Decision support system in KNIME

H1B Visa Prediction

H1B visa

- non-immigrant visa that allows US companies to hire graduates in specialized fields (IT, finance, accounting, architecture, engineering, math, medicine, etc.)
- because of the speed of application, it's more appealing to companies looking to hire non-U.S. workers than the green card
- There are a large number of entries each year (> 2 million) and the number of places is limited
- **It became very important for employers to "know in advance" the chances for an individual candidate, i.e. whether to submit an application at all, which is why numerous research papers on this topic were written:**

O. C. M. Beliz Gunel, "Predicting the Outcome of H-1B Visa Applications.
D. A. Pandya, "Predicting filed H1-B Visa Petitions' Status"
N. N. Madhana Sohan Kumar, "A Predictive Model for H1-B Visa Petition Approval,"

Problem - a large number of rejected applications

65 000

Limit

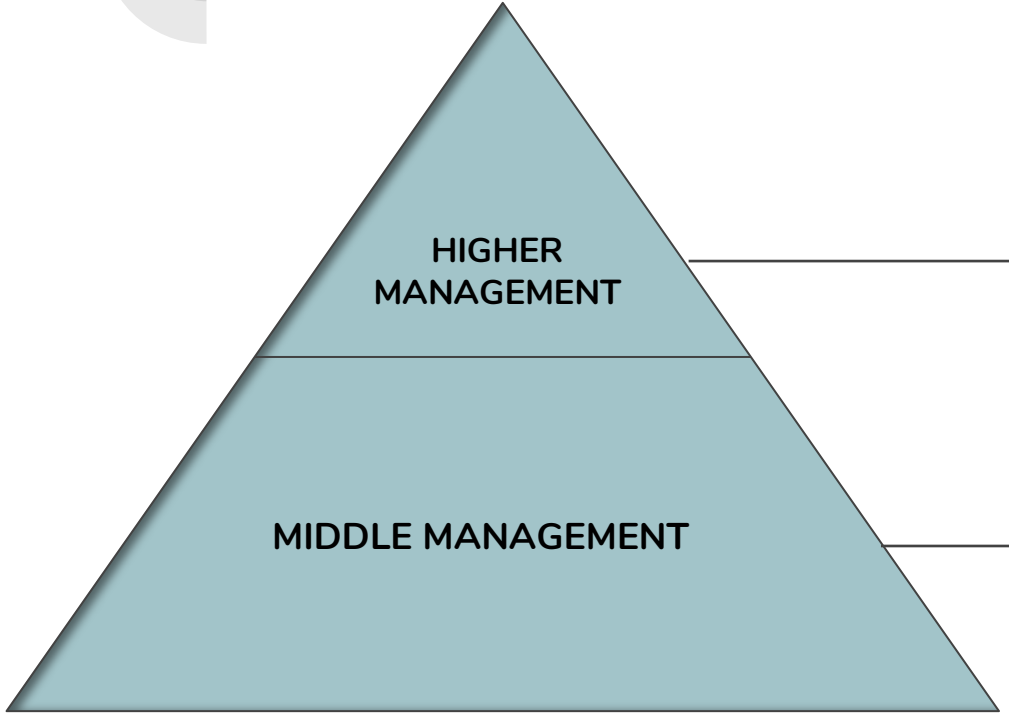
> 2 million applications → **Possibility to obtain a permanent visa**

Goal

The decision support system *H1B Approval Prediction* will help American employers in the following:

* making the decision to apply for a H1B visa on behalf of a potential foreign worker
* analyze labor markets (descriptive statistics)

Decision Support System Users



Strategic decisions (job advert for a specific position for foreign nationals)

The decision to apply for an H1B visa on behalf of the applicant

Dataset

https://public.enigma.com/browse/d582dfbd-4329-4b5e-b0c9-39149f5dd546

Dataset has the following information on individual candidates (2011-2018):

- Position
- The amount of salary
- City / State
- Part / full time
- Etc.

The target variable and the one that will try to predict it is the **case status** variable and it contains two possible values (*certified* or *denied*).

Solution Proposal - Input

Employer Info	Job Info	Wage Info	Other Info
Employer Name	SOC Code	Wage Rate From: 0	☐ Full Time Job
Employer State: Other	Job Title	Wage Rate To: 0	
Employer City	Start date*: Date: 2019-06-01	Wage Rate Unit Of Pay: Bi-Weekly	
Employer Address	End date: Date: 2019-06-01		
Employer Postal Code: 0	Worksite State: Other	PW Wage Source Year: 2000	Naics Code: 0
Total Workers: 1	Worksite City	PW Wage Source Other: Other	Serial ID: 0

*** In order to predict the visa status, it is necessary to enter the data, which is divided into 4 parts: employer info, job data, salary information and other info.

Solution Proposal - Output

Based on input details the system will predict the possible output – whether the candidate will be certified or denied for H1B visa.

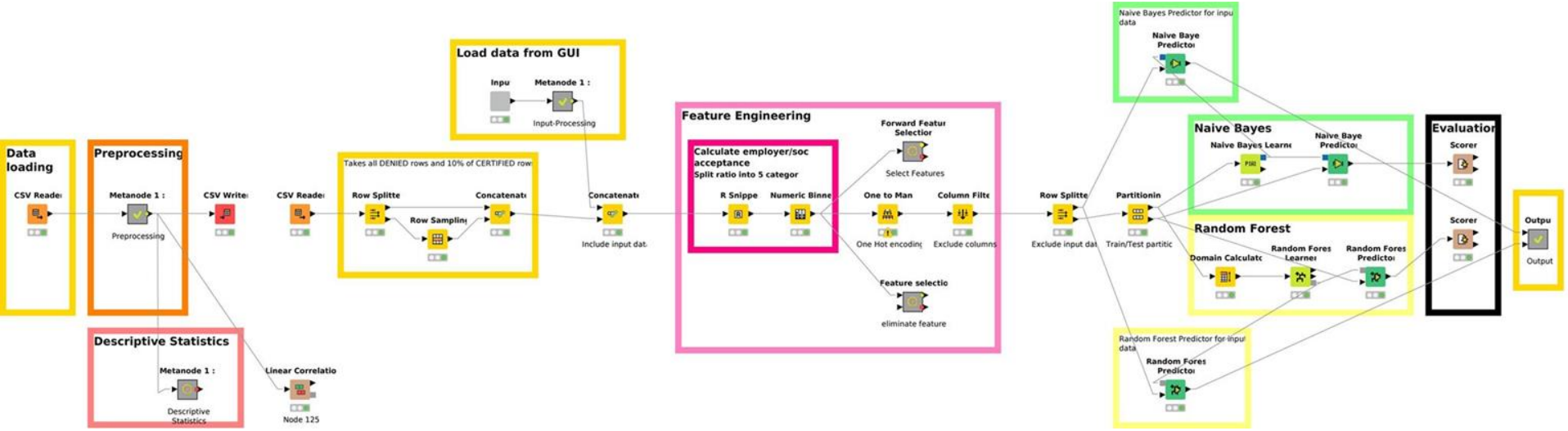
Prediction 1 - Random Forest	Prediction 2 - Naive Bayes
CERTIFIED	DENIED

The output of the first prediction method

The output of the second prediction method

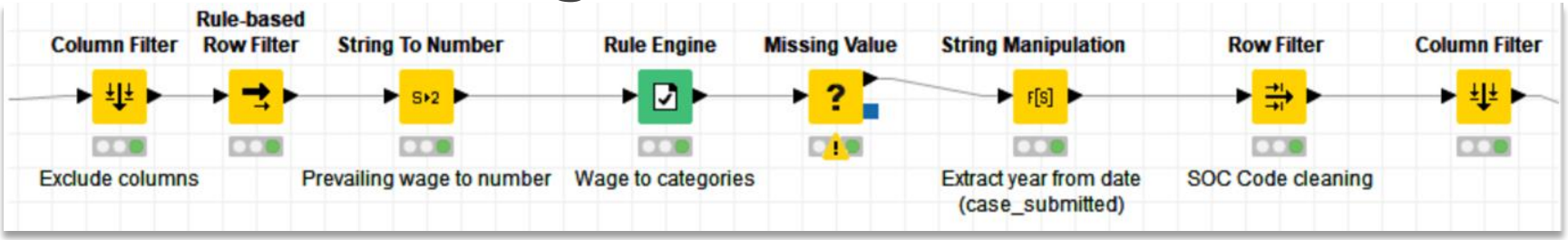
*** It has been found that the chance of visa approval increases with the **salary for a specific occupation** and **employer performance** with previous H1B applications.

System components - KNIME



The system consists of the following components: **Data set loading, preprocessing, feature engineering, partitioning, application of prediction methods, model evaluation**.

Preprocessing



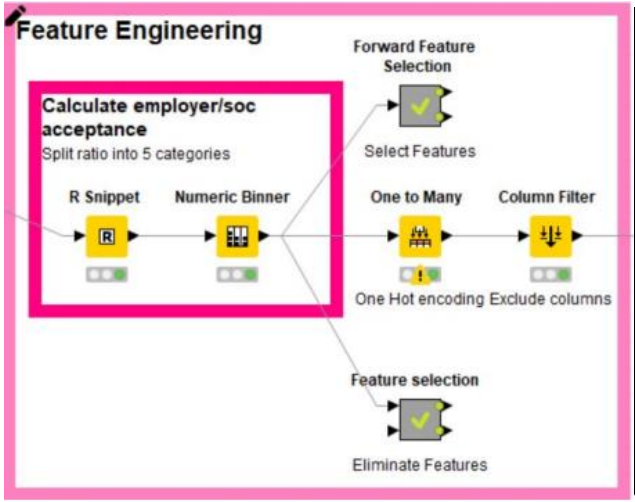
- **Column filter** node was used to manually select the columns to be considered. As the two datasets are joined, they do not have all the same columns, so this node selects the columns that are in the cross section of these two datasets.

- **Rule-based Row filter** node filters columns *pw_unit_of_pay1* and *case_status* so that the raws left satisfy the following: that the unit of payment is years (lines removed where payment was made by the hour) and case status (only include *certified or denied* raws*)*

- **String to Number** node converts the prevailing wage attribute from a string to a number in order to earn annual earnings could be classified using the **Rule Engine** node into one of five categories. Based on the rules it was added new column *pw_category* ranking salaries into one of five categories: *very low, low, medium, high and very high.* This is done because there are many unique values, and these are divided into five categories over which *One-hot coding* will be applied.

- **Missing Value** node removes all rows that have a 'null' value in one of the columns.

- **Column Filter** node removes columns like *job_title* as it is identical to the *soc_name* column, etc.



Low	High	Category
0.0	0.2	VERY LOW ACCEPTANCE
0.2	0.4	LOW ACCEPTANCE
0.4	0.6	MEDIUM ACCEPTANCE
0.6	0.8	HIGH ACCEPTANCE
0.8	1.0	VERY HIGH ACCEPTANCE

- Based on the columns *soc_name, soc_code* and *employer_name*, three new columns *soc_name_acceptance, soc_code_acceptance* and *employer_name_acceptance* were created whose values represent the **ratio of accepted requests and the total number of submitted requests**. This was done in the R programming language, using the **R Snippet** node. The given values are divided into 5 categories using the **Numeric Binner** node.

Feature selection

The following attributes were selected based on both, the *forward* and *backward feature selection* methods:

- **full_time_position,**
- **pw_category,**
- **soc_name_acceptance,**
- **soc_code_acceptance,**
- **employer_name_acceptance,**
- **case_submitted,**
- **worksite_state1.**

Not all features could and should be used within model training. Using forward and backward feature selection those attributes which mostly impact the case status (target variable) will be used for training.

Result Table - 2:20:2 - Feature Selection Loop End (Choose the variable)

File Hilite Navigation View

Table "Result table" - Rows: 10 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	I Nr. of f...	D Accuracy	S Added feature
1	1	0.886	worksite_state1
2	2	0.884	pw_category
3	3	0.882	full_time_position
4	4	0.88	soc_code_acceptance
5	5	0.876	employer_state
6	6	0.873	employer_name_acceptance
7	7	0.868	soc_name_acceptance
8	8	0.864	worksite_city1
9	9	0.851	case_submitted
All	10	0.839	employer_city

Result Table - 2:24:12 - Feature Selection Loop End (collect results)

File Hilite Navigation View

Table "Result table" - Rows: 12 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	I Nr. of f...	D Error	S Removed feature
All	12	0.162	
11	11	0.131	employer_name
10	10	0.116	case_submitted
9	9	0.112	worksite_city1
8	8	0.105	soc_code
7	7	0.103	soc_name
6	6	0.095	soc_name_acceptance
5	5	0.092	worksite_state1
4	4	0.091	soc_code_acceptance
3	3	0.092	employer_state
2	2	0.088	pw_category
1	1	0.092	full_time_position

Prediction methods

Naive Bayes

Naive Bayes was selected for the fact that it performs well in the case of large amounts of data, and for advantages such as training speed and prediction.

Random Forest

In our case, the dataset had highly unbalanced data (0.875: 0.125, certified: denied). The Random Forest algorithm is cited in the literature as a classifier that performs better than others when it comes to the problem of unbalanced datasets, and in terms of ease of implementation and performance.

Evaluation

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas...	D Accuracy	D Cohen'...
CERTIFIED	23656	1588	711	954	0.961	0.937	0.961	0.309	0.949	?	?
DENIED	711	954	23656	1588	0.309	0.427	0.309	0.961	0.359	?	?
Overall	?	?	?	?	?	?	?	?	?	0.906	0.309

Naive Bayes statistics

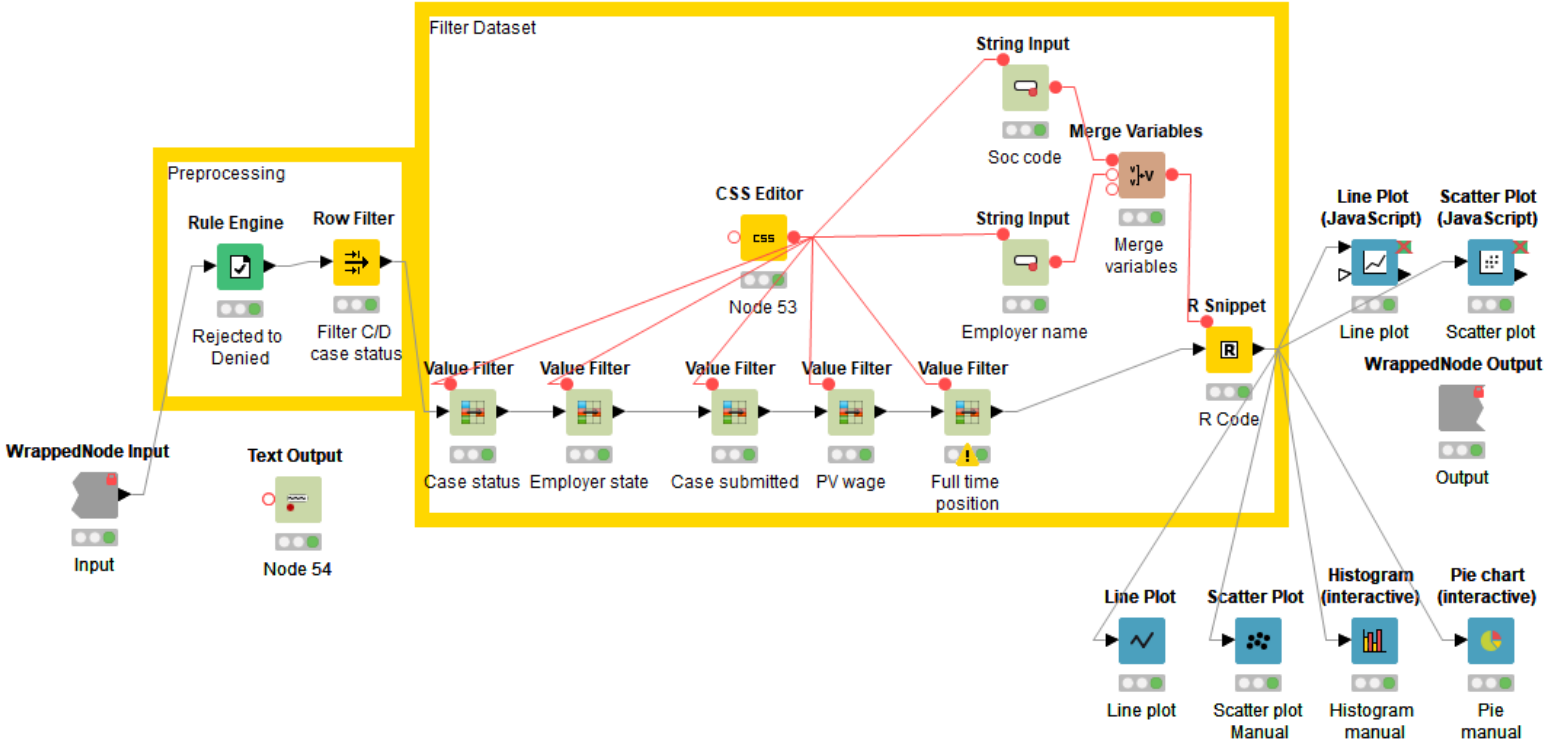
Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

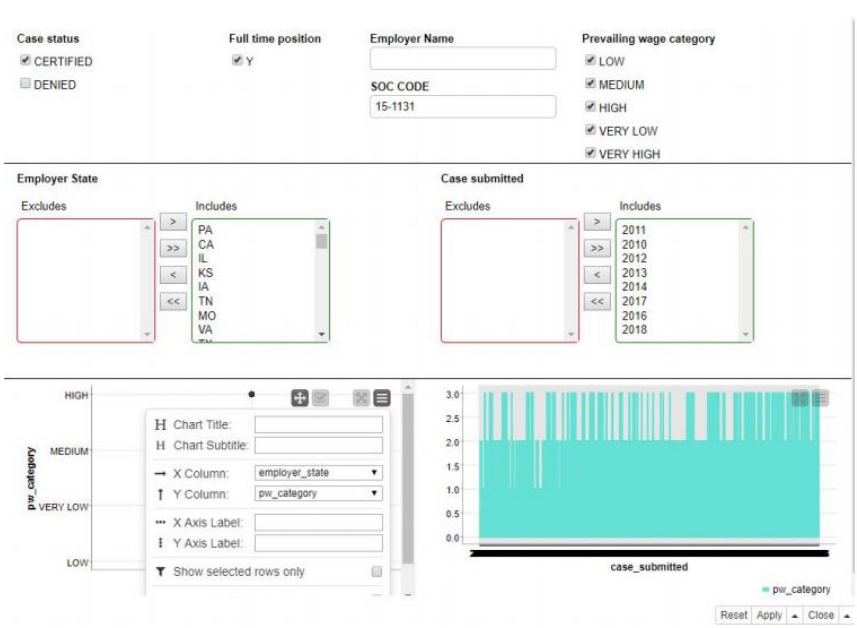
Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas...	D Accuracy	D Cohen'...
CERTIFIED	24432	1946	353	178	0.993	0.926	0.993	0.154	0.958	?	?
DENIED	353	178	24432	1946	0.154	0.665	0.154	0.993	0.249	?	?
Overall	?	?	?	?	?	?	?	?	?	0.921	0.225

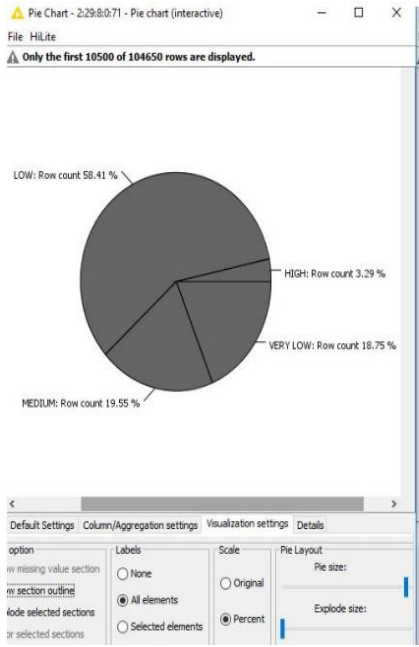
Random Forest statistics

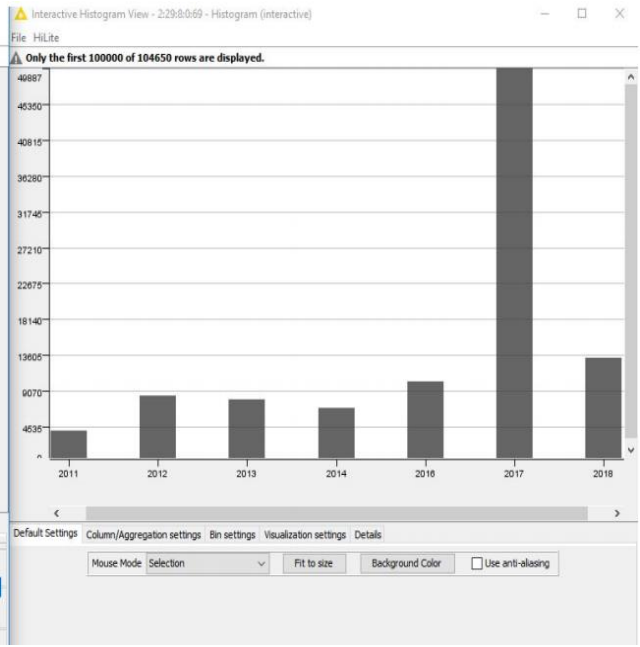
*** Scorer node in KNIME was used to evaluate model, which gives a confusion matrix at the output, and statistics such as overall accuracy, sensitivity, precision, etc.

Descriptive statistics









- The papers described were focused mainly on obtaining and comparing results using different prediction methods. This solution seeks to provide the entire environment to the user, including **user interface** for two types of users to use for **descriptive statistics (market analysis)** and visa status prediction for an individual candidate
- Higher management seeks to get the data as clear as possible over the years so that it **can make decisions regarding the recruitment of candidates** or job creation for particular positions.
- Historical decisions on granting visas to foreign nationals based on parameters such as city of employment, position or salary can be visually displayed to managers who will, based on the same, **make strategic decisions**

Result comparation – Naive Bayes

Autor	Accuracy	Recall	Precision	F1 score
Kumar & Naresh	0.84	0.63	0.91	0.74
Gunel & Mutlu	0.72	0.97	0.73	-
Naše istraživanje	0.906	0.961; 0.309	0.937; 0.427	0.949;0.359

- In terms of the data used, this solution used datasets from 2011-2018 and thus included the largest number of H1B visa data, given that other papers did not include 2018 or used a smaller period of time in their research.
- The difficult was detection of denied status based on the data entered for each candidate. The solution to this problem was partially achieved by using the *Random Forest* method, which itself, like other qualifiers, behaves badly in case of unbalanced datasets. Its variations like *Balanced Random Forest* and *Weighted Random Forest* are better behaved, but do not exist as node in KNIME tools