Data Classification of Data of Diplomas in USA

Tanmai Aurangabadkar, California State University Los Angeles

ABSTRACT **- This paper will provide you with deeper knowledge and understanding that how each variable or sets of variables affect the graduation rate of complete nation and states within it. For evaluation each variables effect on the graduation rate we used some of the "Data Classification Models", which help us evaluate accuracy for predicting the graduation rate. Each of the data classification models have their different methods of prediction and accuracy to the data, the model which fits best to our data is selected and is used. Also, in this research we have looked at the accuracy rate of selected classification model for each state with different variable sets which gives us the understanding that which are the variables in respective states affecting graduation of states. Above things discussed is briefly explained in the following paper.**

Index terms: **Data Classification Model, Graduation rate**

Introduction:

Business Understanding:

In 2001, the US nation high-school graduation rate was 71.7 %, and then over the decade due to development in all sector of industries this rate was increased to 81.4% in 2015. This was outcome of the 1.8 M more student enrolments in high school in 2013 over 2004. Today, GradNation Campaign want this graduation rate to be raised to 90% in 2020 and for this goal they need 300,000 more students in high schools in 2020 than 2013.

As US is the immigrant nation, there are people with different races and languages resident over here. Also, the geographic conditions and locations in this vast nation play a very diverse role in defining the people behavior, thinking and ideas. There are many states in US nation which provide the best facilities and high school education to the people in that specific region, but there are specific regions to which are still facing the problems to compete with another developed state. Due to such diverse environment there are many issues that need to be focused for increasing the graduation rate of US nation. Some issues are as follows:

- Poverty rate of some states.
- Weather Conditions
- Language Barriers
- Pregnancy rate at high school age
- Rural areas and inadequate education system in those area
- Tribal population in some states.
- Uneducated Black people population.

These above issues are not valid to all the states of US nation but there are many states which have all the above issues or a single issue which are affecting their state graduation rate. These issues can be solved by statically analyzing them with the help of data mining, such as;

- Population of specific language or no English-speaking people in state if the affecting the rate then providing facilities for them.
- Looking after the poverty rate one can find the reason in that state and should provide the below poverty level people with education facilities.
- Rate of graduation in black people population will suggest you the uneducated population of them, of course this rate is low because this population get bored from education in their early ages; this view will allow us to get them to the high school by solving their problem and motivating them.
- There are much women population which are not able to complete their high school due to pregnancy at high school ages; this type of region should be figured out and provide the feasible options for such mothers to complete their graduation

Discussion:

Data Preparation:

For the classification of data, we need to decide what value of the individual variable we required so get best prediction for the graduation rate. So, for this we first need to look after the variables which will affect the Graduation rate, following figure show us the variable which are most effective for deciding the Graduation rate. In the above figure we can see that the Graduation Rate (Arate) is most affected by "Cohort, Black cohort, Rural population, Hispanic cohort, Below poverty population, Limited English speaking cohort and No-English speaking population by % years of age". This is evaluated by looking at the angle between the Graduation Rate and variables. As one can see the graduation rate line makes approximately 180-degree angle with "Below poverty population" variable line, so this variable is the most effective variable in getting Graduation rate and same with the other variables. Also, the variables which make angle with rate less than 90 degree are least effective, so variables such as "White Population, Asian Population and High School Graduate Population" are least effective to the Graduation rate. Following table shows the definition of each variables

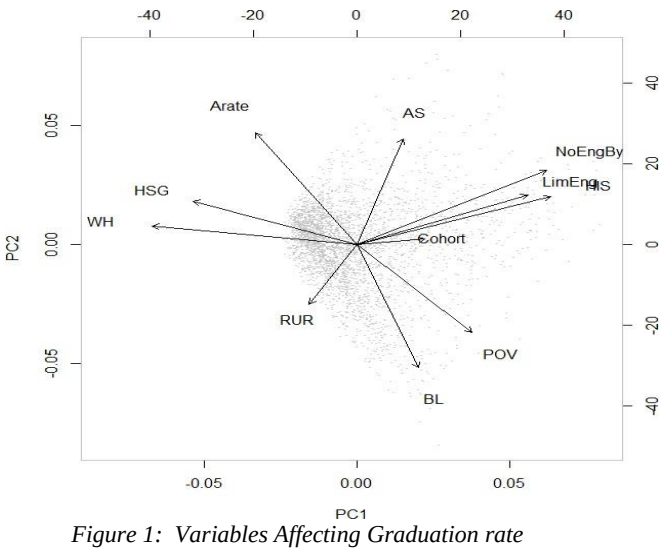

Figure 1: Variables Affecting Graduation rate

Variables	Definition
Arate	Graduation Rate
WH	White Population Rate
AS	Asian Population rate
Cohort	Total Education Cohort
HIS	Hispanic Education Cohort rate
HSG	Total High School Graduate rate
NoEngBy5	Population under 5 years of age speaking no English
BL	Black Education Cohort rate
LimEng	Education Cohort rate of student speaking limited English
POV	Poverty population rate

Table 1: Variables Definition

Class variable required for modeling is the one which we need to predict, so here the Graduation rate is the variable and the other variables are the one which will be classifies according to the class variable.
As per the project need this Graduation rate variable is discretized to the value of the less than 90(<90) and greater than and equal to 90 (>=90). This will split all the other variables which will give us the Schools which have rate less than 90 and greater than 90. Once we get those it will be easy for us to look after the Schools which have graduation rate less than 90.
Class Variable: Arate Discretized: "<90" and ">=90"
As now the class variable is decided, will look after the different set of variables which will be helping to evaluate the best classification models and accuracy to predict the graduation rate. For this the sets of variables are defined based on their intensity to affect graduation rate. The set of variables are as follows:

Set of variables(Names)	Important Predictor to relative importance
1	POV
2	BL
3	POV+BL
4	POV+BL+RUR
5	POV+BL+Cohort+RUR
6	POV+BL+RUR+Cohort+HIS
7	POV+BL+RUR+Cohort+HIS+LimEng
8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5

Table 2: Set of Variables Specified for models

Modeling:

Data Classification Models
The classification models are the regression models which are used to predict the future value of any given variable by considering other variables affecting the given variable and provide the accuracy value which allow us to have faith on the classification models. Following are some of the classification models which are used for Graduation Data.

Decision Tree:
Decision trees are powerful and popular tools for classification and prediction. They are attractive since in contrast to other machine learning techniques such as neural networks, they represent rules that human beings can understand. Decision tree is a classifier in the form of a tree structure where each node is either a leaf node, indicating the value of the target attribute or class of the examples, or a decision node, specifying some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance. These decisions generate rules for classification of a dataset using the statistical criterion: entropy, information gain, Gini index, chi-square test, measurement error, classification rate, etc. There are two stages, tree construction and post pruning, and five tree algorithms are in common use, viz., RPART, CART, CHAID, ID3, C4.5 and C5.0. Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down, greedy search through the space of possible decision trees. The algorithm used for building the models for this thesis is CART i.e., Classification and Regression Tree. In this algorithm, the condition of split is Information Gain and involves the measurement of how much information one can win by choosing a certain variable when deciding upon the variable based on which to split the tree. The dependent variable has been converted into a binary variable and the independent variables have been converted into categorical variables and a binary split is done.
A maximal classification tree gives 100% accuracy on training data, but it is a result of over fitting and would give poor prediction on test data. Tree complexity is a function of the number of leaves, the number of splits and the depth of the tree. A well-fitted tree has low bias and low variance.

Support Vector Machine:
In machine learning, support vector machines are supervised learning models with associated learning algorithms that assess data and recognize patterns, used for classification and regression analysis. Given a set of training cases, each marked for belonging to one of two categories, an SVM training algorithm frames a model that allow new cases into one group or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the cases as points in space, mapped so that the cases of these separate groups are divided by a clear gap that is as wide as possible. New cases are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In extension to executing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, definite mapping their inputs into high-dimensional feature spaces.

Rule-based Classifier:
Rule classification is the way of taking a segmented image and grouping same pixel clusters into groups called classes. A class contains one or more protocols that you can build based on your knowledge of certain features. Each rule contains one or more aspects such as area, length, or texture, which you constrict to a specific range of values. Each class will contain one or more rules (which contain an amalgamation of many aspects) that best define the class.

Artificial Neural Network:
An important application of neural network is pattern recognition which can be implemented using a feed-forward

neural network that has been trained accordingly. During training the network is trained to associate outputs with input patterns. When the network is used, it identifies the input pattern and tries to output the associated output pattern. Neural networks are capable of modeling extremely complex, typically non-linear functions. Each neuron has a certain number of inputs, each of which has a weight assigned to it. The weight is an indication of the importance of the incoming signal for that input. These weighted inputs are added together and if they exceed a pre-set threshold value, the neuron fires. The input value received from a neuron is calculated by summing the weighted input values from its input links. By adjusting the weights on the connections between layers, the perceptron output can be "trained" to match a desired output. Weights are determined by adding an error correction value to the old weight. The amount of correction is determined by multiplying the difference between the actual output (x[j]) and target (t[j]) values by a learning rate constant C. If the input node output (a[j]) is a 1, that connection weight is adjusted, and if it sends 0, it has no bearing on the output and subsequently, there is no need for adjustment. The process can be represented as: Wij(new) = Wij(old) + C (tj – xj)ai , where C = learning rate. The training procedure is repeated until the network performance no longer improves.

Random Forest:

Random forests are an altogether learning method for classification, regression and other tasks, that serve by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees mode of overfitting to their training set.

Classification Model Selection:

As in the above section 3.1 we get two know how the classification models works with the given data. Now we will implement some of the classification models for your data and they are

- Regression partition Decision tree(DTree)
- Support Vector Machine(SVM)
- Rule-based Classifier(PART)
- C 4.5 Decision tree(C4.5)
- Artificial Neural Network(NeuralNet)
- Random Forest(RanFor)

As the class variable used is the Graduation rate, we look after all the above models by using two methods "Bootstrap-validation and Cross-validation".

Bootstrap-validation:

As the result we get different accuracy for each model and when we compare those we got the following figure which tell us which classification model is best for each method. For bootstrap-validation we get the following figure which shows us that the SVM, Dtree and C4.5 classification models are the one who give highest accuracy for our data of complete US nation and will be best suitable for prediction.

```
Accuracy
           Dtree      C45       PART      SVM       NeuralNet RanFor
Dtree                0.020024  0.011596 -0.037524  0.014230 -0.038731
C45        0.0004073           -0.008428 -0.057548 -0.005794 -0.058756
PART       0.3040139 0.4543597           -0.049120  0.002634 -0.050328
SVM        9.363e-13 1.852e-13 8.472e-10            0.051754 -0.001208
NeuralNet 1.0000000 1.0000000 1.0000000 0.0049924            -0.052961
RanFor     4.105e-09 1.234e-11 2.983e-11 1.0000000 0.0066411

Kappa
           Dtree      C45       PART      SVM       NeuralNet RanFor
Dtree                0.021884  0.008815 -0.101597  0.040009 -0.089146
C45        0.462260            -0.013069 -0.123481  0.018125 -0.111030
PART       1.000000  1.000000            -0.110412  0.031194 -0.097961
SVM        3.367e-12 1.064e-14 2.813e-09             0.141606  0.012451
NeuralNet 1.000000  1.000000  1.000000  0.006104             -0.129155
RanFor     4.154e-09 1.696e-10 9.548e-09 1.000000  0.018378
```

Figure 2: Accuracy Comparison for Bootstrap-Validation Classification models

After using the bootstrap-validation methods for all the models selected above we get the accuracy from each for different sets of variables.

	%ACCURACY		
Sets of Variables	Dtree BST	SVM BST	C4.5 BST
1	64.34	61.78	64.26
2	60.79	60.07	62.73
3	67.69	69.25	68.06
4	67.64	69.13	68.35
5	68.04	70.23	67.07
6	67.21	70.35	67.55
7	69.45	70.31	67.96
8	69.48	72.12	66.86

Table 3: %Accuracy for Bootstrap- DTree-SVM-C4.5 models

For selecting the best classification model for bootstrap-validation we look at the following graph which tells us that the accuracy of SVM model is the highest.

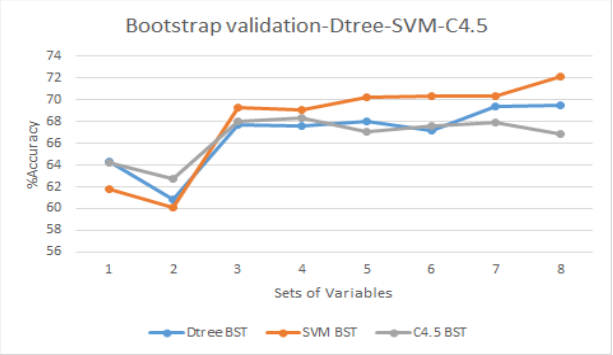

Figure 3: *Graph for Bootstrap-models Comparison*

Cross-validation:

For cross-validation we get the following figure which shows us that the PART, Dtree and Neural Net classification models are the one who give highest accuracy for our data of complete US nation and will be best suitable for prediction.

```
Accuracy
           Dtree    C45       PART      SVM       NeuralNet RanFor
Dtree                0.014214  0.045706 -0.034127  0.043508 -0.052239
C45        1.00000             0.031492 -0.048341  0.029294 -0.066453
PART       1.00000  1.00000             -0.079833 -0.002198 -0.097945
SVM        1.00000  1.00000   0.77640             0.077635 -0.018112
NeuralNet  1.00000  1.00000   1.00000   0.74726             -0.095747
RanFor     1.00000  1.00000   0.01235   1.00000   0.00581

Kappa
           Dtree    C45       PART      SVM       NeuralNet RanFor
Dtree                0.018659  0.111610 -0.070215  0.114478 -0.134280
C45        1.000000            0.092951 -0.088874  0.095819 -0.152939
PART       1.000000 1.000000            -0.181825  0.002868 -0.245889
SVM        1.000000 1.000000  1.000000             0.184693 -0.064064
NeuralNet  1.000000 1.000000  1.000000  0.925910             -0.248757
RanFor     1.000000 1.000000  0.013615  1.000000  0.009956
```

Figure 4: Accuracy Comparison for Cross-validation Models

After using the Cross-validation methods for all the models selected above we get the accuracy from each for different sets of variables.

	%ACCURACY		
Sets of Variables	Dtree CV	PART CV	NeuralNet CV
1	64.97	64.32	65.57
2	61.77	61.11	64.49
3	67.69	68.13	69.3
4	68.1	67.79	70.02
5	68.43	68.93	70.43
6	68.13	68.24	71.29
7	70.03	69.7	71.13
8	70.35	70.97	73.15

Table 4: %Accuracy for Cross-validation PART-Dtree-NeuralNet models

For selecting the best classification model for bootstrap-validation we look at the following graph which tells us that the accuracy of NeuralNet model is the highest.

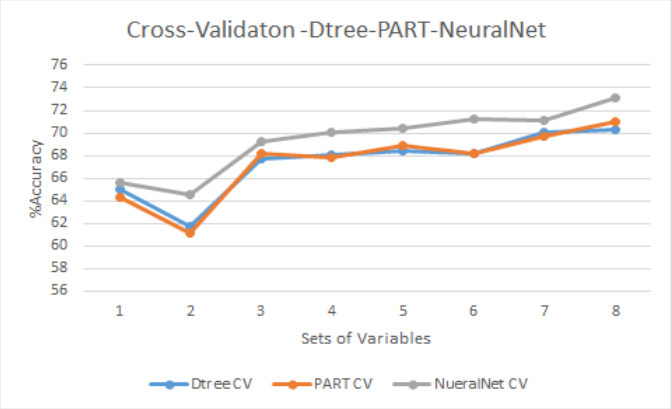

Figure 5: Graph for Bootstrap-models Comparison

Comparing Bootstrap-validation and Cross-validation:

Now we need to decide which model is the best from all the above six discussed models like 3 models for bootstrap-validation and 3 for cross-validation. After comparing each model, we get the following figure which help us to select the best model which gives overall higher accuracy for our data of complete US nation.

	Bootstrap-validation - %Accuracy			Cross-validation - %Accuracy		
Set of Variables	Dtree BST	SVM BST	C4.5 BST	Dtree CV	PART CV	NeuralNet CV
1	64.34	61.78	64.26	64.97	64.32	65.57
2	60.79	60.07	62.73	61.77	61.11	64.49
3	67.69	69.25	68.06	67.69	68.13	69.3
4	67.64	69.13	68.35	68.1	67.79	70.02
5	68.04	70.23	67.07	68.43	68.93	70.43
6	67.21	70.35	67.55	68.13	68.24	71.29
7	69.45	70.31	67.96	70.03	69.7	71.13
8	69.48	72.12	66.86	70.35	70.97	73.15

Table 5: %Accuracy for both Bootstrap and Cross Validation models

For selecting the best classification model from all the six model we look at the following graph which tells us that the accuracy of NeuralNet model is the highest from all six.

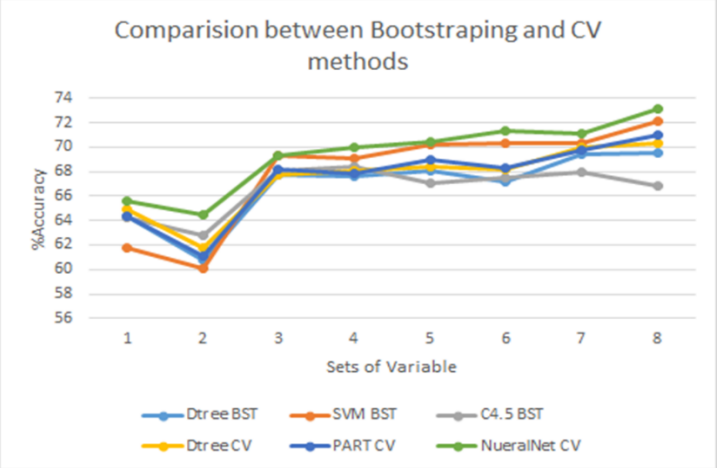

Figure 6: Graph for Comparing and Selecting Classification Model

So, for Classification of our graduation data we will use Artificial Neural Network classification model with cross-validation method.

Evaluation and Deployment

As we had selected our best classification model, now will look after the accuracy of this model toward each states of the USA. This model will help us to understand the variables playing important role for graduation rate in each state, the one variable state with high accuracy is the variable set which define the graduation rate in that state. In the following figure the highlighted accuracy for any sets of variables is the important variables for that respective state graduation rate.

	% Accuracy							
	1	2	3	4	5	6	7	8
States								
Tennessee	79.4	81.4	**82.8**	79.4	81.9	82.5	81.6	79.69
Wisconsin	76.5	76.4	76.5	76.5	80.6	80.7	80.6	**81.4**
Texas	67.4	67.8	69	70.4	70	76.7	77.5	**78.6**
Pennsylvania	72.6	71.8	76.9	77.6	77.0	**78.1**	77.6	77.6
Ohio	69.8	77.5	79.9	80.6	79.7	81	**81.2**	78.23
New Jersey	70.1	75.8	78.6	78.0	77.4	79.5	**81.5**	77.93
Massachusetts	84.3	76.6	82.6	83.5	85.8	87	**87.3**	84
Iowa	66.5	61	71.3	**88.3**	74.3	79.1	78	74.8
Connecticut	71.9	82.7	**85.1**	83.4	83.0	82.9	85.1	80.9
Maryland	**81.6**	81.6	71.6	63.3	66.6	61.6	61.6	66.6
Indiana	69.1	73.1	**75.9**	70.5	69.1	69.2	69.8	67.9
California	63.1	61.8	67.0	67.7	65.5	72.8	72.8	**74.0**
Missouri	60.8	62.4	68.5	**69.6**	64.6	65	63.5	65.5
Arkansas	66.3	66.5	66.5	66.5	68.3	66.3	**68.5**	65.1
Maine	86.66	88.33	88.33	88.33	85	86.66	**88.3**	86.66
Illinois	72.8	70.2	77.3	75.3	73.7	80.3	**81.9**	80.6

North Carolina	94.3	95.4	94.4	95.4	95.2	**95.5**	94.4	95.4
Washington	89.5	89.5	89.5	89.5	89.5	89.5	89.5	**91.5**
Virginia	89.16	89.02	89.16	89.02	89.16	89.02	**89.16**	89.16
Utah	85	85	**86.6**	86.6	86.6	85	86.6	86.6
New York	73.6	76.0	77.7	79.6	79.6	79.2	80.0	**81.1**
New Hampshire	66.6	76.6	76.6	75	70	68.3	63.3	**78.3**
Nebraska	61.6	61.6	66.6	70	**76.6**	70	80	71.6
Minnesota	71.6	70.0	72.6	**72.8**	72.6	74.2	71.4	72.4
Michigan	74.7	75.3	75.2	76.7	74.8	76.	76.7	**77.2**
Kansas	73.5	66	76.8	**77**	70.8	67	67.1	76.8
Alabama	88.6	84.6	85	85.8	**90.3**	88.3	86.6	88.5
Delaware	96.6	95	96.6	96.6	95	96.6	**96.6**	95
Colorado	87	87	87	87	86.6	89	86.6	**89.3**
South Carolina	96.6	96.6	96.6	96.6	**96.9**	96.6	96.6	96.6
Georgia	95.3	95.4	95.2	95.1	96.0	**96.2**	96.2	96.1
Arizona	85	**87.5**	85	85	85	84.5	85	84

Table 6: %Accuracy for each state with sets of variables using NeuralNet Cross-validation model

From the above if we took state "Tennessee" the %Accuracy for set of variables "3" is the highest which is 82.8. This implies that in state "Tennessee" variables "Poverty Population" and "Black Cohort rate" are the important predictor of the state graduation rate. So, by using the above figure we can evaluate important predictor variables for graduation rate in each state.

Conclusion

This report allows to make the decision about the "Graduation rate" predictor variables for each state individually. As we look after each state we get the following figure;

States	Set of Variables	Predictors
Tennessee	3	POV+BL
Wisconsin	8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5
Texas	8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5
Pennsylvania	6	POV+BL+RUR+Cohort+HIS
Ohio	7	POV+BL+RUR+Cohort+HIS+LimEng
New Jersey	7	POV+BL+RUR+Cohort+HIS+LimEng
Massachusetts	7	POV+BL+RUR+Cohort+HIS+LimEng
Iowa	4	POV+BL+RUR
Connecticut	3	POV+BL
Maryland	1	POV
Indiana	3	POV+BL
California	8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5
Missouri	4	POV+BL+RUR
Arkansas	7	POV+BL+RUR+Cohort+HIS+LimEng
Maine	7	POV+BL+RUR+Cohort+HIS+LimEng
Illinois	7	POV+BL+RUR+Cohort+HIS+LimEng
North Carolina	6	POV+BL+RUR+Cohort+HIS
Washington	8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5
Virginia	7	POV+BL+RUR+Cohort+HIS+LimEng
Utah	3	POV+BL
New York	8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5
New Hampshire	8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5
Nebraska	5	POV+BL+Cohort+RUR
Minnesota	4	POV+BL+RUR
Michigan	8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5
Kansas	4	POV+BL+RUR
Alabama	5	POV+BL+Cohort+RUR
Delaware	7	POV+BL+RUR+Cohort+HIS+LimEng
Colorado	8	POV+BL+RUR+Cohort+HIS+LimEng+NoEngBy5
South Carolina	5	POV+BL+Cohort+RUR
Georgia	6	POV+BL+RUR+Cohort+HIS
Arizona	2	BL

Table 7: Predictors of Graduation rate by State

Thus, from above figure we can look after variables which play important role in deciding graduation rate of each state. So, for each state government need to find the solution to increase the graduation in the variables specified by above table.

References:

[1] https://files.eric.ed.gov/fulltext/ED466523.pdf

[2] https://pdfs.semanticscholar.org/2b19/c569a03c5d2d232d13cd3eb1b56dc882d7db.pdf

[3] https://pdfs.semanticscholar.org/df6f/23c834a186dff84b4114eb97034a05ebf695.pdf

[4] Wikipedia