Weight Prediction : Flight Delays

Team Mint Squad

2011004534 윤덕진
2012004407 권규혁
2013012405 김현중
2013012676 이병곤

CONTENTS

01. Introduction

Dataset Description/ Our Goal

02. Data Analysis

Insight / Variables description

03. Pre-processing

Preparation of Data / Feature manufacture

04. Algorithms and Model

Decision Tree / k-NN /SVM /
Random Forest / Naïve Bayes

05. Algorithms Result

Python & R all features / top 6 features

06. Conclusion

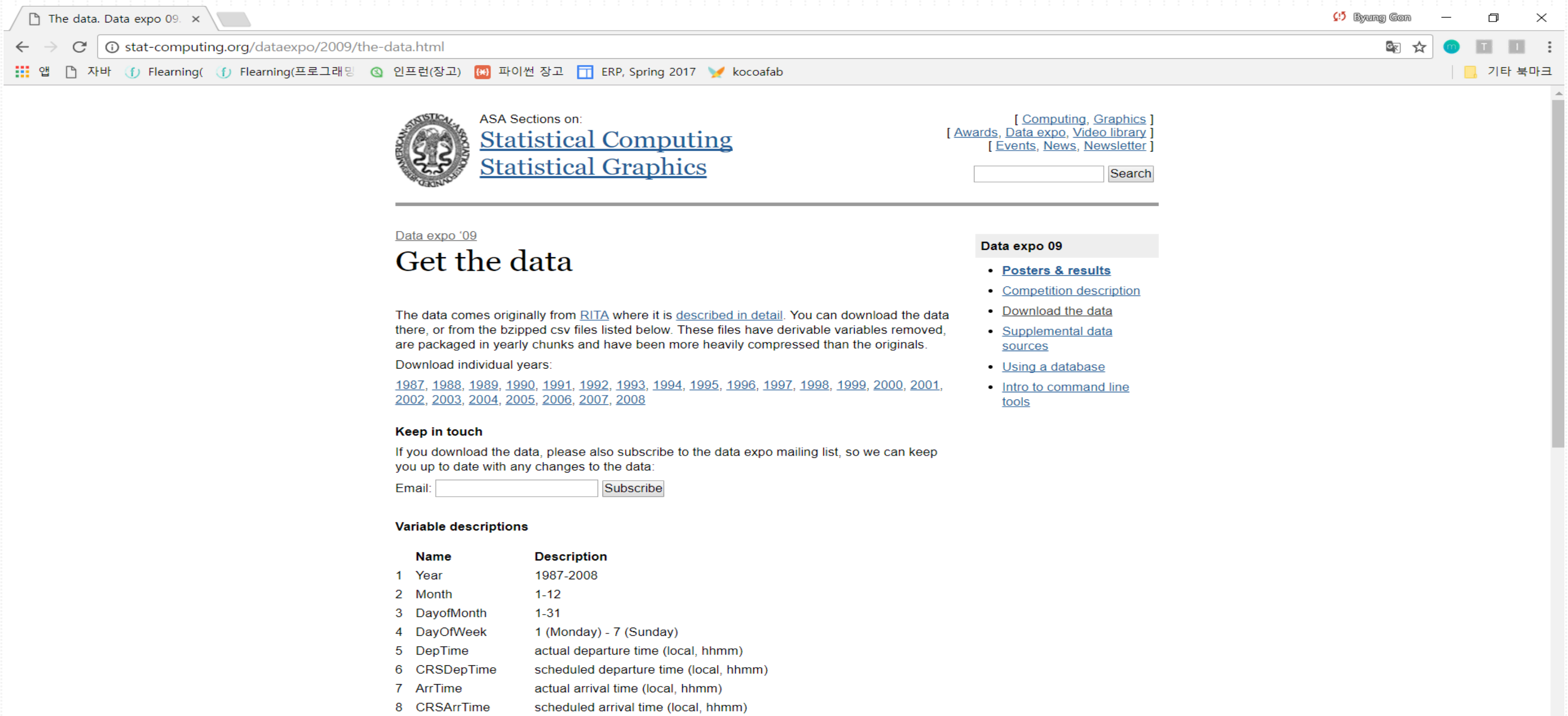
Outro

PART 1

Introduction

1) Dataset of Flight Delays

Dataset provided by Bureau of Transportation Statistics



The data. Data expo 09

stat-computing.org/dataexpo/2009/the-data.html

ASA Sections on:
Statistical Computing
Statistical Graphics

[Computing, Graphics]
[Awards, Data expo, Video library]
[Events, News, Newsletter]

Search

Data expo '09

Get the data

The data comes originally from RITA where it is described in detail. You can download the data there, or from the bzipped csv files listed below. These files have derivable variables removed, are packaged in yearly chunks and have been more heavily compressed than the originals.

Download individual years:

1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008

Keep in touch

If you download the data, please also subscribe to the data expo mailing list, so we can keep you up to date with any changes to the data:

Email: Subscribe

Variable descriptions

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)

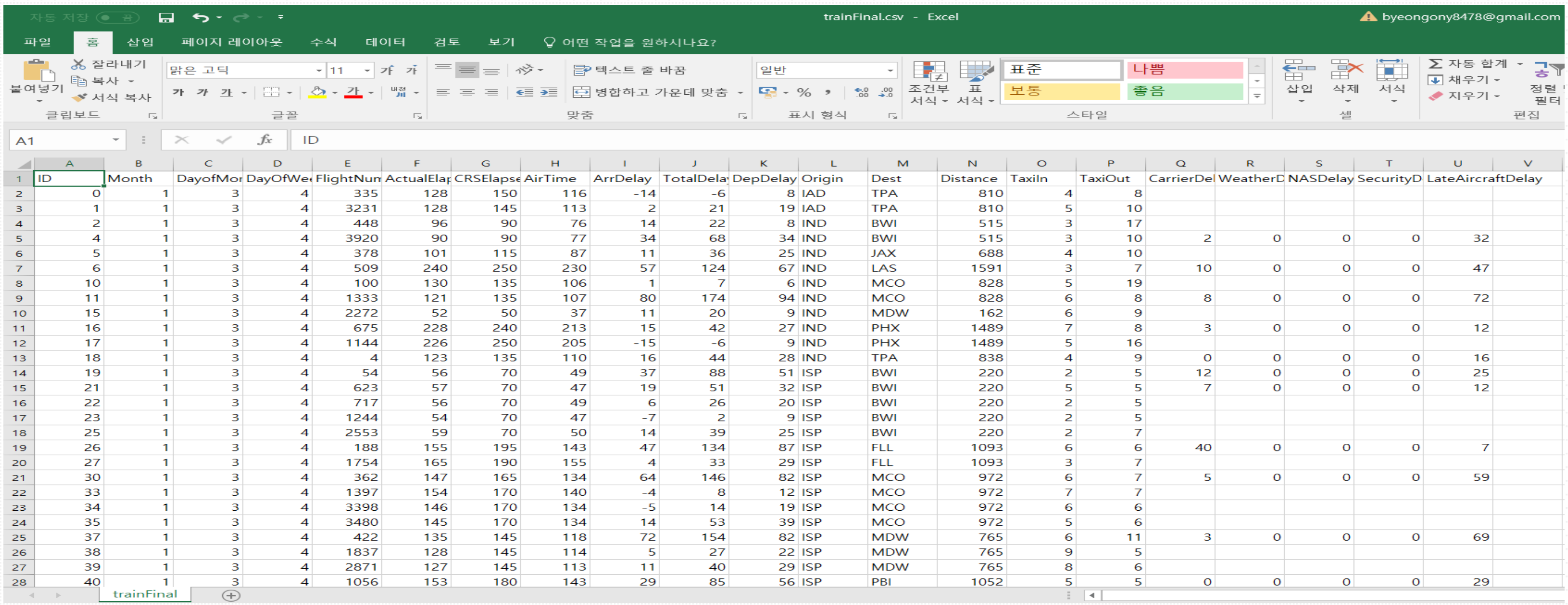
Data expo 09

- Posters & results
- Competition description
- Download the data
- Supplemental data sources
- Using a database
- Intro to command line tools

1) Dataset of Flight Delays

Dataset of Training - trainingFinal.csv

30000 Objects, including Total Delay column



	ID	Month	DayofMor	DayOfWe	FlightNun	ActualElap	CRSElapse	AirTime	ArrDelay	TotalDela	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	CarrierDel	WeatherD	NASDelay	SecurityD	LateAircraftDelay
2	0	1	3	4	335	128	150	116	-14	-6	8	IAD	TPA	810	4	8					
3	1	1	3	4	3231	128	145	113	2	21	19	IAD	TPA	810	5	10					
4	2	1	3	4	448	96	90	76	14	22	8	IND	BWI	515	3	17					
5	4	1	3	4	3920	90	90	77	34	68	34	IND	BWI	515	3	10	2	0	0	0	32
6	5	1	3	4	378	101	115	87	11	36	25	IND	JAX	688	4	10					
7	6	1	3	4	509	240	250	230	57	124	67	IND	LAS	1591	3	7	10	0	0	0	47
8	10	1	3	4	100	130	135	106	1	7	6	IND	MCO	828	5	19					
9	11	1	3	4	1333	121	135	107	80	174	94	IND	MCO	828	6	8	8	0	0	0	72
10	15	1	3	4	2272	52	50	37	11	20	9	IND	MDW	162	6	9					
11	16	1	3	4	675	228	240	213	15	42	27	IND	PHX	1489	7	8	3	0	0	0	12
12	17	1	3	4	1144	226	250	205	-15	-6	9	IND	PHX	1489	5	16					
13	18	1	3	4	4	123	135	110	16	44	28	IND	TPA	838	4	9	0	0	0	0	16
14	19	1	3	4	54	56	70	49	37	88	51	ISP	BWI	220	2	5	12	0	0	0	25
15	21	1	3	4	623	57	70	47	19	51	32	ISP	BWI	220	5	5	7	0	0	0	12
16	22	1	3	4	717	56	70	49	6	26	20	ISP	BWI	220	2	5					
17	23	1	3	4	1244	54	70	47	-7	2	9	ISP	BWI	220	2	5					
18	25	1	3	4	2553	59	70	50	14	39	25	ISP	BWI	220	2	7					
19	26	1	3	4	188	155	195	143	47	134	87	ISP	FLL	1093	6	6	40	0	0	0	7
20	27	1	3	4	1754	165	190	155	4	33	29	ISP	FLL	1093	3	7					
21	30	1	3	4	362	147	165	134	64	146	82	ISP	MCO	972	6	7	5	0	0	0	59
22	33	1	3	4	1397	154	170	140	-4	8	12	ISP	MCO	972	7	7					
23	34	1	3	4	3398	146	170	134	-5	14	19	ISP	MCO	972	6	6					
24	35	1	3	4	3480	145	170	134	14	53	39	ISP	MCO	972	5	6					
25	37	1	3	4	422	135	145	118	72	154	82	ISP	MDW	765	6	11	3	0	0	0	69
26	38	1	3	4	1837	128	145	114	5	27	22	ISP	MDW	765	9	5					
27	39	1	3	4	2871	127	145	113	11	40	29	ISP	MDW	765	8	6					
28	40	1	3	4	1056	153	180	143	29	85	56	ISP	PBI	1052	5	5	0	0	0	0	29

1) Dataset of Flight Delays

Dataset of Test - testFinal.csv

300 Objects, excluding Total Delay column

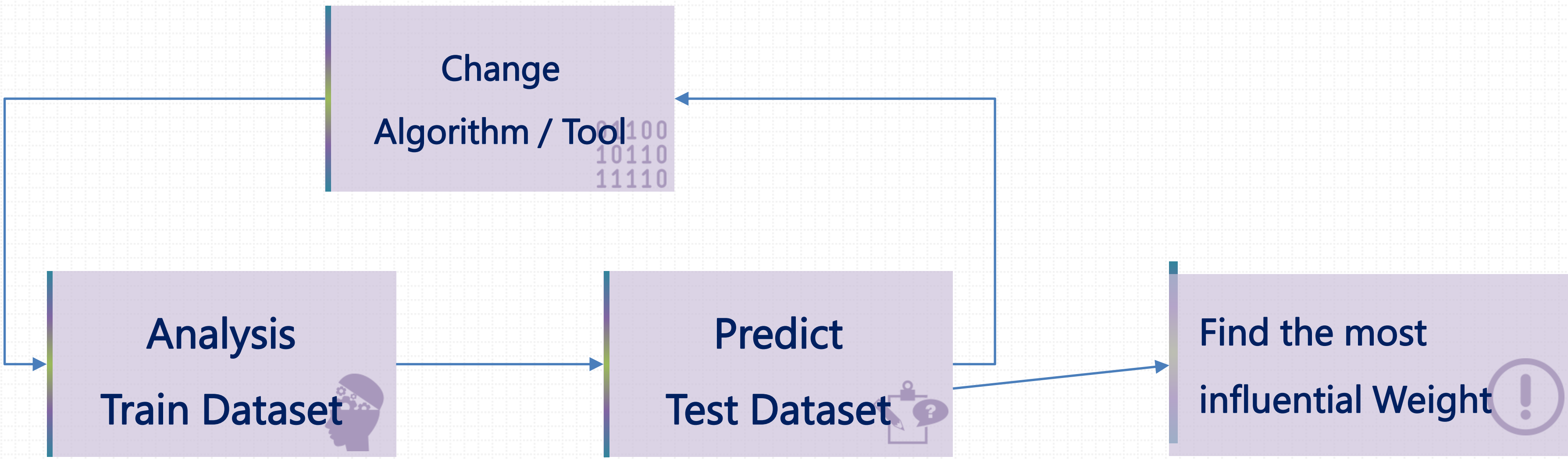
ID	Month	DayofMor	DayOfWe	FlightNun	ActualElap	CRSElapse	AirTime	ArrDelay	TotalDela	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	CarrierDel	WeatherD	NASDelay	SecurityD	LateAircraftDelay
0	1	31	4	1520	224	230	198	192	390	198	MDW	PHX	1444	4	22	0	11	0	0	181
1	1	31	4	1592	252	235	197	85	153	68	MDW	PHX	1444	3	52	19	0	17	0	49
2	1	31	4	3930	310	235	194	194	313	119	MDW	PHX	1444	6	110	9	0	75	0	110
3	1	31	4	207	111	75	60	277	518	241	MDW	PIT	402	8	43	50	0	36	0	191
4	1	31	4	1123	87	75	53	61	110	49	MDW	PIT	402	3	31	49	0	12	0	0
5	1	31	4	1210	191	75	57	267	418	151	MDW	PIT	402	7	127	29	0	116	0	122
6	1	31	4	3069	97	75	55	100	178	78	MDW	PIT	402	4	38	0	56	22	0	22
7	1	31	4	1531	115	125	99	33	76	43	MDW	PVD	842	3	13	7	0	0	0	26
8	1	31	4	1849	112	125	98	-7	-1	6	MDW	PVD	842	4	10					
9	1	31	4	2462	127	125	97	78	154	76	MDW	PVD	842	8	22	0	4	2	0	72
10	1	31	4	3576	139	125	99	85	156	71	MDW	PVD	842	3	37	0	15	14	0	56
11	1	31	4	903	132	105	88	39	51	12	MDW	RDU	632	4	40	12	0	27	0	0
12	1	31	4	2360	133	100	89	144	255	111	MDW	RDU	632	4	40	0	62	33	0	49
13	1	31	4	3187	119	110	96	106	203	97	MDW	RDU	632	3	20	0	16	9	0	81
14	1	31	4	3836	273	255	240	50	82	32	MDW	RNO	1680	6	27	12	0	18	0	20
15	1	31	4	597	234	165	151	334	599	265	MDW	RSW	1105	3	80	0	33	69	0	232
16	1	31	4	620	182	170	153	24	36	12	MDW	RSW	1105	4	25	2	0	12	0	10
17	1	31	4	1424	277	265	241	144	276	132	MDW	SAN	1728	2	34	128	0	12	0	4
18	1	31	4	1766	258	265	230	6	19	13	MDW	SAN	1728	4	24					
19	1	31	4	1777	336	265	232	315	559	244	MDW	SAN	1728	3	101	0	21	71	0	223
20	1	31	4	2113	277	265	234	84	156	72	MDW	SAN	1728	2	41	0	32	12	0	40
21	1	31	4	626	179	175	144	188	372	184	MDW	SAT	1036	2	33	0	20	4	0	164
22	1	31	4	259	96	65	51	83	135	52	MDW	SDF	271	5	40	4	0	31	0	48
23	1	31	4	1111	70	65	51	277	549	272	MDW	SDF	271	6	13	0	46	5	0	226
24	1	31	4	1176	92	65	51	66	105	39	MDW	SDF	271	5	36	6	0	27	0	33
25	1	31	4	1471	75	65	55	392	774	382	MDW	SDF	271	4	16	118	0	10	0	264
26	1	31	4	1108	285	270	250	58	101	43	MDW	SEA	1733	4	31	0	0	15	0	43

1) Dataset of Flight Delays

Variables :

- DayOf Month
- Flight Number
- CRS Elapsed Time
- Actual Elapsed Time
- Distance
- TaxiIn
- TaxiOut
- Carrier Delay
- Airtime
- NAS Delay
- Security Delay
- Late AirCraft Delay
- Weather Delay

2) Our Goal



Our Goal

Predict total delay of flights without weather conditions.

PART 2

Data Analysis

1) Dataset of Flight Delays

Variables :

- DayOf Month
- Flight Number
- CRS Elapsed Time
- Actual Elapsed Time
- Distance
- TaxiIn
- TaxiOut
- Carrier Delay
- Airtime
- NAS Delay
- Security Delay
- Late AirCraft Delay
- Weather Delay

2) Variables Description

DayOfMonth – the impacts by the day of month.

Flight Number – The influence of different types of airplanes.

CRS/Actual Elapsed Time – Difference between scheduled time and actual departure time.

Distance – Flight distance.

TaxiIn/Out – Time to get on and off the plane.

Carrier Time – Time to find luggage.

Air Time – Flight time.

2) Variables Description

NAS Delay – Time it takes in National Air System delay

Security Delay – Time it takes to perform the security check.

Late Aircraft Delay – time it take to maintain an airplane.

Weather Delay – Time delayed due to weather.

PART 3

Pre-Processing

1) Variables Format

DayOfMonth	1-12	**CarrierDelay**	In minutes
FlightNumber	Factor	**AirTime**	In minutes
CRSElapsedTime	In minutes	**NASDelay**	In minutes
ActualElapsedTime	In minutes	**SecurityDelay**	In minutes
Distance	In miles	**LateAirCraftDelay**	In minutes
TaxiIn/Out	Taxi in time, in minutes	**WeatherDelay**	In minutes

1) Variables Format

```
# Train Attributes 20
train_data$ID <- as.character(train_data$ID)
train_data$FlightNum <- as.factor(train_data$FlightNum)
train_data$DayofMonth <- as.factor(train_data$DayofMonth)

# Test Attributes 19
test_data$ID <- as.character(test_data$ID)
test_data$FlightNum <- as.factor(test_data$FlightNum)
test_data$DayofMonth <- as.factor(test_data$DayofMonth)
```

Description

We set non-numeric values in categories.

2) Incomplete dataset

```
> sum(is.na(train_data$DayOfWeek))
[1] 0
> sum(is.na(train_data$FlightNum))
[1] 0
> sum(is.na(train_data$CRSElapsedTime))
[1] 0
> sum(is.na(train_data$ActualElapsedTime))
[1] 78
> sum(is.na(train_data$Distance))
[1] 0
> sum(is.na(train_data$TaxiIn))
[1] 78
> sum(is.na(train_data$TaxiOut))
[1] 0
> sum(is.na(train_data$CarrierDelay))
[1] 13554
> sum(is.na(train_data$AirTime))
[1] 78
> sum(is.na(train_data$NASDelay))
[1] 13554
> sum(is.na(train_data$SecurityDelay))
[1] 13554
> sum(is.na(train_data$LateAircraftDelay))
```

Train.csv

```
> sum(is.na(test_data$DayOfWeek))
[1] 0
> sum(is.na(test_data$FlightNum))
[1] 0
> sum(is.na(test_data$CRSElapsedTime))
[1] 0
> sum(is.na(test_data$ActualElapsedTime))
[1] 3
> sum(is.na(test_data$Distance))
[1] 0
> sum(is.na(test_data$TaxiIn))
[1] 3
> sum(is.na(test_data$TaxiOut))
[1] 0
> sum(is.na(test_data$CarrierDelay))
[1] 62
> sum(is.na(test_data$AirTime))
[1] 3
> sum(is.na(test_data$NASDelay))
[1] 62
> sum(is.na(test_data$SecurityDelay))
[1] 62
> sum(is.na(test_data$LateAircraftDelay))
```

Test.csv

Problem

We have found that NULL values exist.

3) complete dataset

```
# Null of ActualElapsedTime
train_data$ActualElapsedTime[is.na(train_data$ActualElapsedTime)] <- median(train_data$ActualElapsedTime, na.rm=TRUE)
test_data$ActualElapsedTime[is.na(test_data$ActualElapsedTime)] <- median(test_data$ActualElapsedTime, na.rm=TRUE)

# Null of AirTime
train_data$AirTime[is.na(train_data$AirTime)] <- median(train_data$AirTime, na.rm=TRUE)
test_data$AirTime[is.na(test_data$AirTime)] <- median(test_data$AirTime, na.rm=TRUE)

#Null of TaxiIn
train_data$TaxiIn[is.na(train_data$TaxiIn)] <- median(train_data$TaxiIn, na.rm=TRUE)
test_data$TaxiIn[is.na(test_data$TaxiIn)] <- median(test_data$TaxiIn, na.rm=TRUE)
```

```
# Train 5가지 Delay 빈값 = 0
train_data$CarrierDelay[is.na(train_data$CarrierDelay)] <- 0
train_data$WeatherDelay[is.na(train_data$WeatherDelay)] <- 0
train_data$NASDelay[is.na(train_data$NASDelay)] <- 0
train_data$SecurityDelay[is.na(train_data$SecurityDelay)] <- 0
train_data$LateAircraftDelay[is.na(train_data$LateAircraftDelay)] <- 0

# Test 5가지 Delay 빈값 = 0
test_data$CarrierDelay[is.na(test_data$CarrierDelay)] <- 0
test_data$WeatherDelay[is.na(test_data$WeatherDelay)] <- 0
test_data$NASDelay[is.na(test_data$NASDelay)] <- 0
test_data$SecurityDelay[is.na(test_data$SecurityDelay)] <- 0
test_data$LateAircraftDelay[is.na(test_data$LateAircraftDelay)] <- 0
```

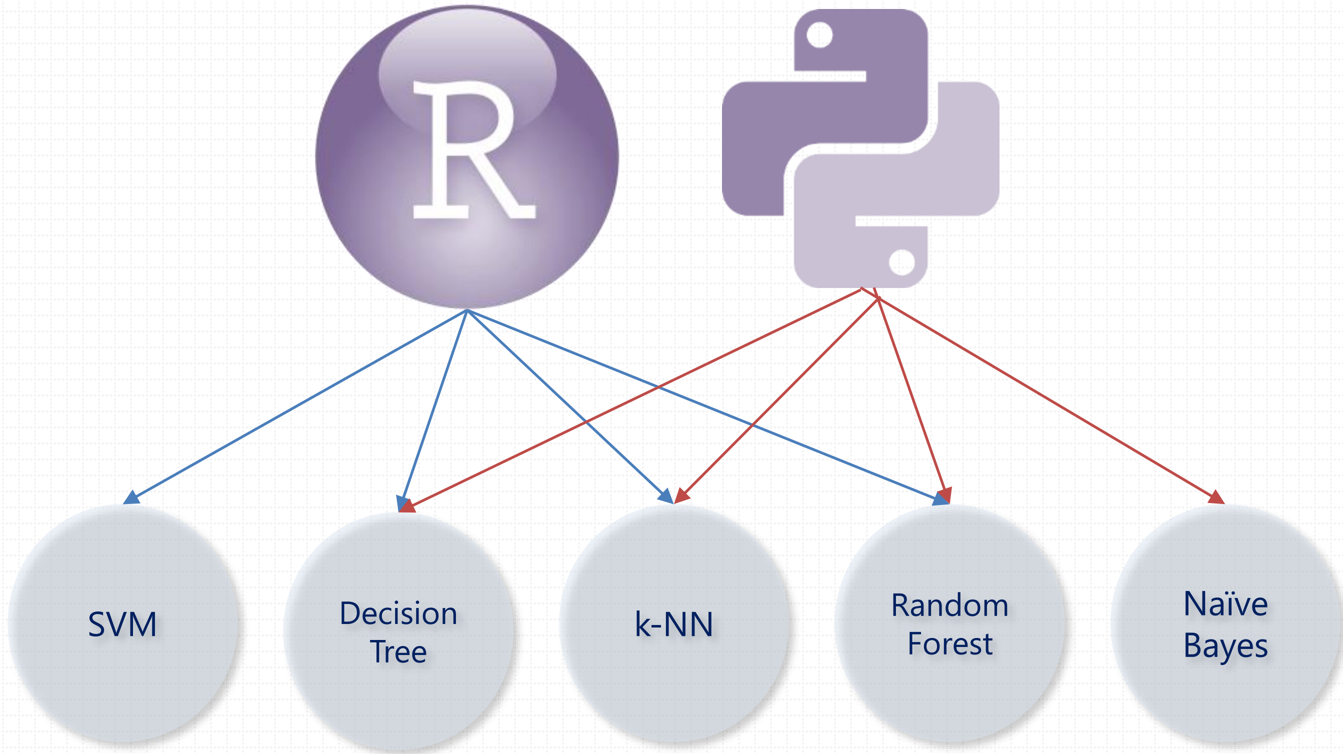
Solution

We have replaced Null values with average values.

PART 4

Algorithm and Models

1) 2 Tools & 5 Algorithms



2) Decision Tree

[Decision Tree Classifier]

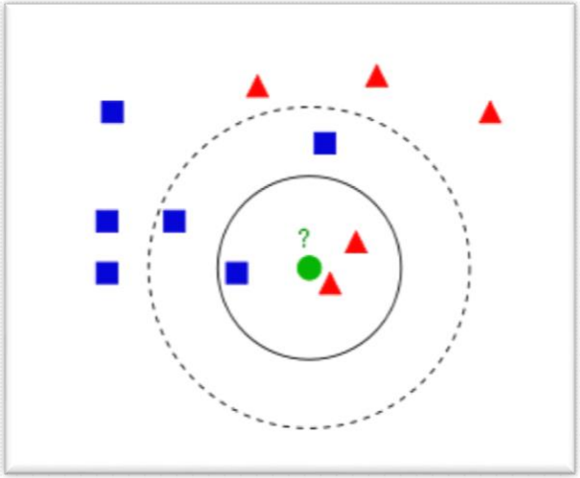
Decision tree analysis is the most popular decision support technique in data mining field. From the root, input training set is divided recursively at split point of input features(attributes) making a tree. The benefit of using decision tree is simple understanding of classification process.

Python	R
Library : sklearn.tree.DecisionTreeClassifier	Library : library(rpart)
Accuracy Result : **79.57%**	Accuracy Result : **73.81%**

3) k-NN (k-Nearest Neighbors)

[k-NN(k-nearest neighbors)]

kNN is a prediction method for classification as well as regression type prediction problems. Its one of the simplest machine-learning algorithm, that test case is simply assigned to the class that most k nearest training cases are found. The result get different depending on k value you set.



Python

Library : sklearn.neighbors.KNeighborsClassifier

Accuracy Result : **55.89%**

R

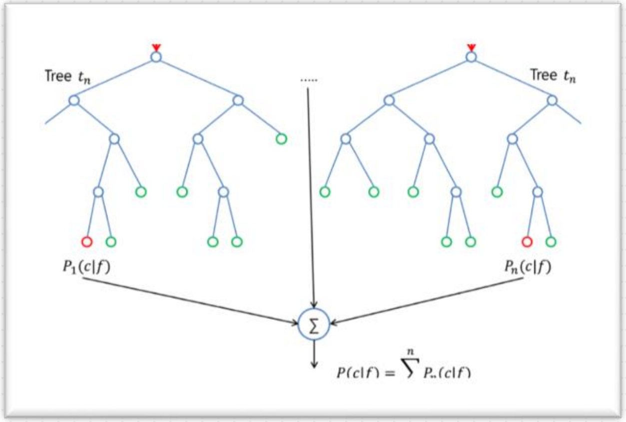
Library : library(kknn)

Accuracy Result : **34.16%**

4) Random Forest

[Random Forest Classifier]

A random forest is a meta classifier that has a number of decision tree classifiers on various sub-samples of the dataset. It uses mode of the classes for classification and mean prediction for regression to improve the predictive accuracy and control over-fitting.



Python

Library : sklearn.ensemble.RandomForestClassifier

Accuracy Result : **73.99%**

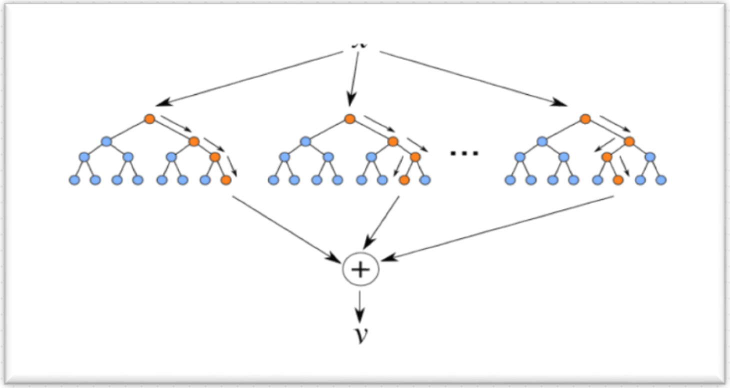
R

Library : library(randomForest)

Accuracy Result : **86.86%**

5) Naïve Bayes

[Naive Bayes Net Classifier]



Naive Bayes Classification is a simple probabilistic classification based on applying Bayes theorem using the naive independence assumptions. In this algorithm, there is an assumption that every features are independent between each other.

Python

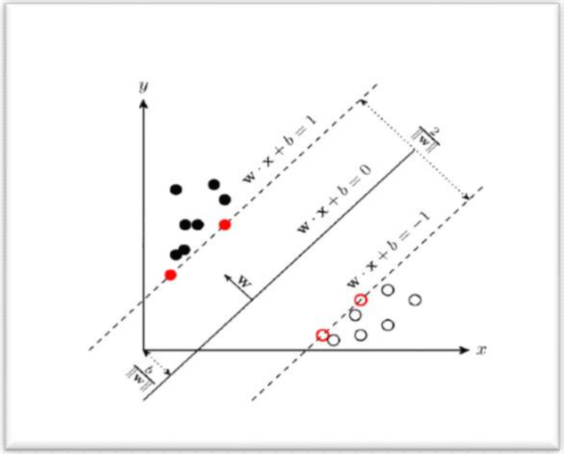
Library : sklearn.naive.bayes.GaussianNB

Accuray Result : **45.94%**

6) SVM (Support Vector Machine)

[Support Vector Machine]

SVMs are supervised learning algorithms that are mostly used for classification and regression. SVMs produce linear classifiers called hyperplane that separate the data into multiple subsections.



R

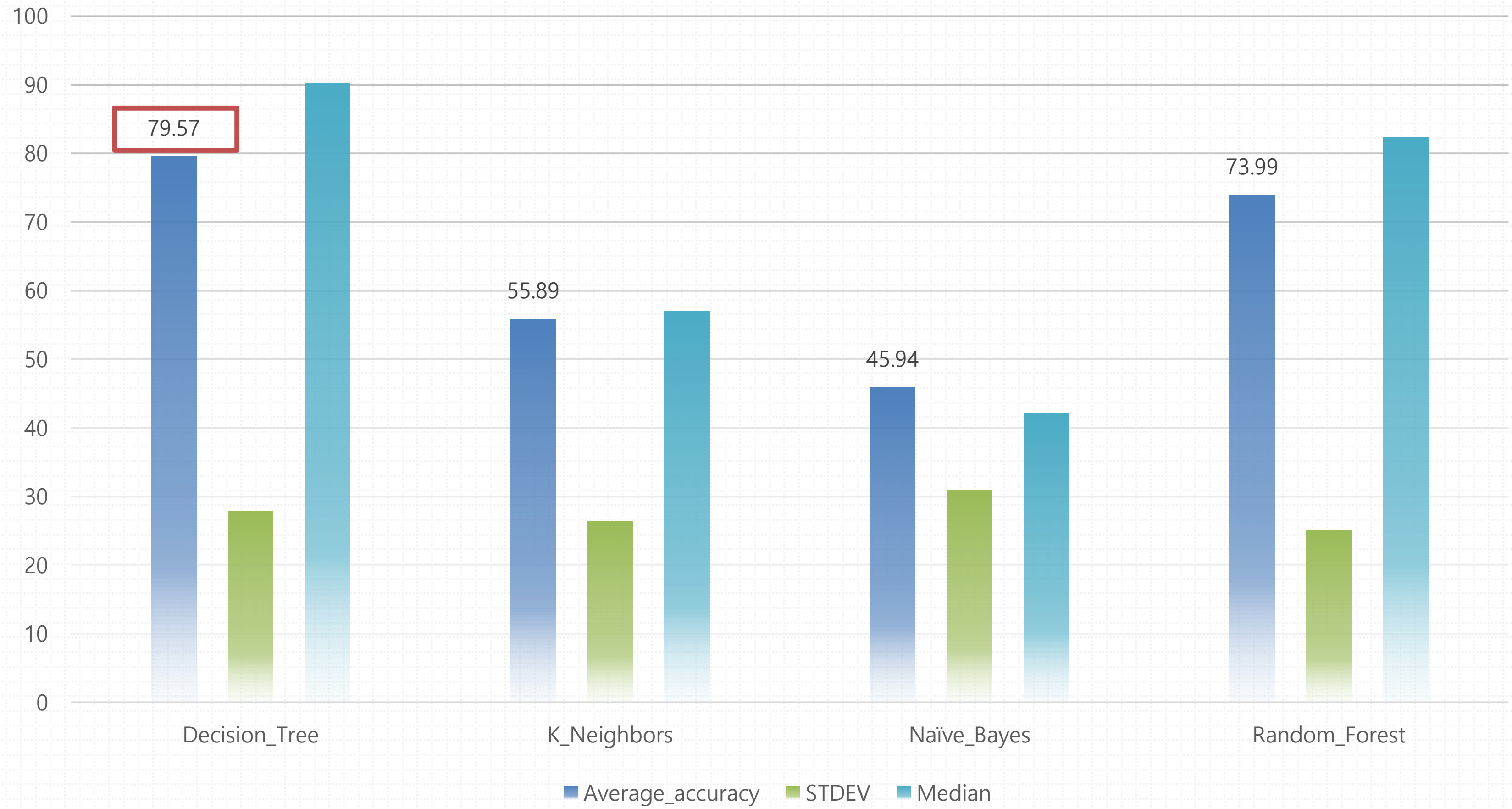
Library : library(kernlab)

Accuray Result : **56.68%**

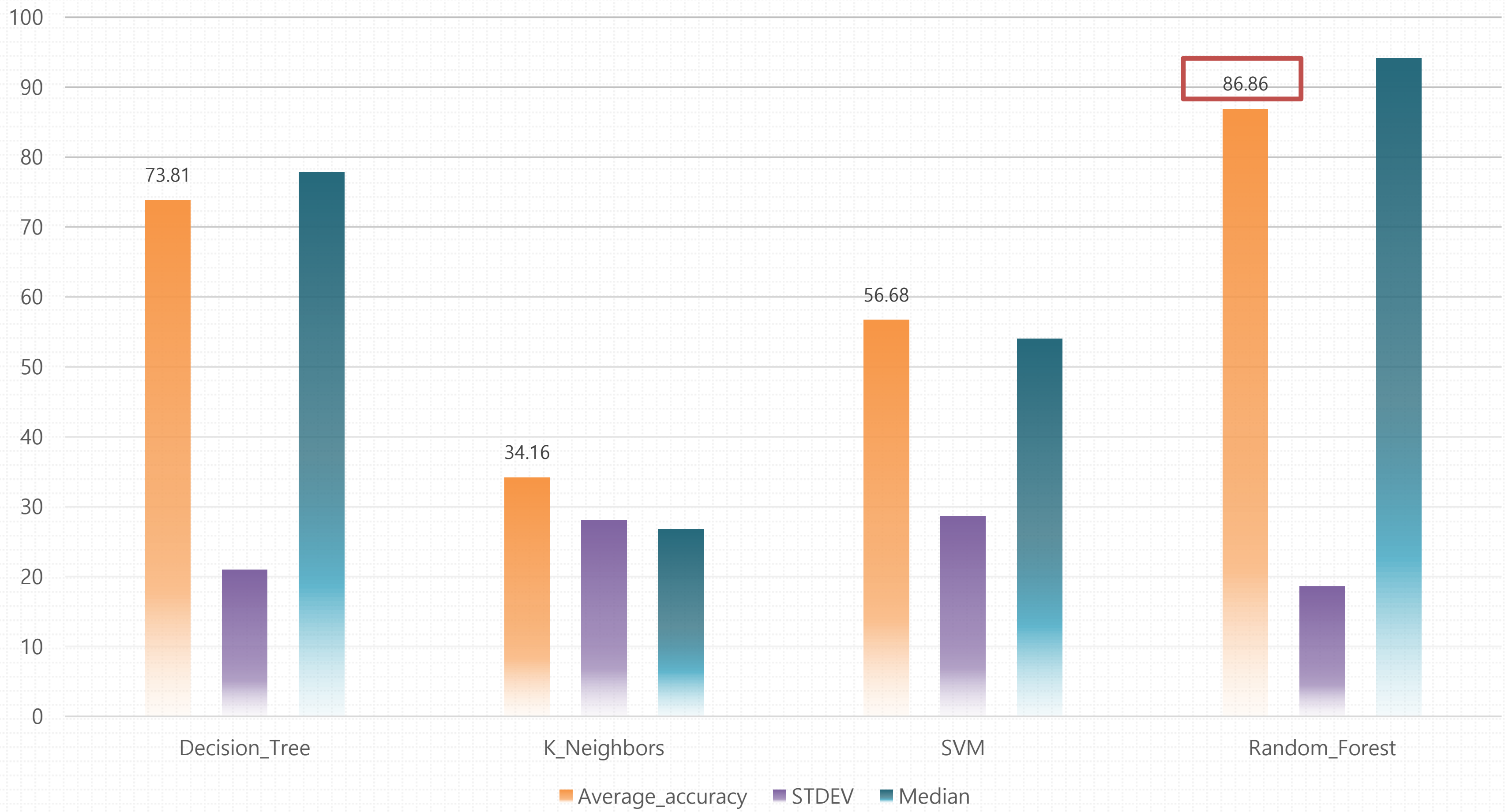
PART 5

Algorithm Results

1) Python


100
90
80
70
60
50
40
30
20
10
0
79.57
55.89
45.94
73.99
Decision_Tree
K_Neighbors
Naïve_Bayes
Random_Forest
Average_accuracy
STDEV
Median

2) R


100
90
80
70
60
50
40
30
20
10
0
73.81
34.16
56.68
86.86
Decision_Tree
K_Neighbors
SVM
Random_Forest
Average_accuracy
STDEV
Median

3) Compare Accuracy Python & R

R	Algorithm	Python
73.81%	Decision Tree	79.57%
86.86%	Random Forest	73.99%
34.16%	K-NN	55.89%
	Naïve Bayes	45.94%
56.58%	SVM	

Result

Algorithms with highest accuracy that fit our dataset are
<u>Decision Tree & RandomForest</u>

4) Select the most 6 influential weight

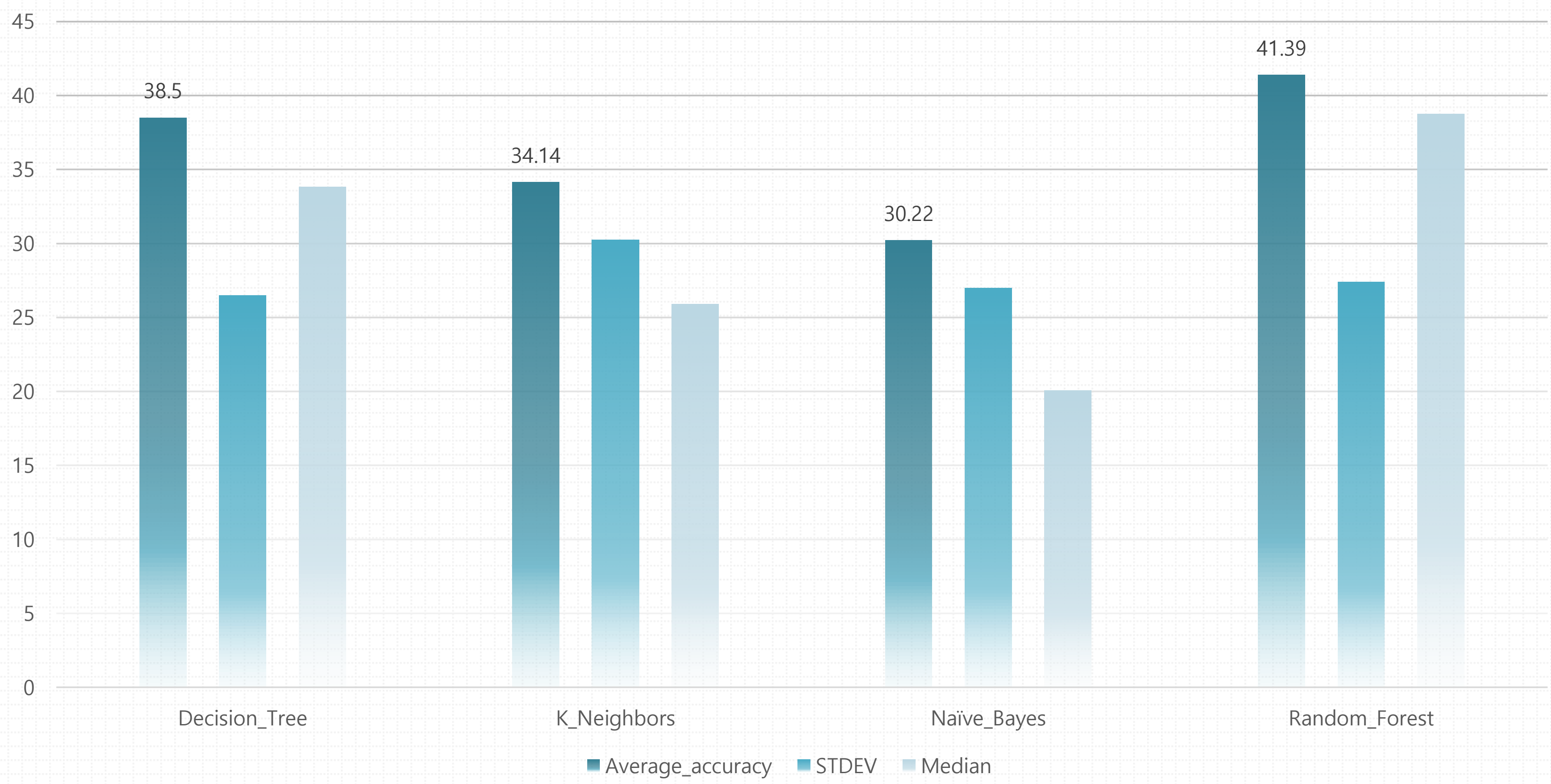
Python

```
  ('Month', 0.0)
2 ('DayofMonth', 0.11247138377205146)
  ('DayOfWeek', 0.07661932841390516l)
1 ('FlightNum', 0.13885770294697292)
  ('ActualElapsedTime', 0.08808191007658024 2)
  ('CRSElapsedTime', 0.05635498763132218 4)
5 ('AirTime', 0.09098313846028803 5)
3 ('Distance', 0.10704811910243951)
  ('TaxiIn', 0.08555248328253611 8)
6 ('TaxiOut', 0.09210559753297162 3)
4 ('CarrierDelay', 0.04789023673503397 4)
  ('WeatherDelay', 0.00707999934066018 93)
  ('NASDelay', 0.04428588918220805 6)
  ('SecurityDelay', 0.00159401500622216 61)
  ('LateAircraftDelay', 0.05107520851680830 8)
  1.0
```

R

```
> fit$variable.importance
 FlightNum DayofMonth    TaxiOut   Distance    AirTime     TaxiIn
42369268.7 16412143.4  1544937.4  1051584.0   737314.7   735131.1
```

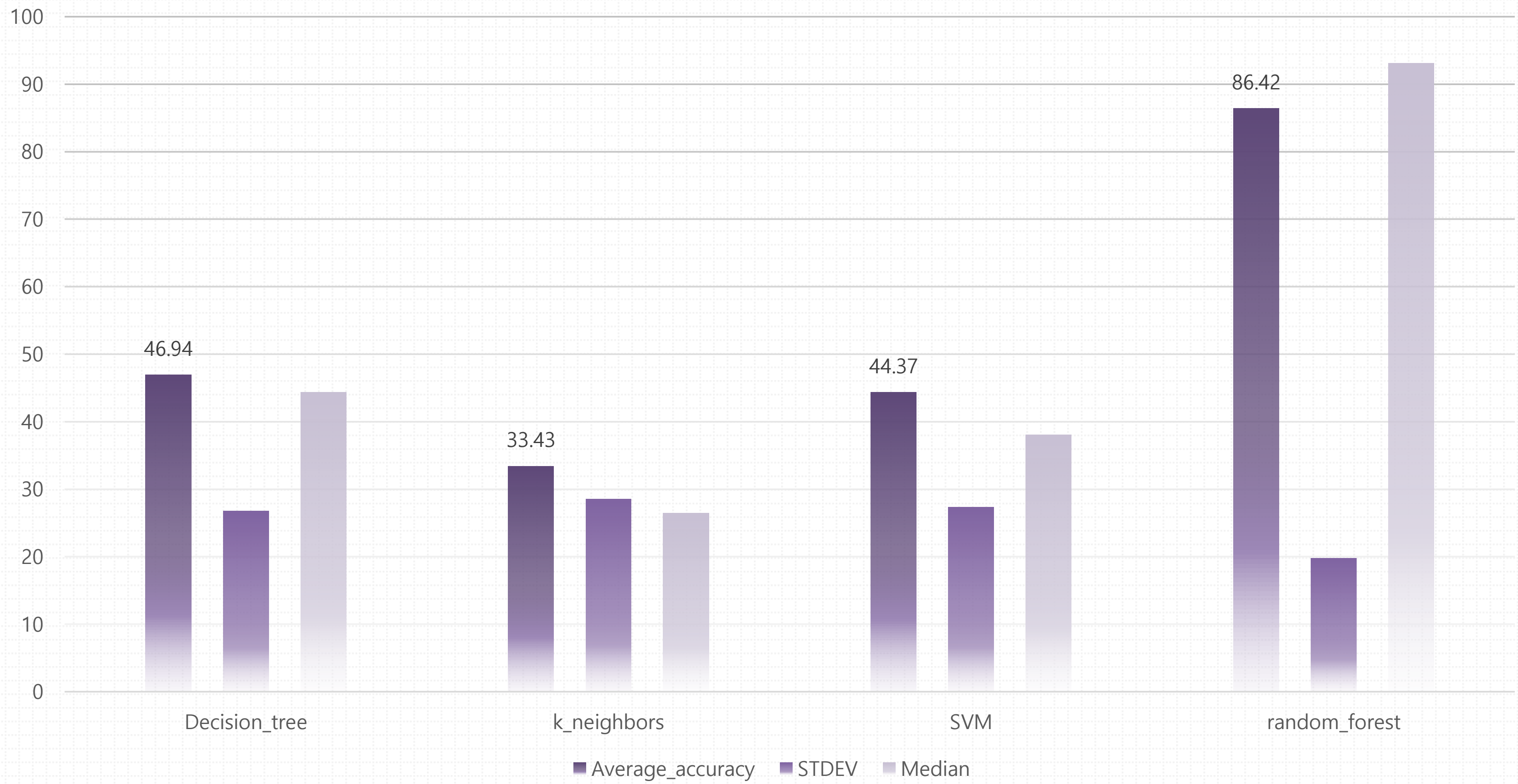
4) Python with Top6 features



5) Compare python with Top 6 & with all features

With all features	Algorithm	With Top 6 features
79.57%	Decision Tree	38.5%
73.99%	Random Forest	34.14%
55.89%	K-NN	30.22%
45.94%	Naïve Bayes	41.49%

6) R with Top6 features



7) Compare R with Top 6 & with all features

With all features	Algorithm	With Top 6 features
73.81%	Decision Tree	46.94%
86.86%	Random Forest	86.42%
34.16%	K-NN	33.43%
56.58%	SVM	44.37%

Result

As a result of selecting and running the top 6 variables, the accuracy was lower than all the other variables.

8) Conclusion

1. After comparing the accuracy using several algorithms, **Decision Tree and RandomForest fit our dataset the best.**

2. As a result of checking the weights of all variables through the above two algorithms, **we confirmed that the top 6 variables (FlightNumber, DayOfMonths, Distance, TaxiOut, AirTime, TaxiIn) affect 65% of the total delay.**

3. As a result of estimating the delay with only the top 6 variables, **we confirmed that considering less variables leads to a considerable loss of accuracy.**

Our Conclusion

After having reviewed the results of our project, we have concluded that considering more variables for the machine increases accuracy. Furthermore, the dataset itself is not appropriate for machine learning. So the more concrete data is needed to our study.


감사합니다.