Feature Selection of Post- graduation Income of College Students in United States

Qiang Hao
Western Washington University

Ewan Wright, Khaled Rasheed, Yan Liu

Rationales

- Out-dated Literatures
  - 1980s – 1990s
  - Small sample sizes (100 - 200)
  - Limits of applied methods

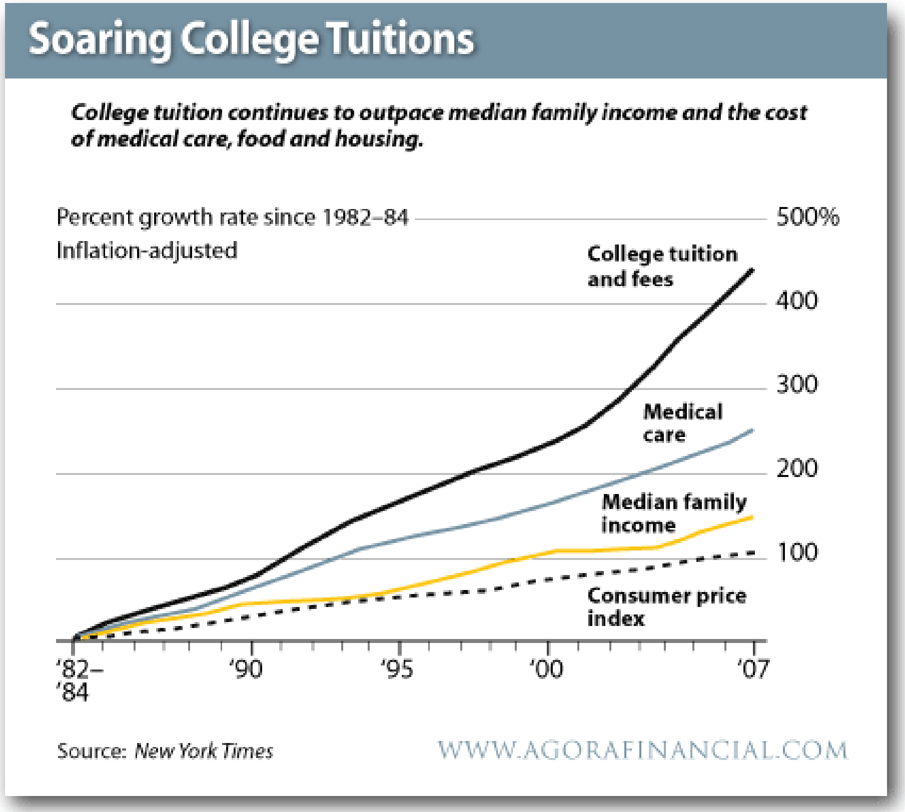
Rationales

- Out-dated Literatures
  - 1980s – 1990s
  - Small sample sizes (100 - 200)
  - Limits of applied methods

- Old Trends of Universities / Research
  - Not to focus on post-graduation income
  - Reluctant to disclose information

Rationales

- Demands from parents
  - Investment (Autor 2014; Goldin & Katz 2009; Hout 2012)
  - High cost



Soaring College Tuitions

College tuition continues to outpace median family income and the cost of medical care, food and housing.

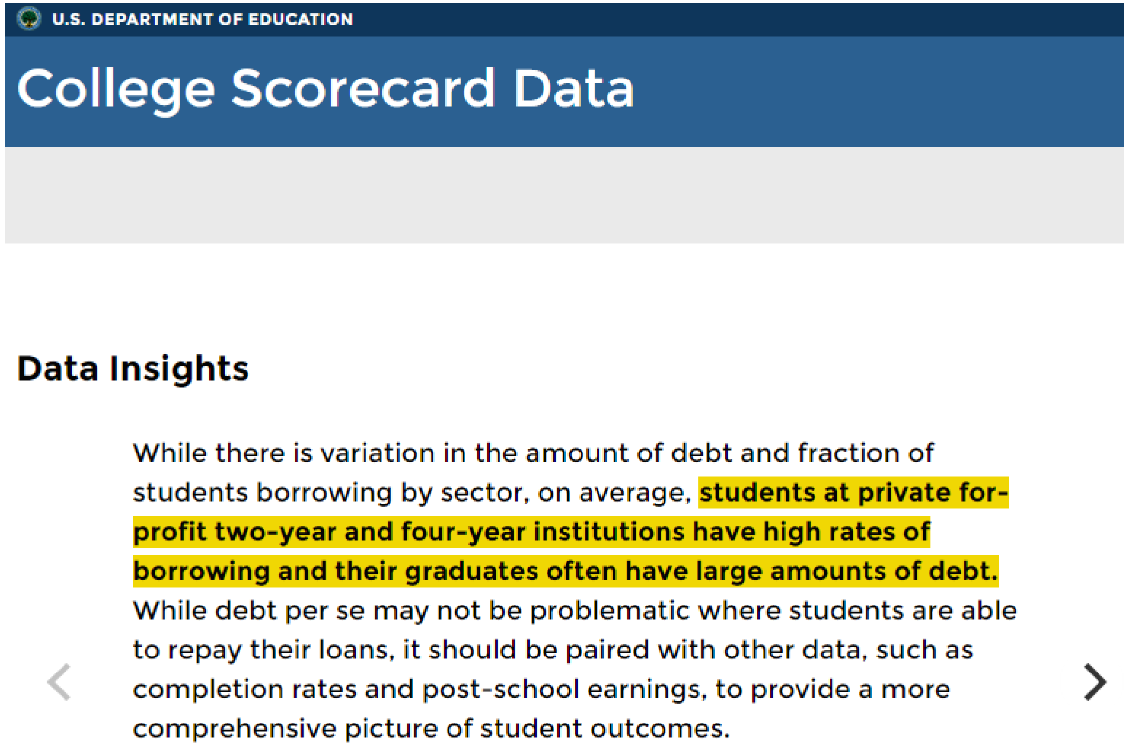
Source: *New York Times*

WWW.AGORAFINANCIAL.COM

Rationales

- Demands from parents
  - Investment (Autor 2014; Goldin & Katz 2009; Hout 2012)
  - High cost

- Newly release data
  - U.S. Department of Education
  - College Scorecard



U.S. DEPARTMENT OF EDUCATION

College Scorecard Data

Data Insights

While there is variation in the amount of debt and fraction of students borrowing by sector, on average, **students at private for-profit two-year and four-year institutions have high rates of borrowing and their graduates often have large amounts of debt.** While debt per se may not be problematic where students are able to repay their loans, it should be paired with other data, such as completion rates and post-school earnings, to provide a more comprehensive picture of student outcomes.

Research Questions

- What are the most important attributes of post-graduation income of college students who graduate with debt repayment obligations?
- To what extent can the selected attributes classify post-graduation income of college students who graduate with debt repayment obligations?

Data

- Release in October, 2015 by College ScoreCard under the United States Department of Education (https://collegescorecard.ed.gov/data/)
- Students who used financial aid during their college study period
- Organized by student cohorts at a university

Data

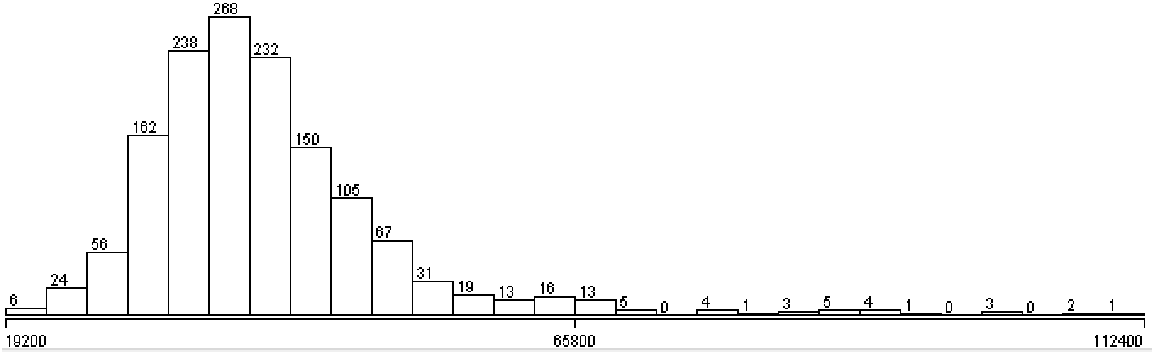
- Target
    - Mean value of 6-year post-graduation Income
    - 1997, 1999, 2001, 2003 and 2005

Data

- Target
  - Mean value of 6-year post-graduation Income
  - ~~1997, 1999,~~ 2001, 2003 and 2005

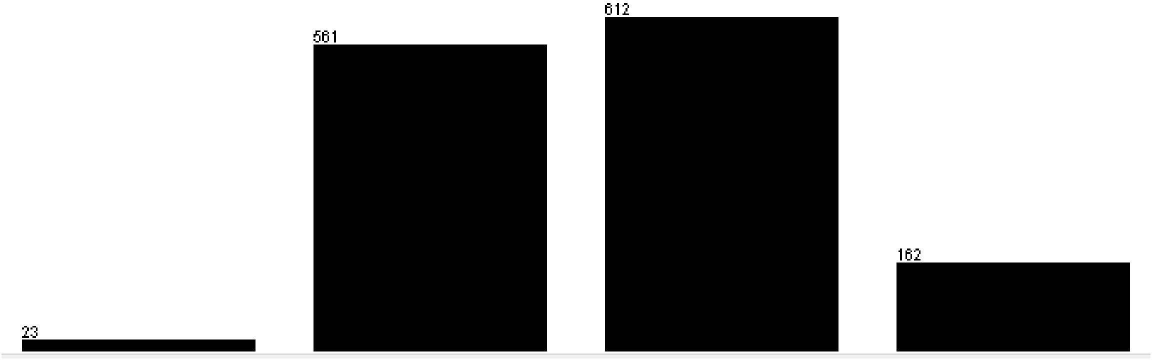
Data

- Target
  - Mean value of 6-year post-graduation Income
  - ~~1997, 1999,~~ 2001, 2003 and 2005
  - Discretized into four groups based on the information from American Individual Income Distribution (U.S. Census Bureau, 2010):
    1. Very low: From 0 to 25000
    2. Low: From 25000 to 37500
    3. Middle: From 37500 to 50000
    4. High: Above 50000



Data

- Target
  - Mean value of 6-year post-graduation Income
  - ~~1997, 1999,~~ 2001, 2003 and 2005
  - Discretized into four groups based on the information from American Individual Income Distribution (U.S. Census Bureau, 2010):
    1. Very low: From 0 to 25000
    2. Low: From 25000 to 37500
    3. Middle: From 37500 to 50000
    4. High: Above 50000



Data

- Attributes
    - Preselected based on domain knowledge
    - Exclude irrelevant attributes, such as *latitude of the institution, accreditor of the institution, or percent of students who passed away within 6 years after graduation*
    - Include 30 attributes in 5 groups:
        - School
        - Admission
        - Cost
        - Student Cohort
        - Socioeconomic Status of Students' Family

Data

- Attributes
  - Preselected based on domain knowledge
  - Exclude irrelevant attributes, such as *latitude of the institution, accreditor of the institution, or percent of students who passed away within 6 years after graduation*
  - Include 30 attributes in 5 groups:
    - School
    - Admission
    - Cost
    - Student Cohort
    - Socioeconomic Status of Students’ Family
  - Standardization (28 numeric attributes) / One-hot encoding (2 nominal attributes)

Data

- Targets
- Attributes
- 1429 student cohorts were included

Data Analysis – Feature Selection

- **Filter methods**

- **Stepwise wrapper methods**

- **Naturally inspired algorithms**

Data Analysis – Feature Selection

- **Filter methods**
  - OneR algorithm
  - Relief-based selection
  - Chi-square selection
  - Gain-ratio-based selection
  - Information-gain-based selection
- **Stepwise wrapper methods**
- **Naturally inspired algorithms**

Data Analysis – Feature Selection

- **Filter methods**
    - OneR algorithm
    - Relief-based selection
    - Chi-square selection
    - Gain-ratio-based selection
    - Information-gain-based selection
- **Stepwise wrapper methods**
    - Forward / backward selection
    - Logistic Regression
- **Naturally inspired algorithms**

Data Analysis – Feature Selection

- **Filter methods**
  - OneR algorithm
  - Relief-based selection
  - Chi-square selection
  - Gain-ratio-based selection
  - Information-gain-based selection
- **Stepwise wrapper methods**
  - Forward / backward selection
  - Logistic Regression
- **Naturally inspired algorithms**
  - Genetic Algorithm
  - Logistic Regression

Data Analysis – Feature Selection

- **Filter methods (13 Attributes)**
  - OneR algorithm
  - Relief-based selection
  - Chi-square selection
  - Gain-ratio-based selection
  - Information-gain-based selection
- **Stepwise wrapper methods (9 Attributes)**
  - Forward / backward selection
  - Logistic Regression
- **Naturally inspired algorithms (22 Attributes)**
  - Genetic Algorithm
  - Logistic Regression

Data Analysis – Feature Selection

- **Filter methods (13 Attributes)**
  - OneR algorithm
  - Relief-based selection
  - Chi-square selection
  - Gain-ratio-based selection
  - Information-gain-based selection
- **Stepwise wrapper methods (9 Attributes)**
  - Forward / backward selection
  - Logistic Regression
- **Naturally inspired algorithms (22 Attributes)**
  - Genetic Algorithm
  - Logistic Regression

- Logistic Regression
- Support Vector Machine (*Pearson VII function kernel*)

Data Analysis – Feature Selection

- **Filter methods (13 Attributes)**
  - OneR algorithm
  - Relief-based selection
  - Chi-square selection
  - Gain-ratio-based selection
  - Information-gain-based selection
- **Stepwise wrapper methods (9 Attributes)**
  - Forward / backward selection
  - Logistic Regression
- **Naturally inspired algorithms (22 Attributes)**
  - Genetic Algorithm
  - Logistic Regression

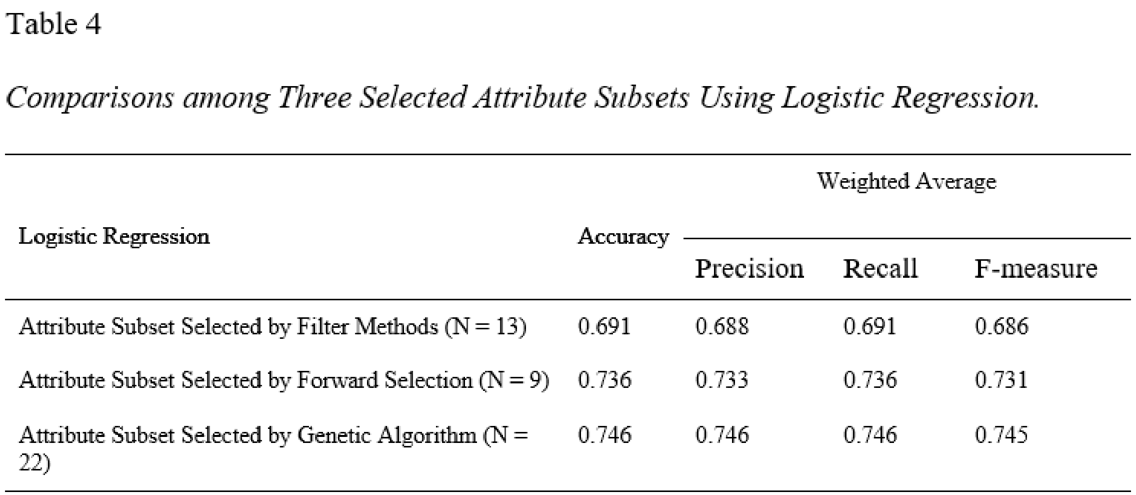


Table 4

Comparisons among Three Selected Attribute Subsets Using Logistic Regression.

		Weighted Average		
Logistic Regression	Accuracy	Precision	Recall	F-measure
Attribute Subset Selected by Filter Methods (N = 13)	0.691	0.688	0.691	0.686
Attribute Subset Selected by Forward Selection (N = 9)	0.736	0.733	0.736	0.731
Attribute Subset Selected by Genetic Algorithm (N = 22)	0.746	0.746	0.746	0.745

Data Analysis – Feature Selection

- **Filter methods (13 Attributes)**
  - OneR algorithm
  - Relief-based selection
  - Chi-square selection
  - Gain-ratio-based selection
  - Information-gain-based selection
- **Stepwise wrapper methods (9 Attributes)**
  - Forward / backward selection
  - Logistic Regression
- **Naturally inspired algorithms (22 Attributes)**
  - Genetic Algorithm
  - Logistic Regression

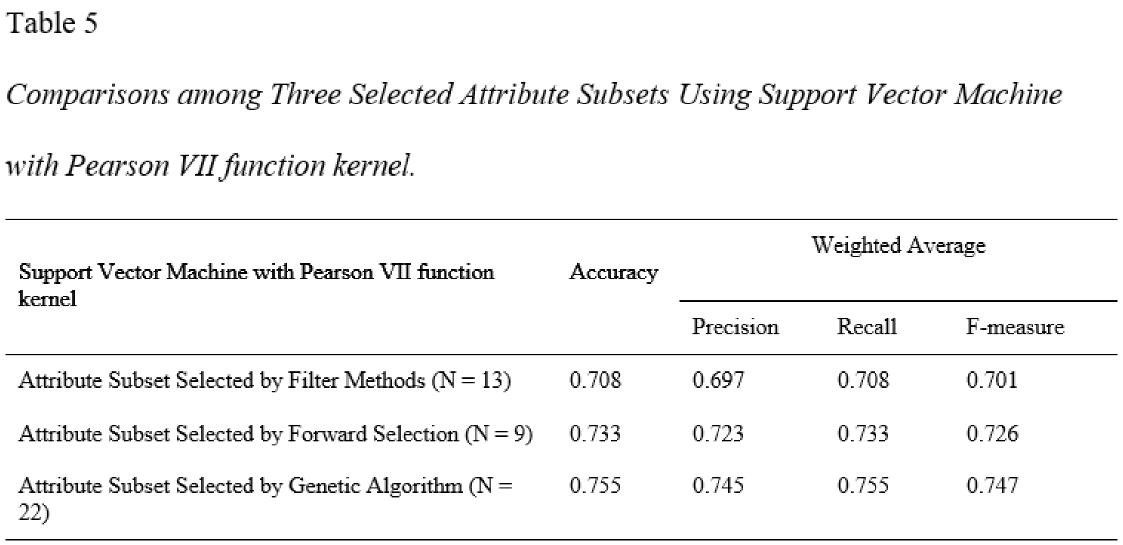


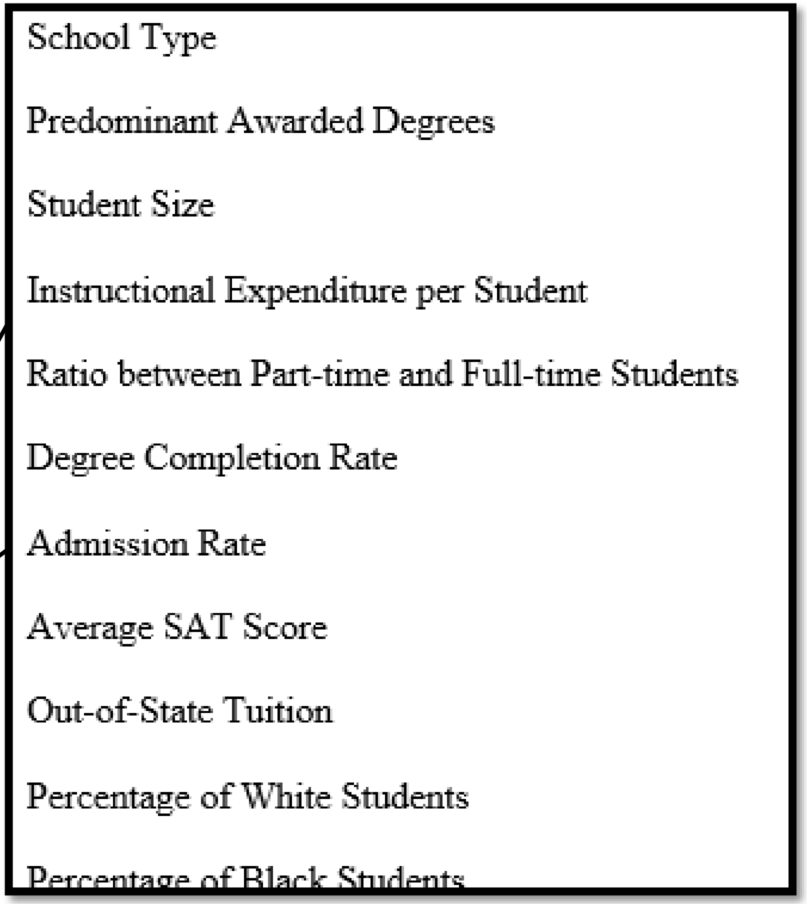
Table 5

Comparisons among Three Selected Attribute Subsets Using Support Vector Machine with Pearson VII function kernel.

Support Vector Machine with Pearson VII function kernel	Accuracy	Weighted Average		
		Precision	Recall	F-measure
Attribute Subset Selected by Filter Methods (N = 13)	0.708	0.697	0.708	0.701
Attribute Subset Selected by Forward Selection (N = 9)	0.733	0.723	0.733	0.726
Attribute Subset Selected by Genetic Algorithm (N = 22)	0.755	0.745	0.755	0.747

Data Analysis – Feature Selection

- **Filter methods (13 Attributes)**
  - OneR algorithm
  - Relief-based selection
  - Chi-square selection
  - Gain-ratio-based selection
  - Information-gain-based selection
- **Stepwise wrapper methods (9 Attributes)**
  - Forward / backward selection
  - Logistic Regression
- **Naturally inspired algorithms (22 Attributes)**
  - Genetic Algorithm
  - Logistic Regression



School Type

Predominant Awarded Degrees

Student Size

Instructional Expenditure per Student

Ratio between Part-time and Full-time Students

Degree Completion Rate

Admission Rate

Average SAT Score

Out-of-State Tuition

Percentage of White Students

Percentage of Black Students

Data Analysis – Classification

- Single Learners
  - Bayes-based algorithms
  - Function-based algorithms
  - Instance-based algorithms
  - Tree-based algorithms
  - Rule-based algorithms
- Ensemble Learning
  - Bagging
  - Randomization
  - Bosting

Data Analysis – Classification

Single Learners

- ***Bayes-based algorithms***:

  Naive Bayes Update, Bayes Net

- ***Function-based algorithms***:

  Logistic Regression, Support Vector Machine, Multilayer Perceptron

- ***Instance-based algorithms***:

  Distance-weighted K-Nearest Neighbor

- ***Tree-based algorithms***:

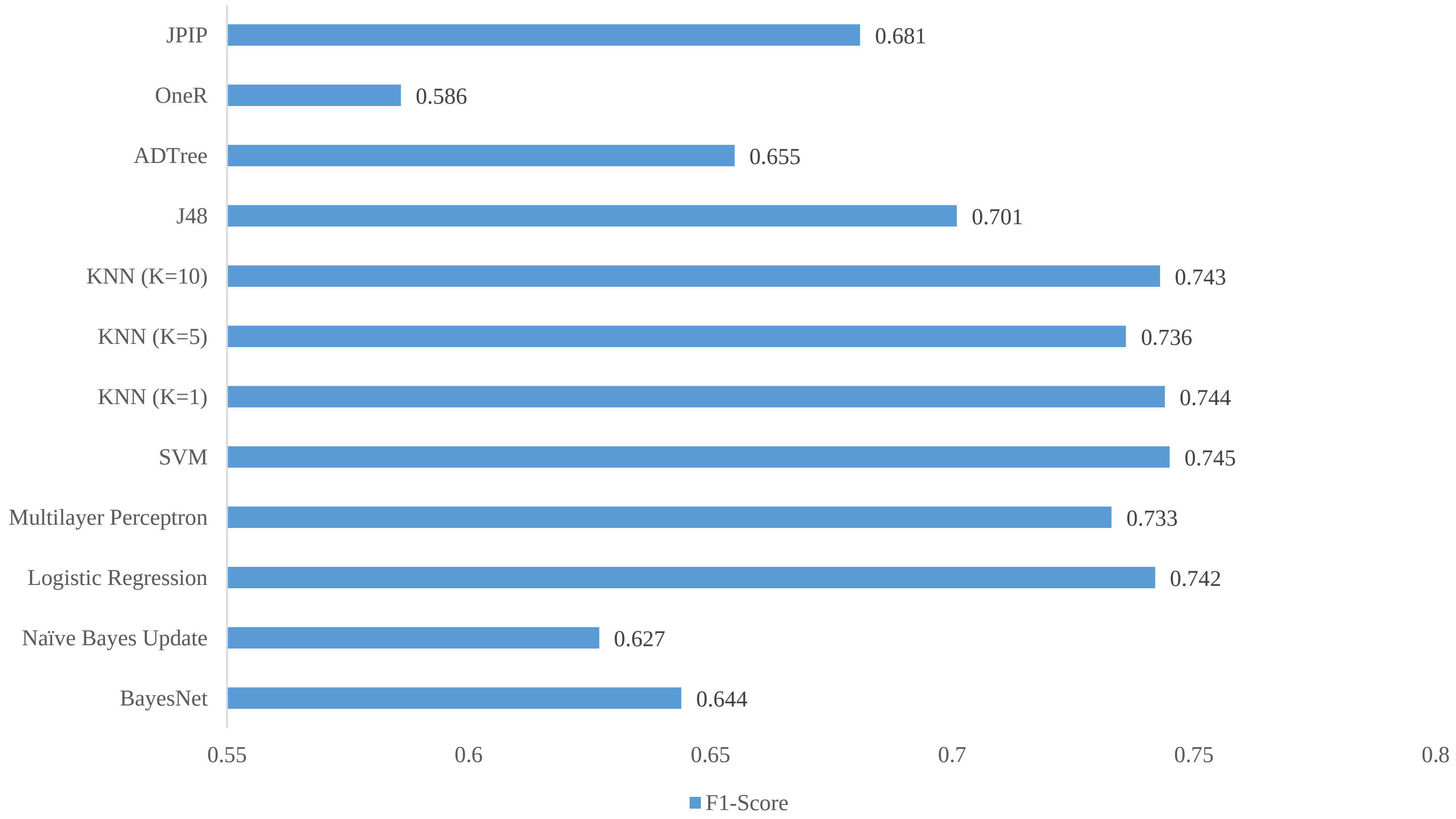
  J48, Multiclass Alternating Decision Tree

- ***Rule-based algorithms***:

  OneR, JRIP

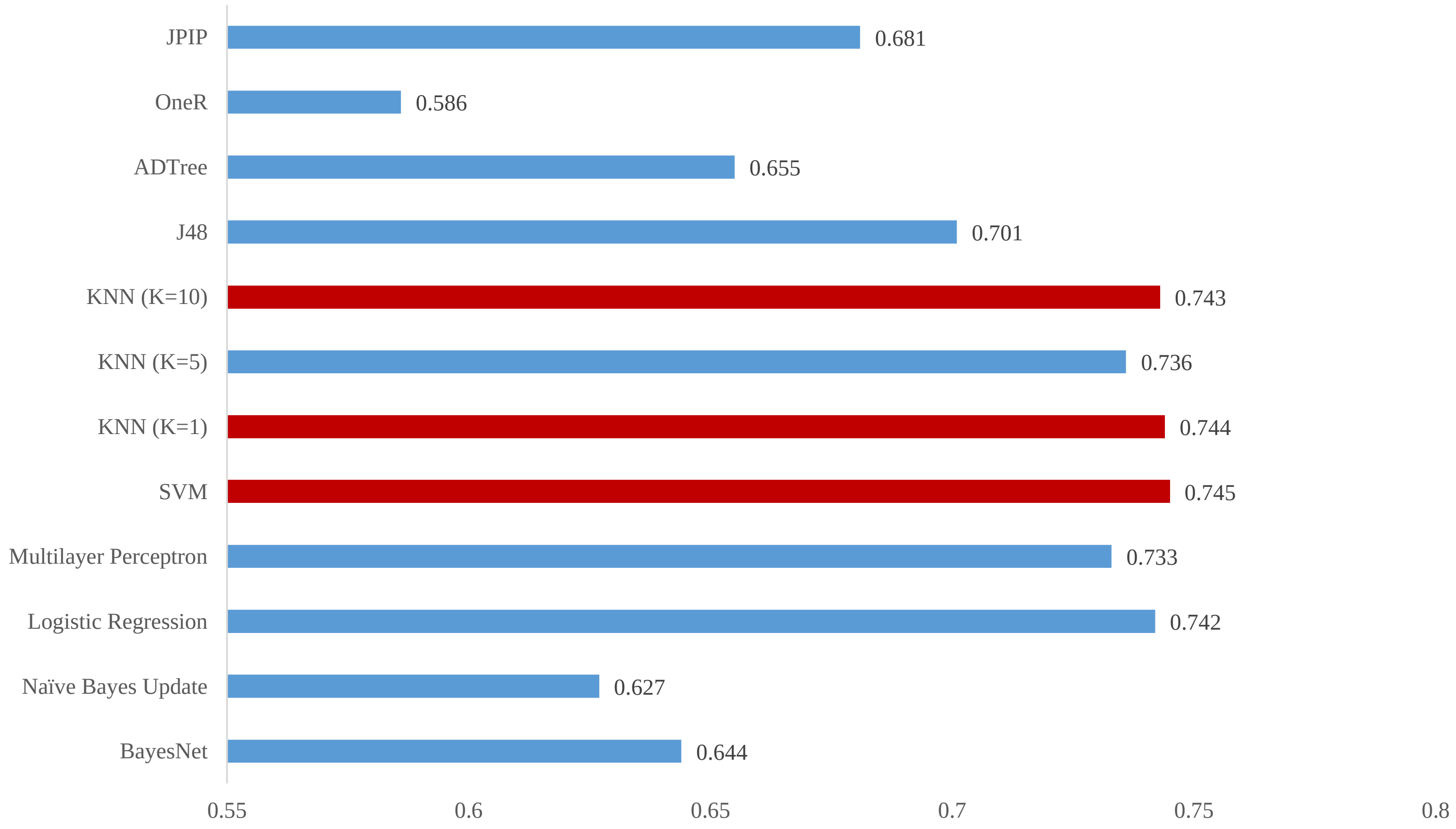
Data Analysis – Classification

- Single Learners



Data Analysis – Classification

- Single Learners



Data Analysis – Classification

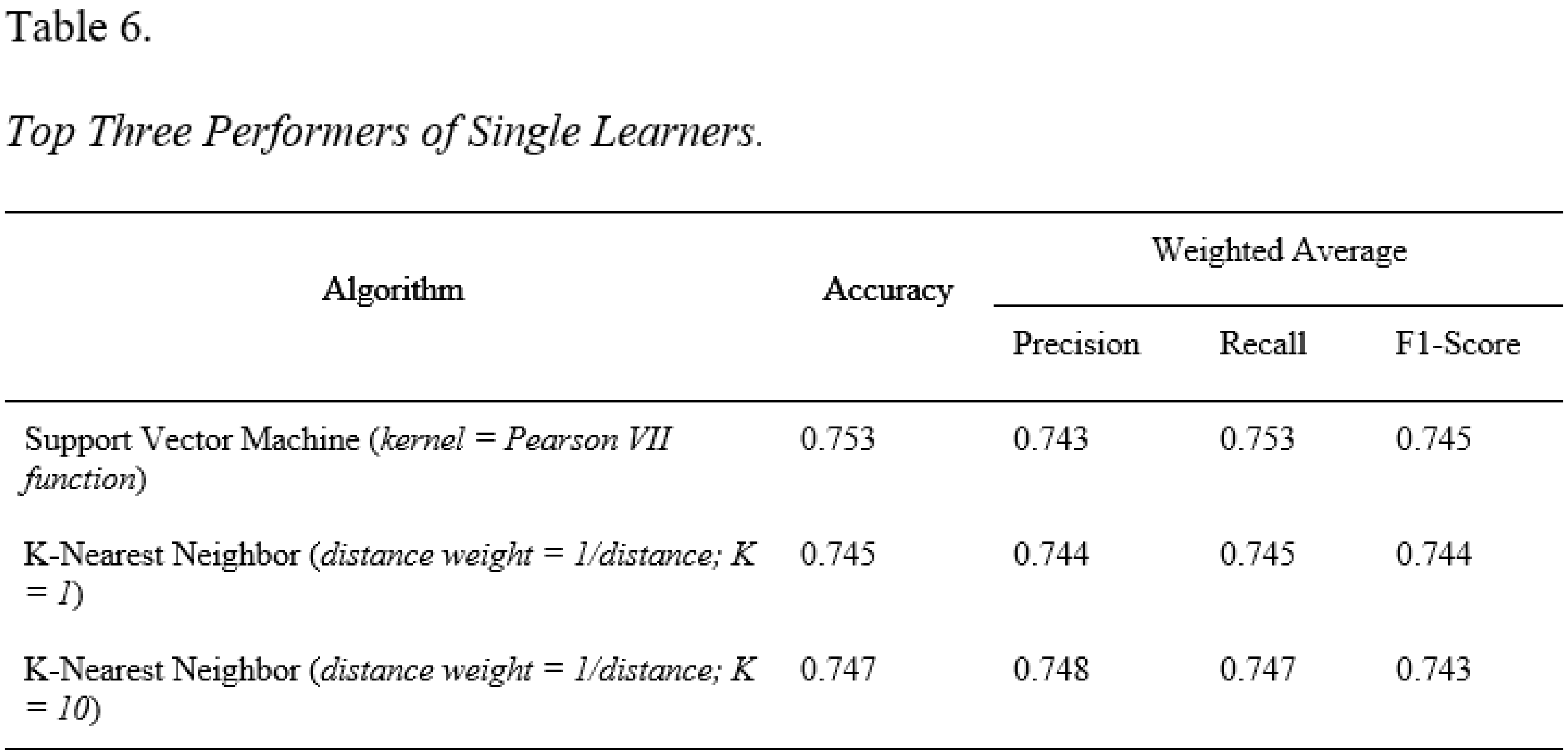


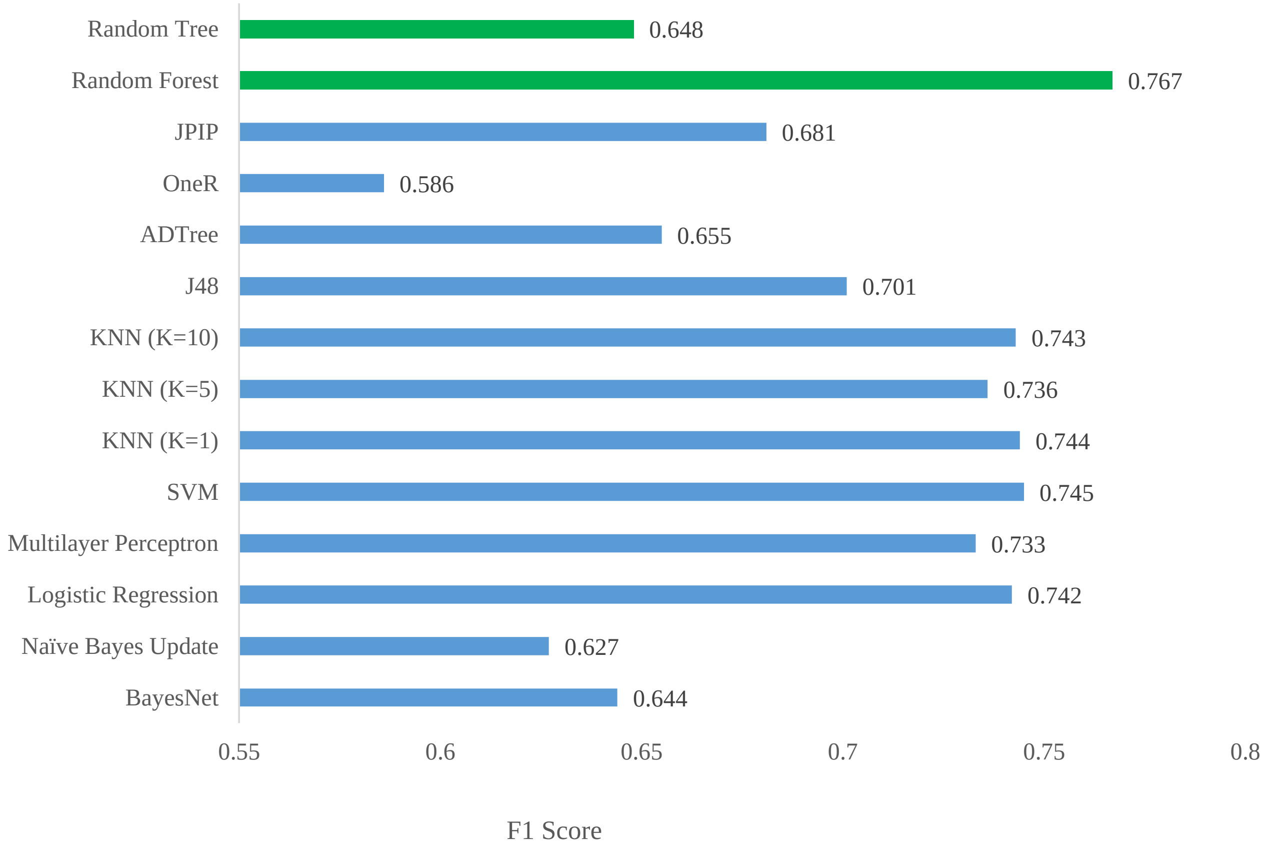
Table 6.

Top Three Performers of Single Learners.

Algorithm	Accuracy	Weighted Average		
		Precision	Recall	F1-Score
Support Vector Machine (*kernel = Pearson VII function*)	0.753	0.743	0.753	0.745
K-Nearest Neighbor (*distance weight = 1/distance; K = 1*)	0.745	0.744	0.745	0.744
K-Nearest Neighbor (*distance weight = 1/distance; K = 10*)	0.747	0.748	0.747	0.743

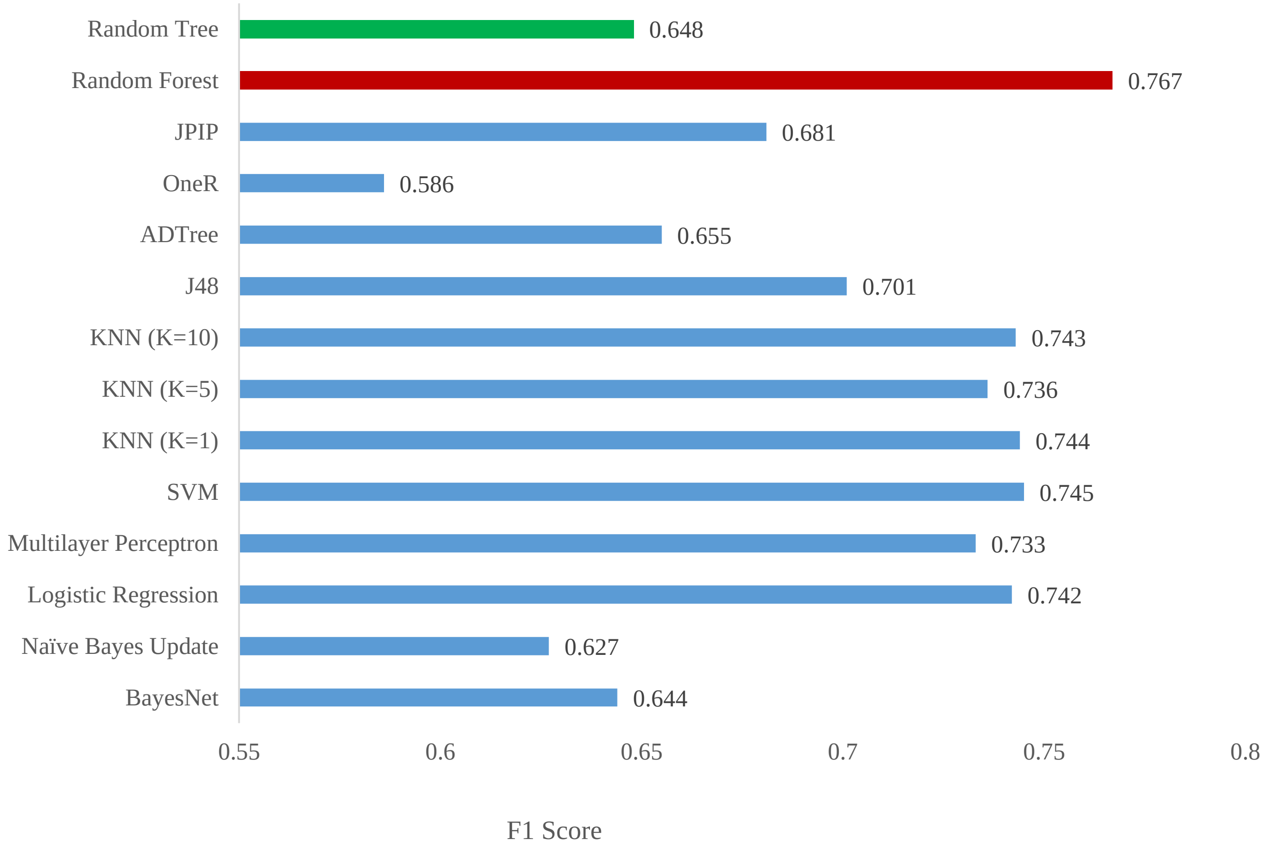
Data Analysis – Classification

- Randomization



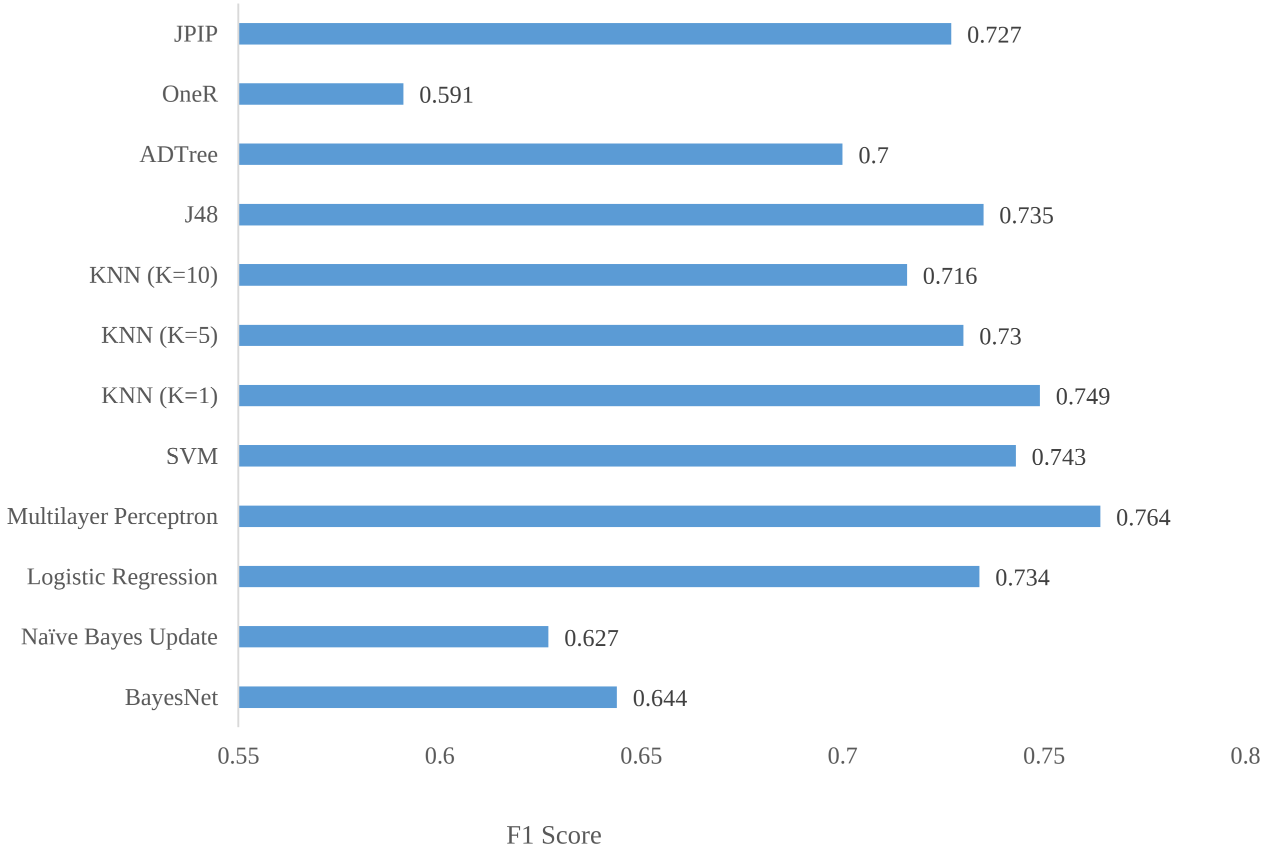
Data Analysis – Classification

- Randomization



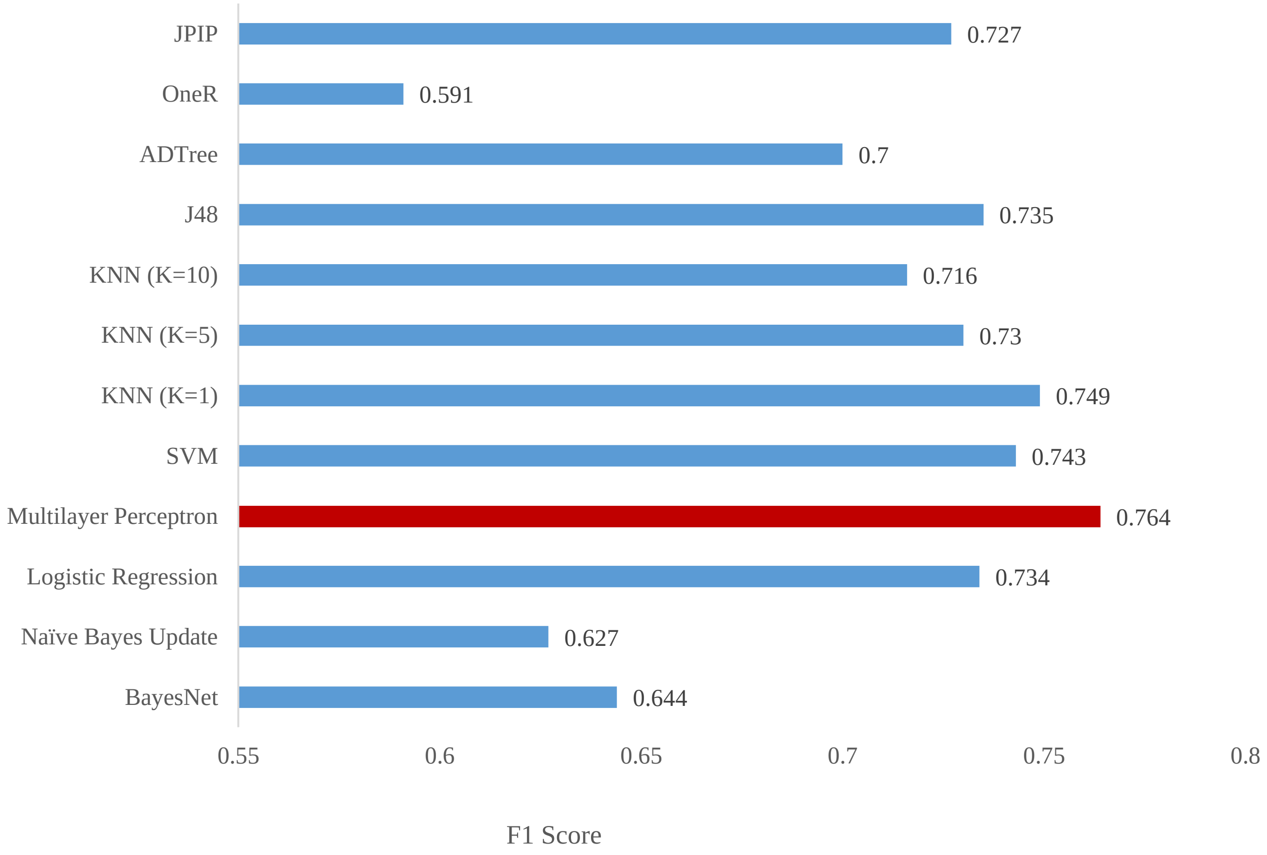
Data Analysis – Classification

- Bagging



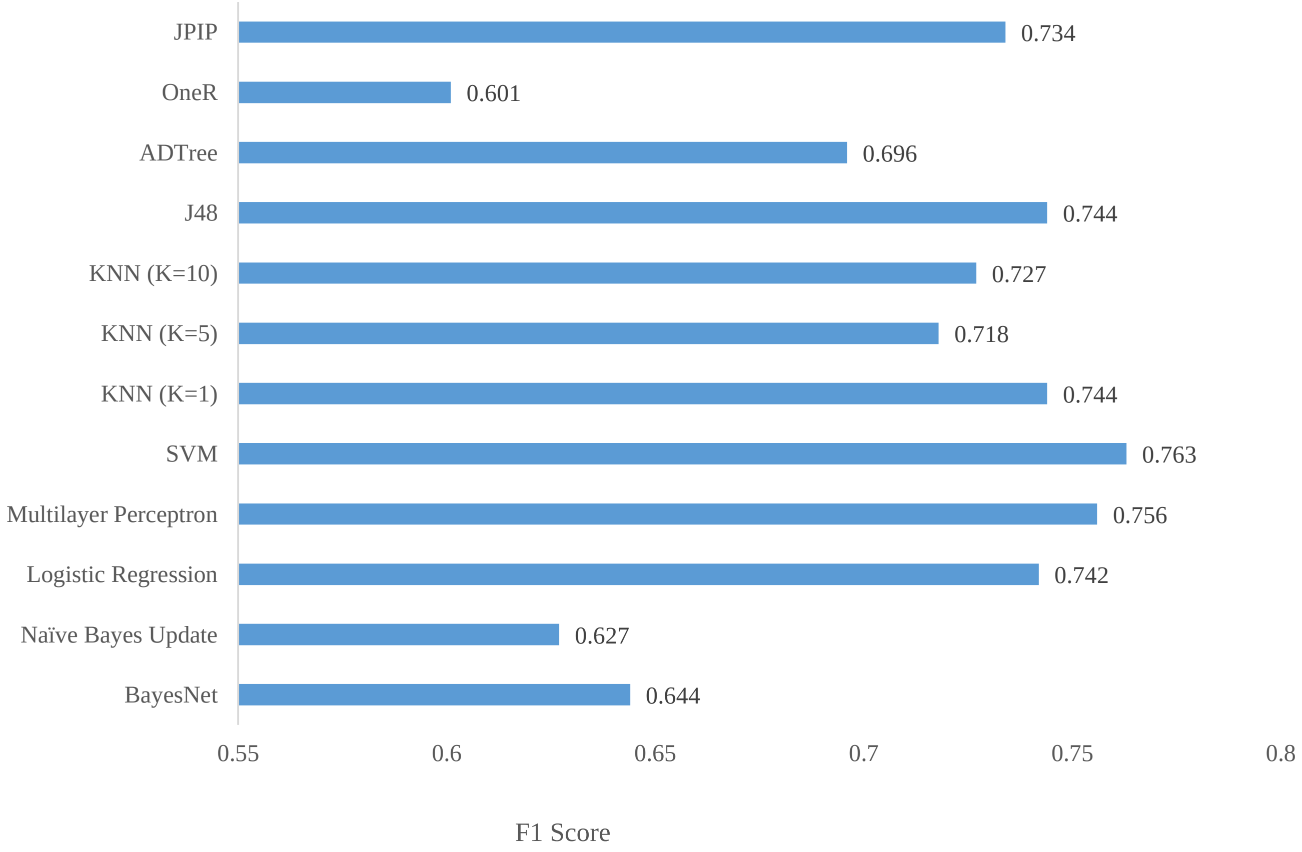
Data Analysis – Classification

- Bagging



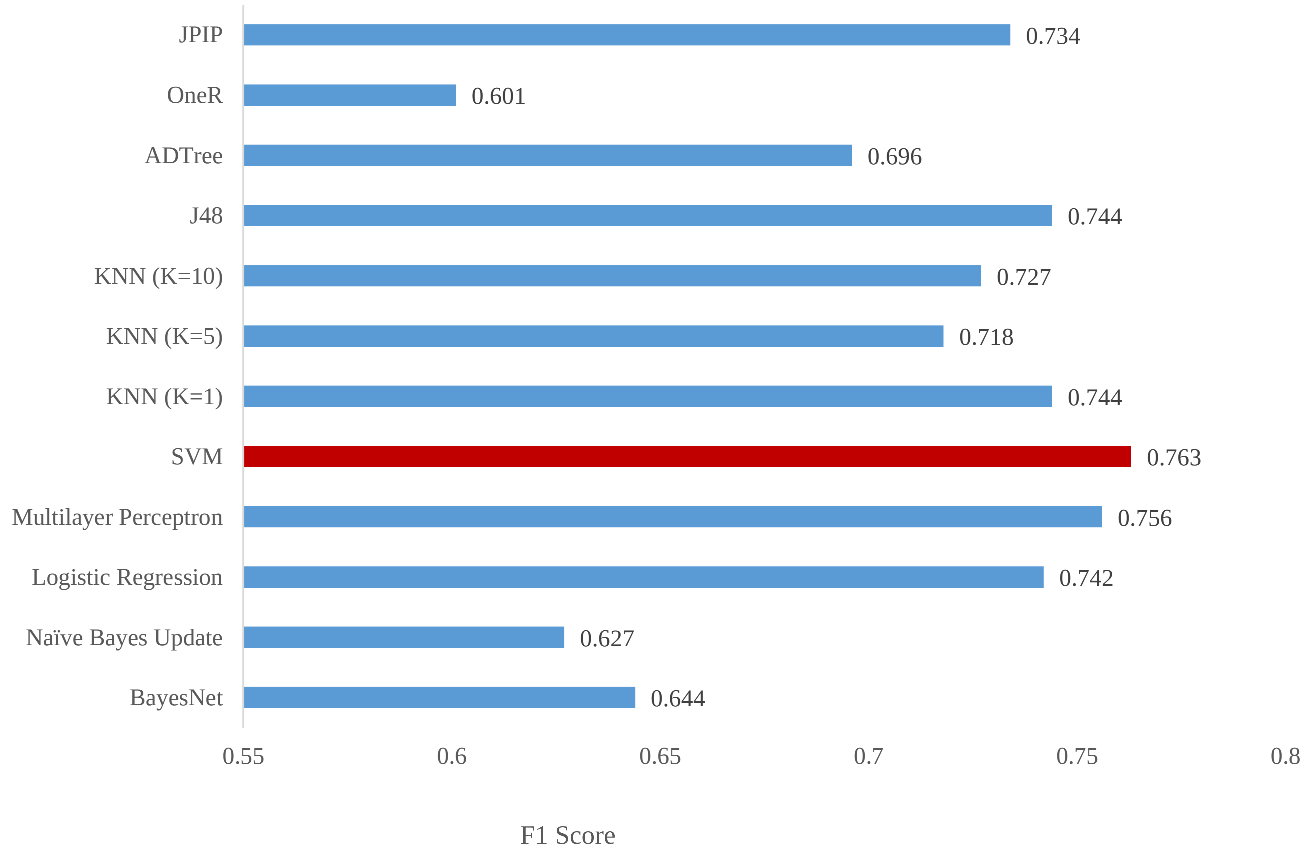
Data Analysis – Classification

- Boosting



Data Analysis – Classification

- Boosting



Data Analysis – Classification

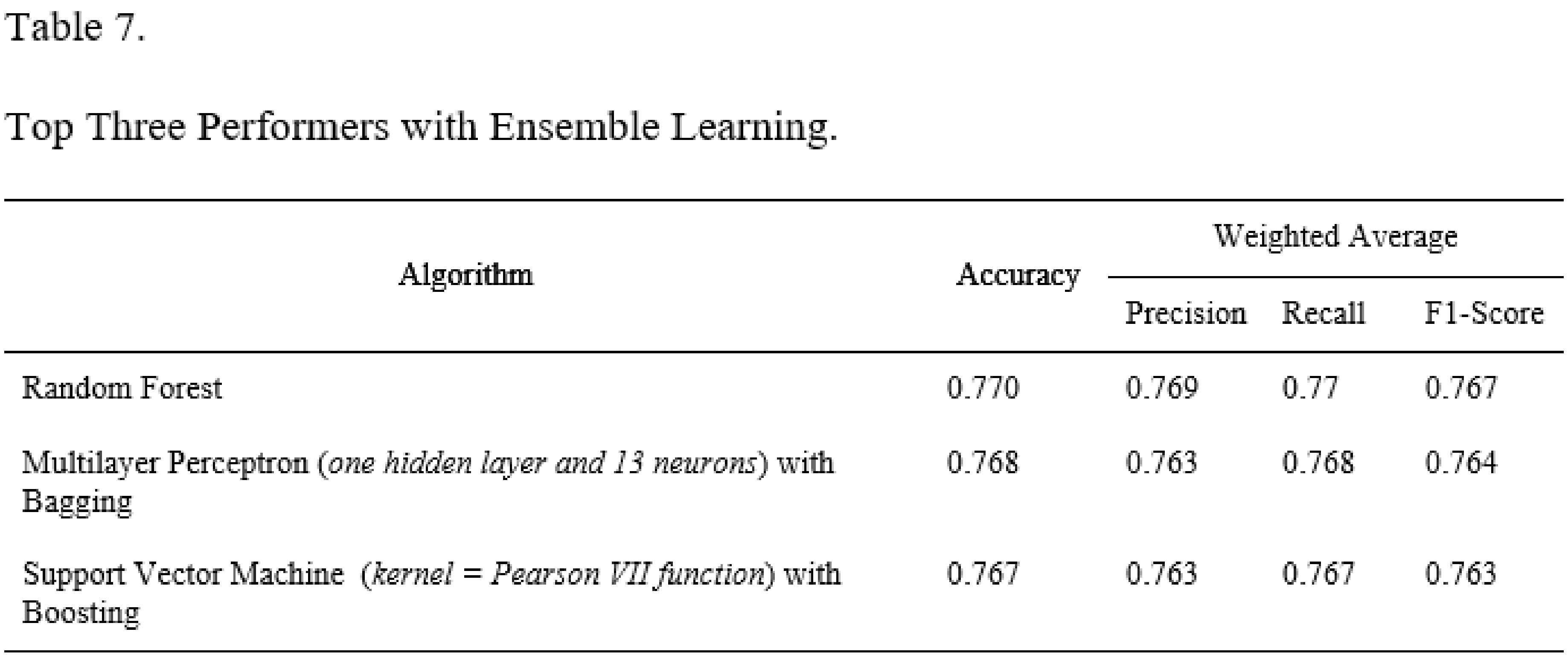
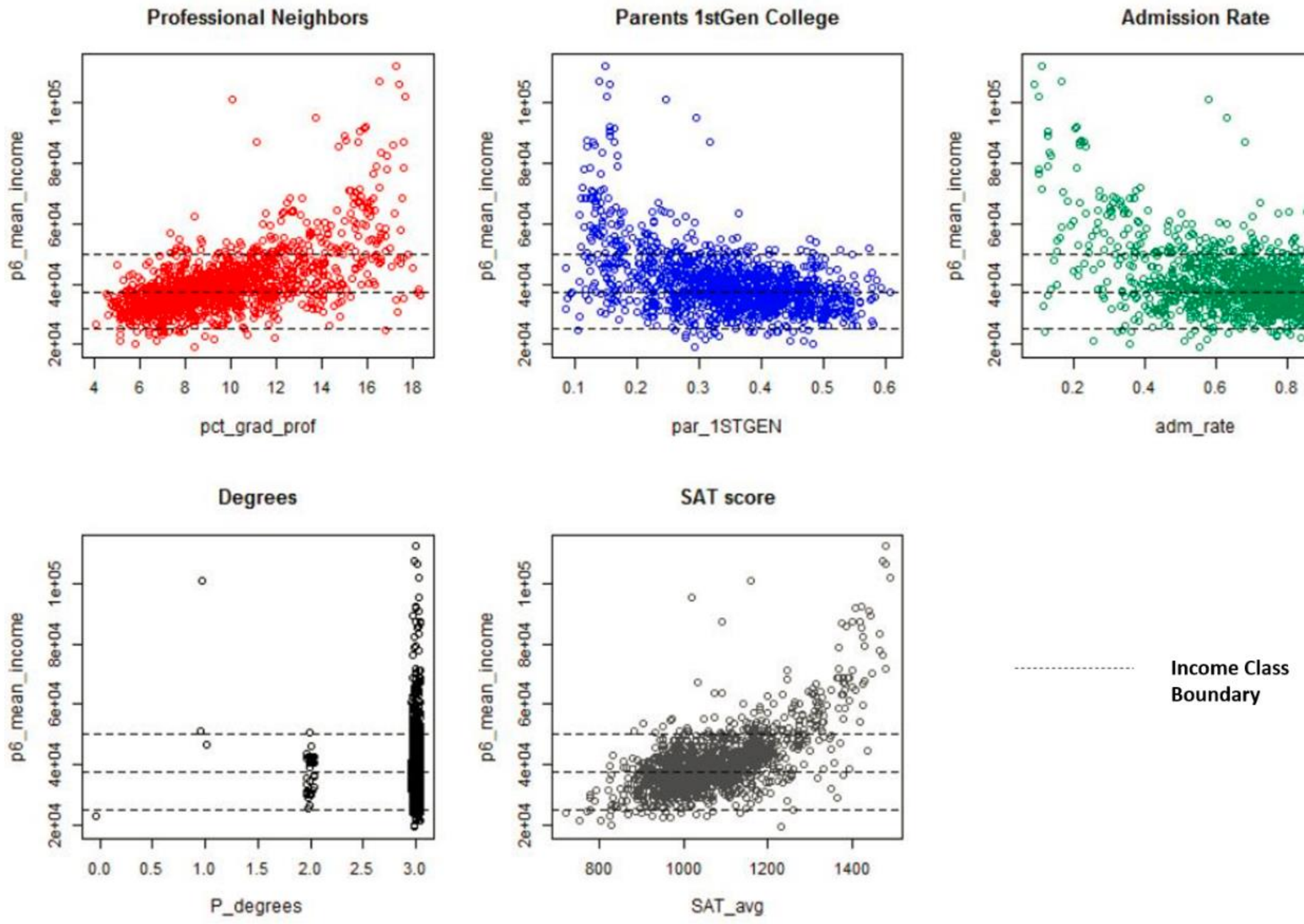


Table 7.

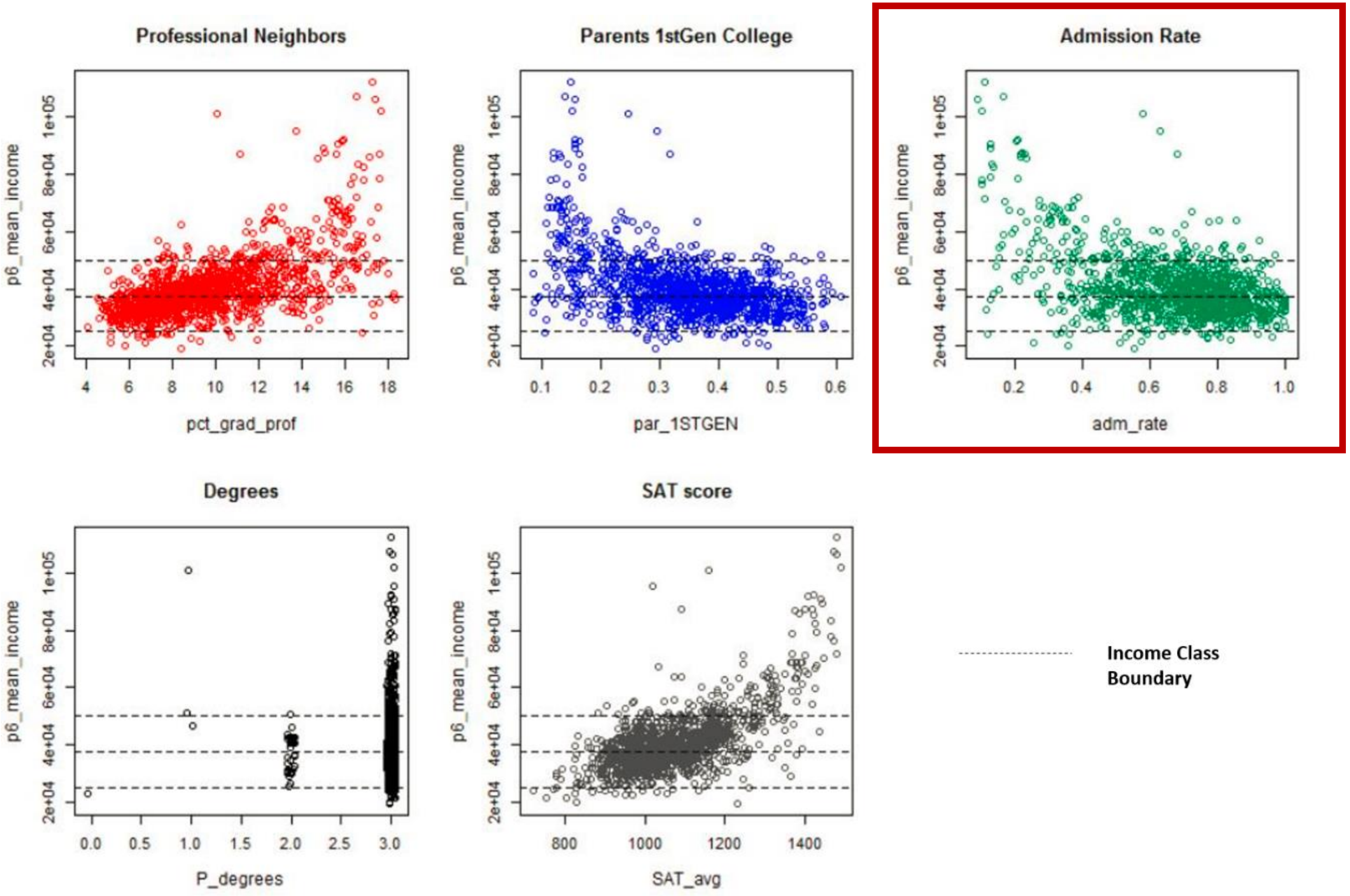
Top Three Performers with Ensemble Learning.

Algorithm	Accuracy	Weighted Average		
		Precision	Recall	F1-Score
Random Forest	0.770	0.769	0.77	0.767
Multilayer Perceptron (*one hidden layer and 13 neurons*) with Bagging	0.768	0.763	0.768	0.764
Support Vector Machine (*kernel = Pearson VII function*) with Boosting	0.767	0.763	0.767	0.763

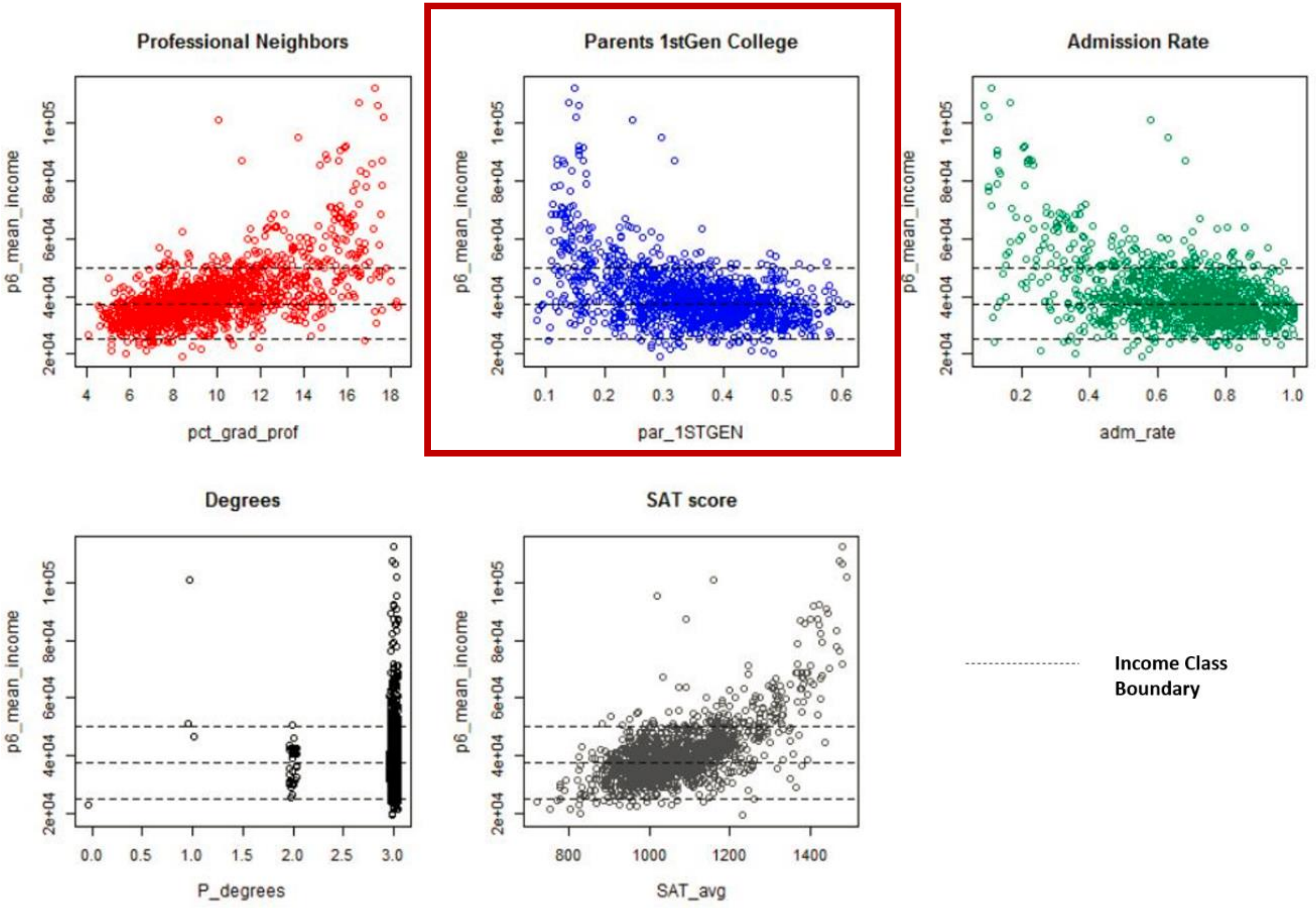
Discussion


Professional Neighbors
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
4
6
8
10
12
14
16
18
pct_grad_prof
Parents 1stGen College
p6_mean_income
0.1
0.2
0.3
0.4
0.5
0.6
par_1STGEN
Admission Rate
p6_mean_income
0.2
0.4
0.6
0.8
1.0
adm_rate
Degrees
p6_mean_income
0.0
0.5
1.0
1.5
2.0
2.5
3.0
P_degrees
SAT score
p6_mean_income
800
1000
1200
1400
SAT_avg
Income Class Boundary

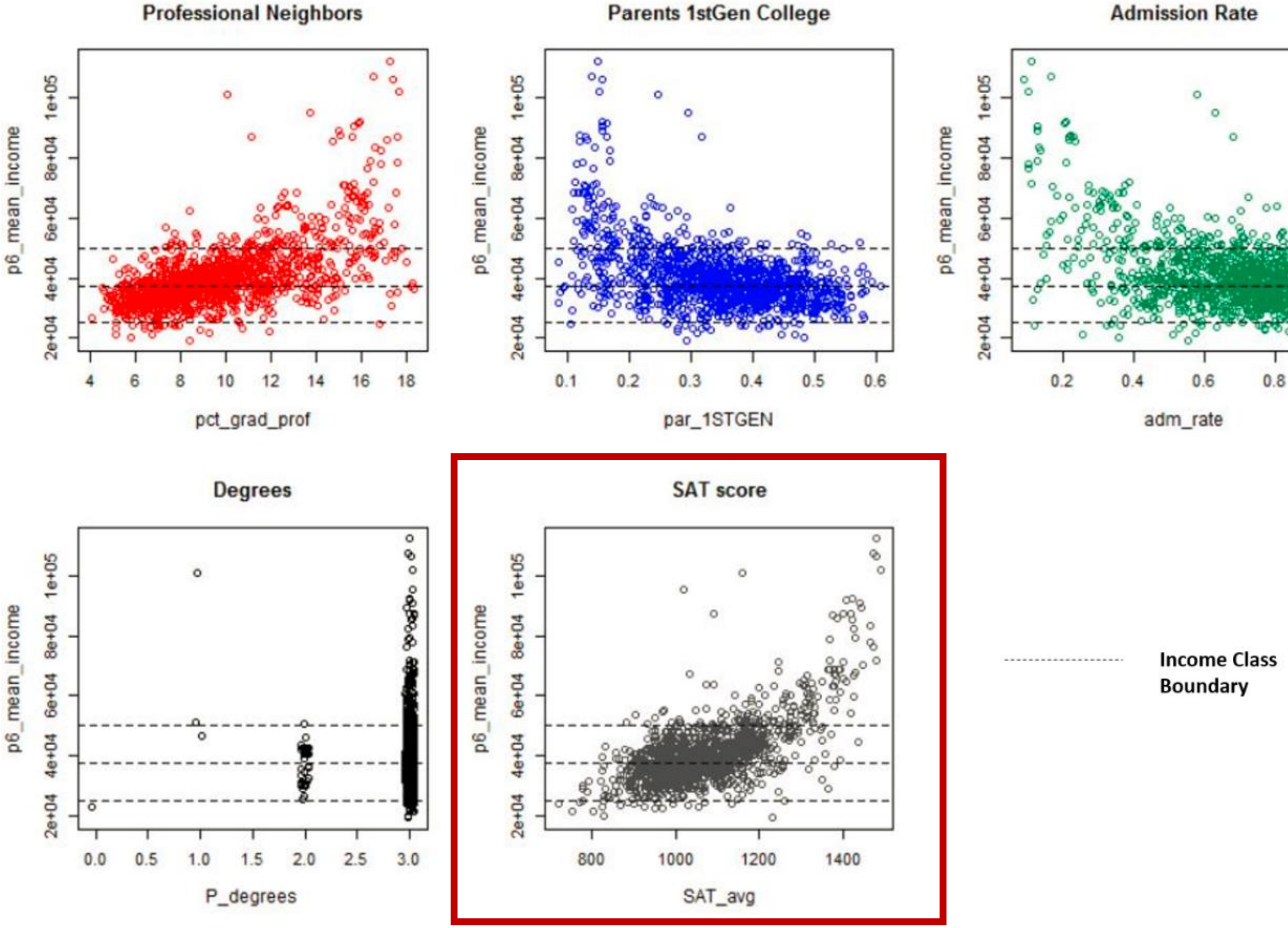
Discussion


Professional Neighbors
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
4 6 8 10 12 14 16 18
pct_grad_prof
Parents 1stGen College
p6_mean_income
0.1 0.2 0.3 0.4 0.5 0.6
par_1STGEN
Admission Rate
p6_mean_income
0.2 0.4 0.6 0.8 1.0
adm_rate
Degrees
p6_mean_income
0.0 0.5 1.0 1.5 2.0 2.5 3.0
P_degrees
SAT score
p6_mean_income
800 1000 1200 1400
SAT_avg
Income Class Boundary

Discussion

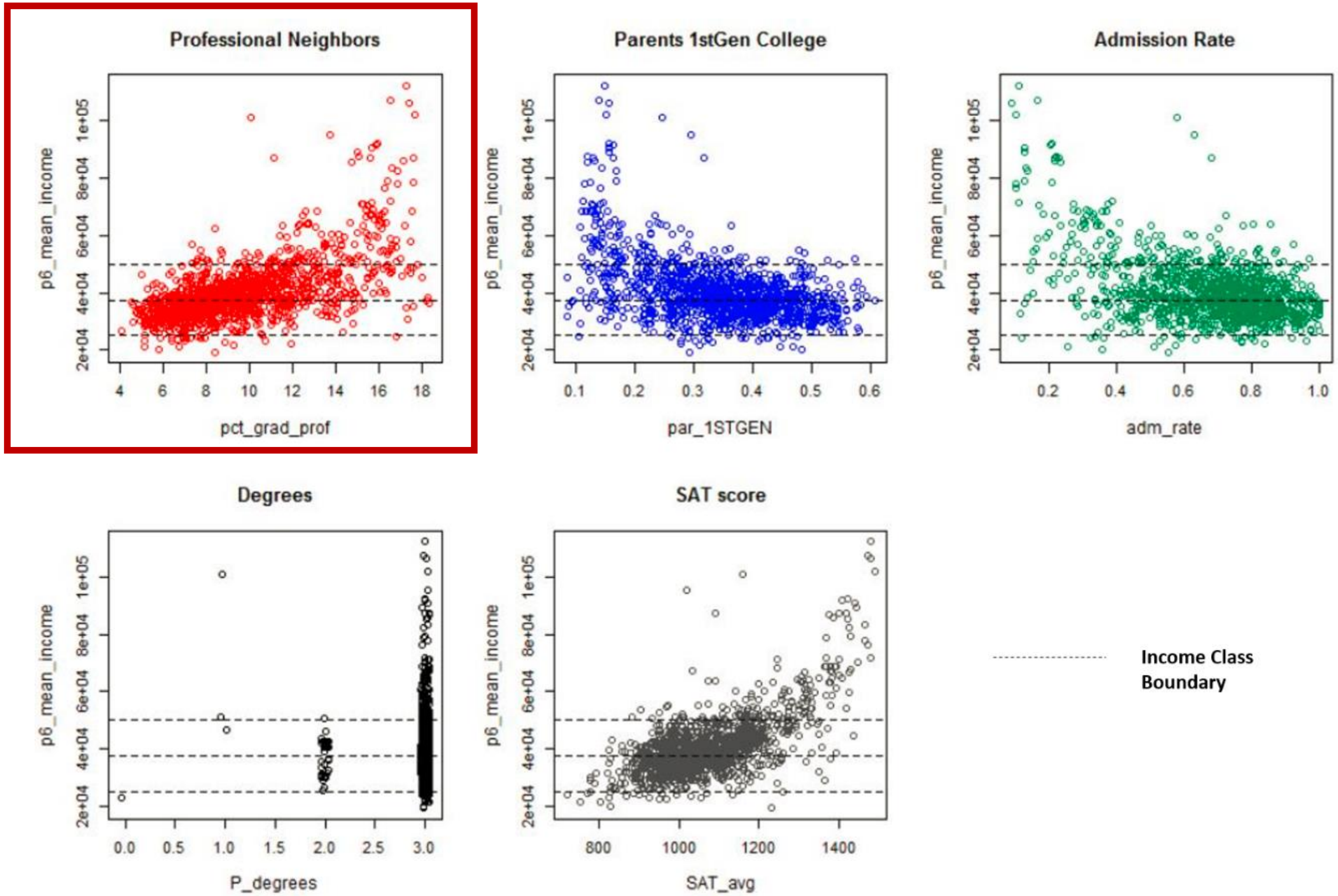

Professional Neighbors
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
4
6
8
10
12
14
16
18
pct_grad_prof
Parents 1stGen College
p6_mean_income
0.1
0.2
0.3
0.4
0.5
0.6
par_1STGEN
Admission Rate
p6_mean_income
0.2
0.4
0.6
0.8
1.0
adm_rate
Degrees
p6_mean_income
0.0
0.5
1.0
1.5
2.0
2.5
3.0
P_degrees
SAT score
p6_mean_income
800
1000
1200
1400
SAT_avg
Income Class
Boundary

Discussion



Professional Neighbors
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
4
6
8
10
12
14
16
18
pct_grad_prof
Parents 1stGen College
p6_mean_income
0.1
0.2
0.3
0.4
0.5
0.6
par_1STGEN
Admission Rate
p6_mean_income
0.2
0.4
0.6
0.8
1.0
adm_rate
Degrees
p6_mean_income
0.0
0.5
1.0
1.5
2.0
2.5
3.0
P_degrees
SAT score
p6_mean_income
800
1000
1200
1400
SAT_avg
Income Class Boundary

Discussion


Professional Neighbors
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
4
6
8
10
12
14
16
18
pct_grad_prof
Parents 1stGen College
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
0.1
0.2
0.3
0.4
0.5
0.6
par_1STGEN
Admission Rate
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
0.2
0.4
0.6
0.8
1.0
adm_rate
Degrees
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
0.0
0.5
1.0
1.5
2.0
2.5
3.0
P_degrees
SAT score
p6_mean_income
2e+04
4e+04
6e+04
8e+04
1e+05
800
1000
1200
1400
SAT_avg
Income Class Boundary

Thanks!