918
File 5	68.3479	59.5978	68.4032	60.1504
File 6	68.7691	59.5662	68.8328	59.7028
File 7	69.6852	60.8719	68.8865	60.4726
File 8	68.927	59.3766	68.618	60.6695
File 9	68.8533	60.3033	68.5643	60.3831
File 10	68.5374	60.3348	68.7612	60.5621
Average	**68.77329**	**60.08424**	**68.70034**	**60.41531**
Variance	**0.139545078**	**0.389471043**	**0.041169125**	**0.189712557**
Std Dev.	**0.37355732**	**0.6240761**	**0.2029017**	**0.4355600**

7.5 Summary of Results

With the benchmark(ZeroR): 24.52%, table below gives the summary of the results archived from the three algorithms

Algorithms	C1	C2	C3	C4
OneR	46.5905	42.88092	47.72287	43.92769
Naïve Bayes	66.541	60.461	66.9763	60.92373
J48	68.77329	60.08424	68.70034	60.41531

From this table above, we can say that out of all three algorithms, J48 gives the highest accuracy.

8 Summary of Results

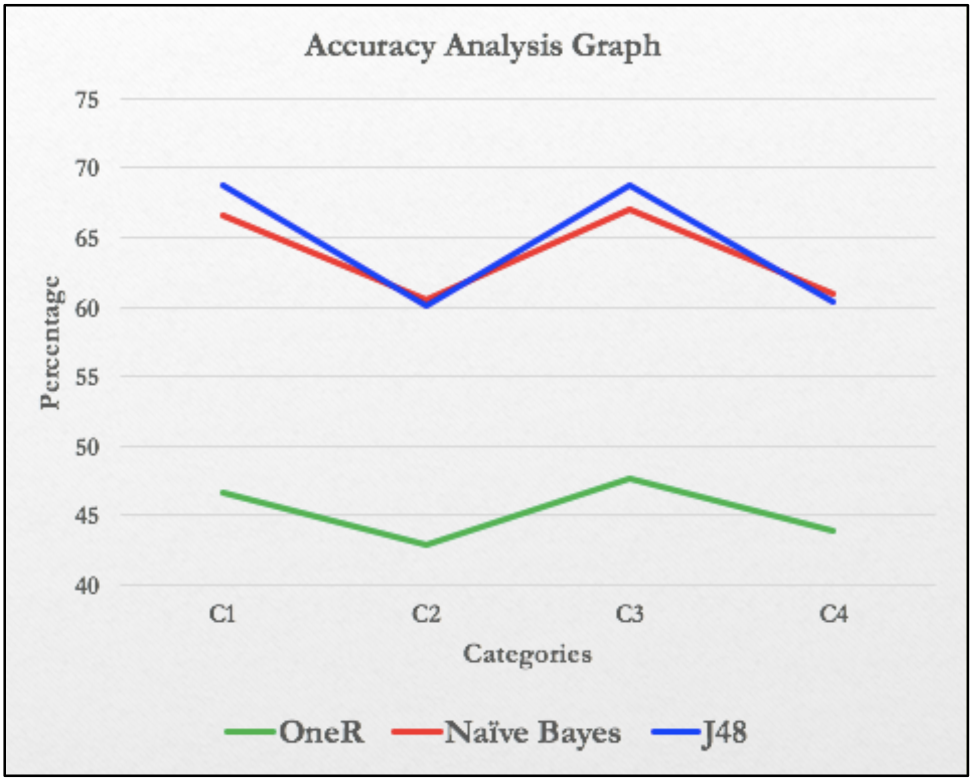
Using the four categories in 2*2 experimental design, that is C1= 66/34 split with no noise, C2= 66/34 split with 10% noise, C3= 80/20 split with no noise and C4= 80/20 split with 10% noise we calculated the accuracy and variance for all three algorithms. The accuracy analysis and variance analysis are described below.

8.1 Accuracy Analysis

Algorithms	C1	C2	C3	C4
OneR	46.5905	42.88092	47.72287	43.92769
Naïve Bayes	66.541	60.461	66.9763	60.92373
J48	68.77329	60.08424	68.70034	60.41531

8.1.1 Graphical Representation for Accuracy

The obtained results are represented in the graphical formats by the following graphs:



8.1.2 Observations for Accuracy

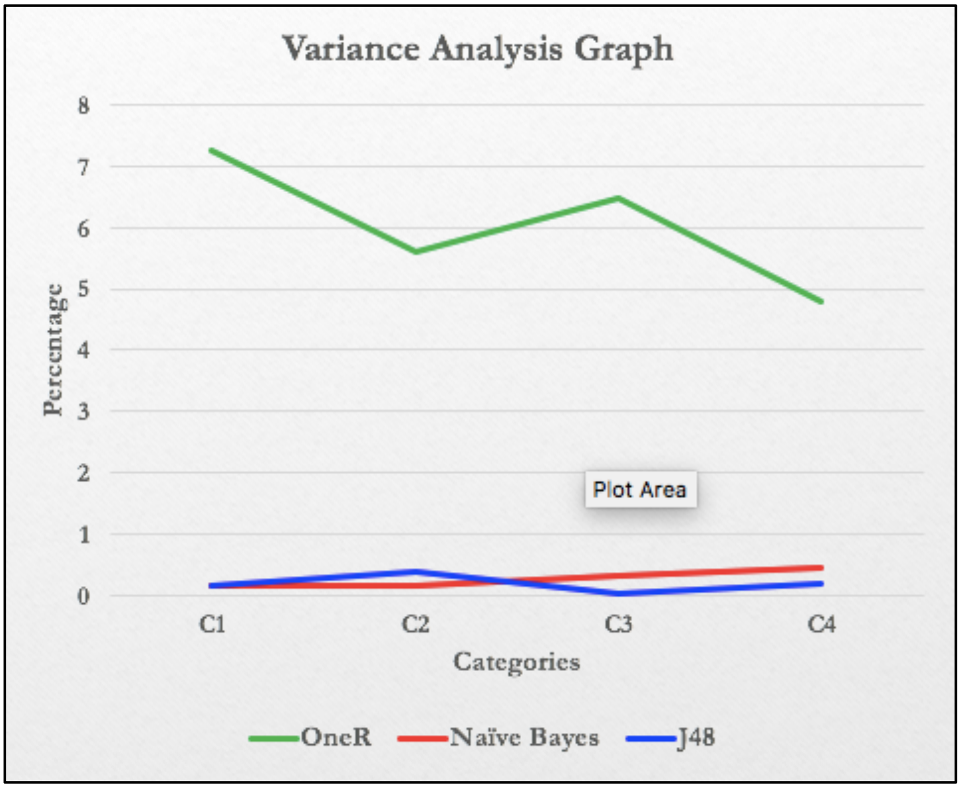
From the above accuracy graph, we observe that J48 algorithm gives us the highest accuracy in category C3, which is 80/20 percentage split and no noise. The accuracy is highest for all three algorithms at C1 and C2 when there is no noise. We can see that the accuracy decreases when we add noise to the target variable. J48 and Naïve Bayes algorithm gives us similar accuracy. The lowest accuracy is for OneR algorithm.

8.2 Variance Analysis

Algorithms	C1	C2	C3	C4
OneR	7.268497813	5.604065666	6.462883467	4.793777417
Naïve Bayes	0.150445092	0.160837233	0.330034497	0.450785331
J48	0.139545078	0.389471043	0.041169125	0.189712557

8.2.1 Graphical Representation for Variance

The obtained results are represented in the graphical formats by the following graphs:



8.2.2 Observations for Variance

From the above variance graph, we observe that OneR algorithm has the highest variance at category C1, which is 66/34 split with no noise. J48 algorithm gives us the lowest variance at category C3. The variance for J48 and Naïve Bayes algorithm are similar. Variance for Naïve Bayes algorithm increases as we move from category C1 to C4.

8.3 Attribute Analysis

Using the Attribute Evaluator, 'CfsSubsetEval' in WEKA, we found that the most important attributes in determining the case status of LCA application are PW_Source_Year, Case_Submitted_Year, SOC_Name, Employer_Name.

We evaluated the results provided by the WEKA Attribute Evaluator by comparing it to J48 algorithm, which was the best model based on our experimental design. We visualized the J48 decision tree to determine the most significant attributes. For the J48 decision tree, the root node is 'PW_Source_Year' attribute. The root node of the tree is the most important attribute. The root node is then split into three branches. Those three branches are 'Case_Submitted_Year', 'SOC_Name', 'Employer_Name' attributes.

The attribute 'PW_Source_Year' is the year the prevailing wage source was issued. Our dataset contains case certification determinations processed by the Office of Foreign Labor Certification (OFLC) where the date of the determination was issues on or after October 1, 2016 and on or before June 30, 2017. The cases in our dataset were from year 2011 through 2017. The cases which were submitted before the year 2015, were either withdrawn or denied which is why the 'PW_Source_Year' is an important attribute followed by 'Case_Submitted_Year'. Attribute 'SOC_Name' is the occupation name and attribute 'Employer_Name' is the name of the employer filing the Labor Condition Application. Based on the J48 decision tree, these are most important attributes:

1. PW_Source_Year
2. Case_Submitted_Year
3. SOC_Name
4. Employer_Name

8.4 Conclusion

Based on our accuracy and variance analysis, we observe that J48 algorithm is the best model for our dataset providing the highest accuracy and lowest variance. OneR has the lowest accuracy and highest variance among all three algorithms. According to our research, job occupation, employer name, case submitted year and the year prevailing wage source was issued are important factors in determining the case status of the Labor Condition Application filed by a perspective H1B employer to employ non-immigrant workers.

9 References

- UNITED STATES DEPARTMENT OF LABOR . (2009, January 15). OFLC Performance Data. (www.dol.gov) Retrieved September 09, 2017, from UNITED STATES DEPARTMENT OF LABOR Employment & Training Administration: https://www.foreignlaborcert.doleta.gov/performancedata.cfm
- Wikipedia. (2017, September 09). *Data cleansing - Wikipedia.* Retrieved from en.wikipedia.org: https://en.wikipedia.org/wiki/Data_cleansing