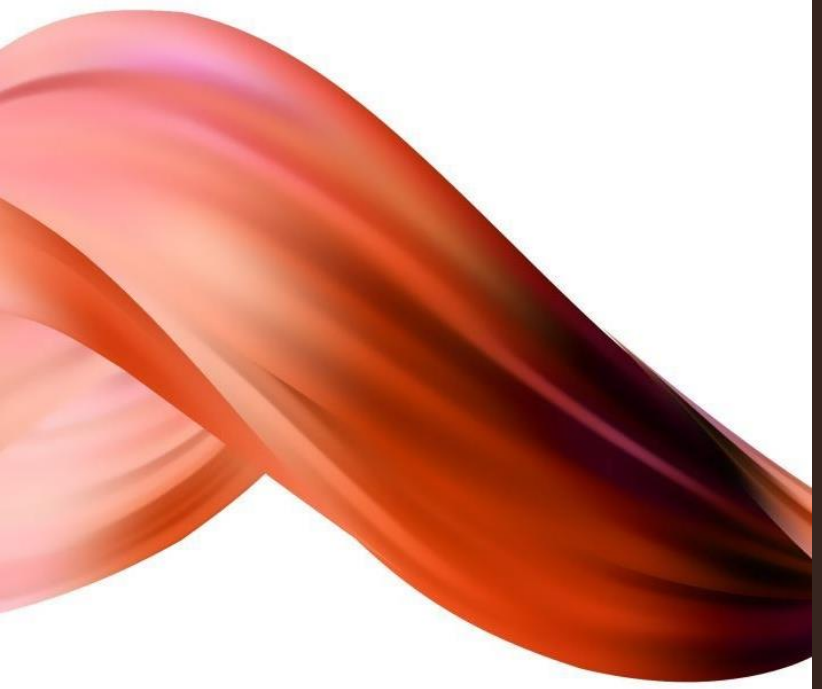


AWS PROJECT PHASE 2

Phase 2

OMKAR VARTAK

YEAR 1987

SELECTING RANDOM 1000 RECORDS FROM YEAR 1987

In [1]:
```
import pandas as pd
```

In [11]:
```
df=pd.read_csv('C:/Users/omkar/Desktop/Project Files/1987 (4).csv.bz2')
df
```

Out[11]:

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	...	TaxiIn	TaxiOut	Cancelled	(
0	1987	10	14	3	741.0	730	912.0	849	PS	1451	...	NaN	NaN	0	
1	1987	10	15	4	729.0	730	903.0	849	PS	1451	...	NaN	NaN	0	
2	1987	10	17	6	741.0	730	918.0	849	PS	1451	...	NaN	NaN	0	
3	1987	10	18	7	729.0	730	847.0	849	PS	1451	...	NaN	NaN	0	
4	1987	10	19	1	749.0	730	922.0	849	PS	1451	...	NaN	NaN	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1311821	1987	12	11	5	1530.0	1530	1825.0	1823	CO	638	...	NaN	NaN	0	
1311822	1987	12	13	7	1530.0	1530	1815.0	1823	CO	638	...	NaN	NaN	0	
1311823	1987	12	14	1	1530.0	1530	1807.0	1823	CO	638	...	NaN	NaN	0	
1311824	1987	12	1	2	1525.0	1525	1643.0	1638	CO	639	...	NaN	NaN	0	
1311825	1987	12	2	3	1540.0	1525	1706.0	1638	CO	639	...	NaN	NaN	0	

1311826 rows × 29 columns

Adding the Delay Column, And if the record has a value of zero or less in the ArrDelay and DepDelay updating columns by Inserting a “N” in the Delay column Otherwise update it With "Y".

In [4]:
```
delay = []
for i in range(len(df)):
    if df.loc[i,'ArrDelay'] <= 0 and df.loc[i,'DepDelay'] <= 0:
        delay.append('N')
    else :
        delay.append('Y')
```

In [5]:
```
df['delay'] = delay
df
```

Out[5]:

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	...	TaxiOut	Cancelled	Cancella
0	1987	10	14	3	741.0	730	912.0	849	PS	1451	...	NaN	0	
1	1987	10	15	4	729.0	730	903.0	849	PS	1451	...	NaN	0	
2	1987	10	17	6	741.0	730	918.0	849	PS	1451	...	NaN	0	
3	1987	10	18	7	729.0	730	847.0	849	PS	1451	...	NaN	0	
4	1987	10	19	1	749.0	730	922.0	849	PS	1451	...	NaN	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1311821	1987	12	11	5	1530.0	1530	1825.0	1823	CO	638	...	NaN	0	
1311822	1987	12	13	7	1530.0	1530	1815.0	1823	CO	638	...	NaN	0	
1311823	1987	12	14	1	1530.0	1530	1807.0	1823	CO	638	...	NaN	0	
1311824	1987	12	1	2	1525.0	1525	1643.0	1638	CO	639	...	NaN	0	
1311825	1987	12	2	3	1540.0	1525	1706.0	1638	CO	639	...	NaN	0	

1311826 rows × 30 columns

In [6]:
```
df[['ArrDelay', 'DepDelay', 'delay']]
```

Out[6]:

	ArrDelay	DepDelay	delay
0	23.0	11.0	Y
1	14.0	-1.0	Y
2	29.0	11.0	Y
3	-2.0	-1.0	N
4	33.0	19.0	Y
...	...	...	...
1311821	2.0	0.0	Y
1311822	-8.0	0.0	N
1311823	-16.0	0.0	N
1311824	5.0	0.0	Y
1311825	28.0	15.0	Y

1311826 rows × 3 columns

In [8]:
```
df_unknown = df[df.TailNum != 'UNKNOW']
```

In [10]:
```
Omkar_sample = df_unknown.sample(1000)
```

In [11]:
```
Omkar_sample
```

Out[11]:

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	...	TaxiOut	Cancelled
567923	1987	11	17	2	738.0	730	1024.0	1005	HP	1	...	NaN	0
1268783	1987	12	18	5	1820.0	1820	1927.0	1923	US	212	...	NaN	0
51085	1987	10	22	4	1506.0	1508	2005.0	2016	UA	344	...	NaN	0
1208455	1987	12	5	6	1429.0	1430	1831.0	1806	AA	167	...	NaN	0
672815	1987	11	29	7	800.0	800	850.0	908	CO	1110	...	NaN	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1258263	1987	12	21	1	710.0	645	920.0	847	US	14	...	NaN	0
1086032	1987	12	17	4	1413.0	1410	1454.0	1445	PI	807	...	NaN	0
946933	1987	12	13	7	1646.0	1647	1931.0	1927	UA	888	...	NaN	0
1125551	1987	12	18	5	2039.0	2013	2203.0	2124	DL	300	...	NaN	0
893141	1987	12	7	1	1145.0	1128	1352.0	1328	TW	245	...	NaN	0

SUCCESSFULLY IMPORTING THE UPDATED FILE TO THE SYSTEM.

SAVING TO THE LOCAL SYSTEM.

1311826 rows × 3 columns

In [21]:
```
df_unknown = df[df.TailNum != 'UNKNOW']
```

In [22]:
```
Omkar_sample = df_unknown.sample(1000)
```

In [23]:
```
Omkar_sample
```

Out[23]:

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	...	TaxiOut	Cancelled	Cancella
954219	1987	12	31	4	1303.0	1305	1401.0	1357	UA	1135	...	NaN	0	
877387	1987	12	23	3	843.0	845	947.0	941	PS	1652	...	NaN	0	
496629	1987	11	5	4	1326.0	1330	1505.0	1504	UA	319	...	NaN	0	
1000162	1987	12	15	2	NaN	1240	NaN	1310	HP	250	...	NaN	1	
243486	1987	10	8	4	2030.0	2025	2142.0	2154	CO	1272	...	NaN	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
245367	1987	10	16	5	1635.0	1635	1655.0	1656	CO	1631	...	NaN	0	
531189	1987	11	8	7	1214.0	1215	1326.0	1322	UA	1427	...	NaN	0	
150779	1987	10	23	5	930.0	930	1107.0	1100	NW	370	...	NaN	0	
21	1987	10	8	4	932.0	915	1033.0	1001	PS	1451	...	NaN	0	
628530	1987	11	14	6	911.0	902	1223.0	1142	PI	107	...	NaN	0	

1000 rows × 30 columns

In [ ]:
```
Omkar_sample.to_csv(r'C::\Users\omkar\Desktop\Project Files\Omkar_sample_1987.csv', index=False, header=True)
```

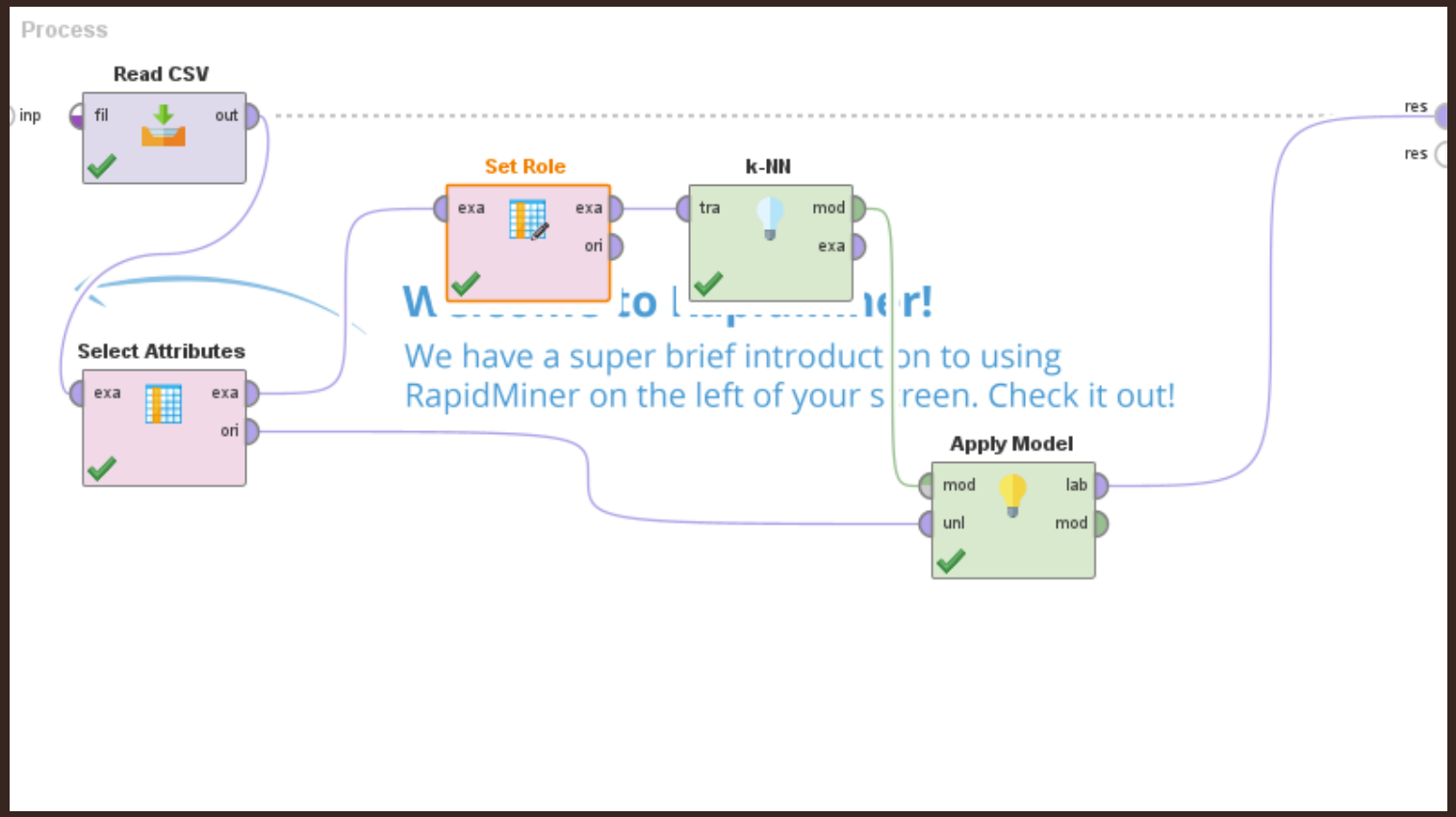
COMBINING THE DATA FROM ALL THE USERS YEAR 1987

A1 | Year

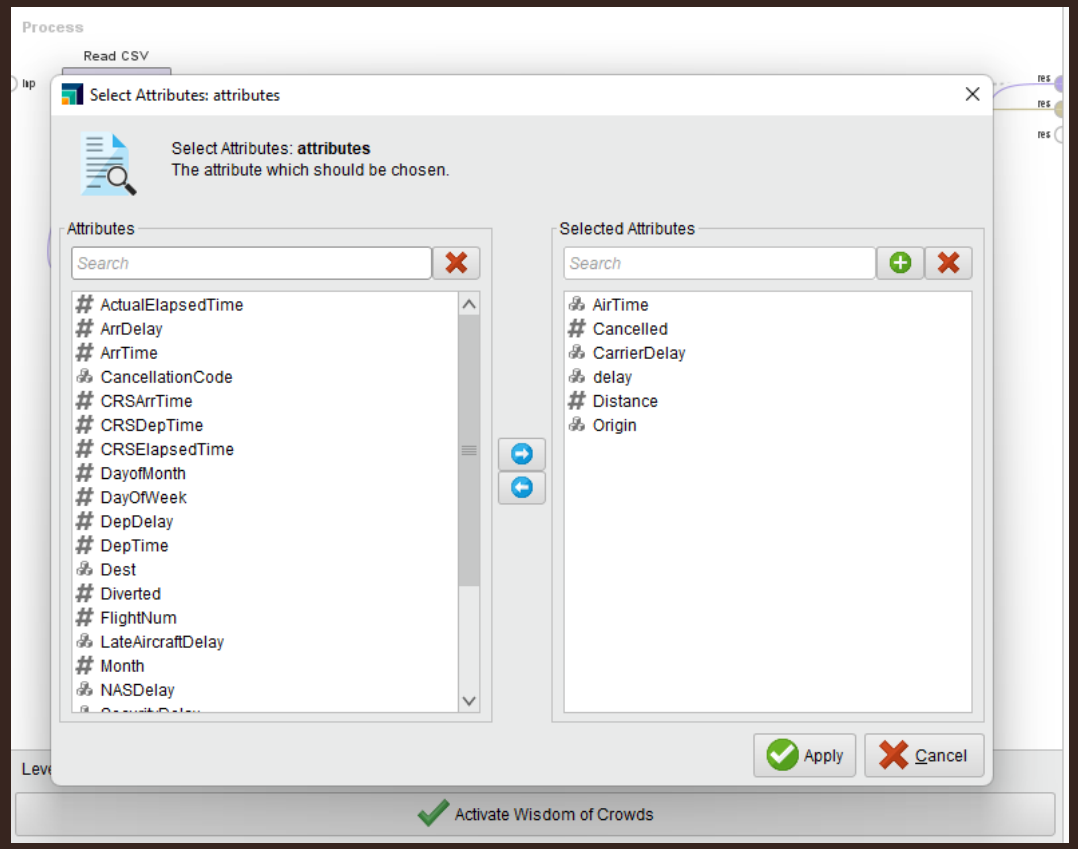
Year	Month	DayofMor	DayOfWe	DepTime	CRSDepTi	ArrTime	CRSArrTin	UniqueCa	FlightNun	TailNum	ActualElap	CRSElapse	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Cancelled	Cancellati	Di
1987	11	17	2	738	730	1024	1005	HP	1		226	215		19	8	ORD	PHX	1440			0		
1987	12	18	5	1820	1820	1927	1923	US	212		67	63		4	0	GSO	PIT	304			0		
1987	10	22	4	1506	1508	2005	2016	UA	344		179	188		-11	-2	DEN	IAD	1452			0		
1987	12	5	6	1429	1430	1831	1806	AA	167		362	336		25	-1	LAX	HNL	2556			0		
1987	11	29	7	800	800	850	908	CO	1110		50	68		-18	0	IAH	MSY	305			0		
1987	10	10	6	1400	1400	1446	1455	WN	31		46	55		-9	0	DAL	HOU	239			0		
1987	12	9	3	1639	1640	1755	1753	UA	730		76	73		2	-1	ABQ	DEN	349			0		
1987	10	14	3	830	830	915	925	WN	10		45	55		-10	0	HOU	DAL	239			0		
1987	11	8	7	1735	1730	1823	1823	CO	525		108	113		0	5	DEN	LAS	629			0		
1987	10	4	7	1029	1029	1100	1102	AA	820		31	33		-2	0	CLT	FAY	118			0		
1987	11	13	5	2015	2015	2135	2124	TW	212		80	69		11	0	STL	ORD	258			0		
1987	10	30	5	1324	1320	1600	1601	AA	456		96	101		-1	4	ORD	SYR	607			0		
1987	10	2	5	856	850	1157	1155	UA	91		361	365		2	6	BOS	LAX	2611			0		
1987	10	9	5	1436	1422	1533	1521	PI	80		57	59		12	14	CLT	ORF	290			0		
1987	11	16	1		700		812	CO	1643			132				AUS	DEN	775			1		
1987	12	2	3	1741	1740	1907	1905	CO	93		146	145		2	1	DEN	SAN	853			0		
1987	11	25	3	2130	2115	2255	2240	NW	212		85	85		15	15	DTW	PHL	453			0		
1987	10	30	5	1215	1215	1500	1444	CO	164		165	149		16	0	FLL	IAD	901			0		
1987	11	10	2	853	850	905	900	HP	17		72	70		5	3	PHX	LAX	370			0		
1987	12	29	2	2007	1955	2143	2131	US	89		96	96		12	12	PHL	DAY	477			0		
1987	11	6	5	1254	1251	1455	1445	AA	1027		121	114		10	3	BOS	RDU	612			0		
1987	12	7	1	1545	1545	1654	1655	PI	817		69	70		-1	0	MIA	JAX	334			0		
1987	11	22	7	2358	2359	456	510	UA	954		178	191		-14	-1	GEG	ORD	1498			0		
1987	11	12	4	800	800	830	829	TW	517		30	29		1	0	CMI	PIA	86			0		
1987	12	19	6	1817	1810	1907	1910	EA	385		110	120		-3	7	ATL	MCI	692			0		
1987	11	8	7	756	750	1544	1539	PI	8		288	289		5	6	SFO	CLT	2296			0		
1987	12	31	4	843	843	955	959	AA	2163		72	76		-4	0	ONT	SJC	333			0		
1987	12	25	5		930		1206	UA	606			96				ORD	DCA	612			1		
1987	12	9	3	1210	1210	1329	1325	AA	2404		79	75		4	0	SMF	LAX	373			0		
1987	11	29	7	1145	1145	1226	1225	AS	203		41	40		1	0	SAN	LAX	109			0		
1987	10	26	1	808	800	903	851	US	47		55	51		12	8	BUF	PIT	186			0		
1987	10	10	6	1923	1905	2019	2005	WN	66		56	60		14	18	BHM	MSY	321			0		

Combined_Data

IMPORTING THE DATA INTO RAPID MINER FOR YEAR 1987



ATTRIBUTES FOR CASE 1



ACCURACY FOR COMBINED CASES

Table View Plot View

accuracy: 51.45% +/- 1.93% (micro average: 51.45%)

	true Y	true N	class precision
pred. Y	1844	1077	63.13%
pred. N	1836	1243	40.37%
class recall pred. N	50.11%	53.58%	

RECALL FOR COMBINED CASES

Table View Plot View

weighted_mean_recall: 51.84% +/- 1.30% (micro average: 51.84%), weights: 1, 1

	true Y	true N	class precision
pred. Y	1844	1077	63.13%
pred. N	1836	1243	40.37%
class recall	50.11%	53.58%	

MEAN FOR COMBINED CASES

Table View Plot View

weighted_mean_precision: 51.77% +/- 1.25% (micro average: 51.75%), weights: 1, 1

	true Y	true N	class precision
pred. Y	1844	1077	63.13%
pred. N	1836	1243	40.37%
class recall	50.11%	53.58%	

THE RESULT THAT WE GOT AFTER EXECUTING

Result History | ExampleSet (Apply Model)

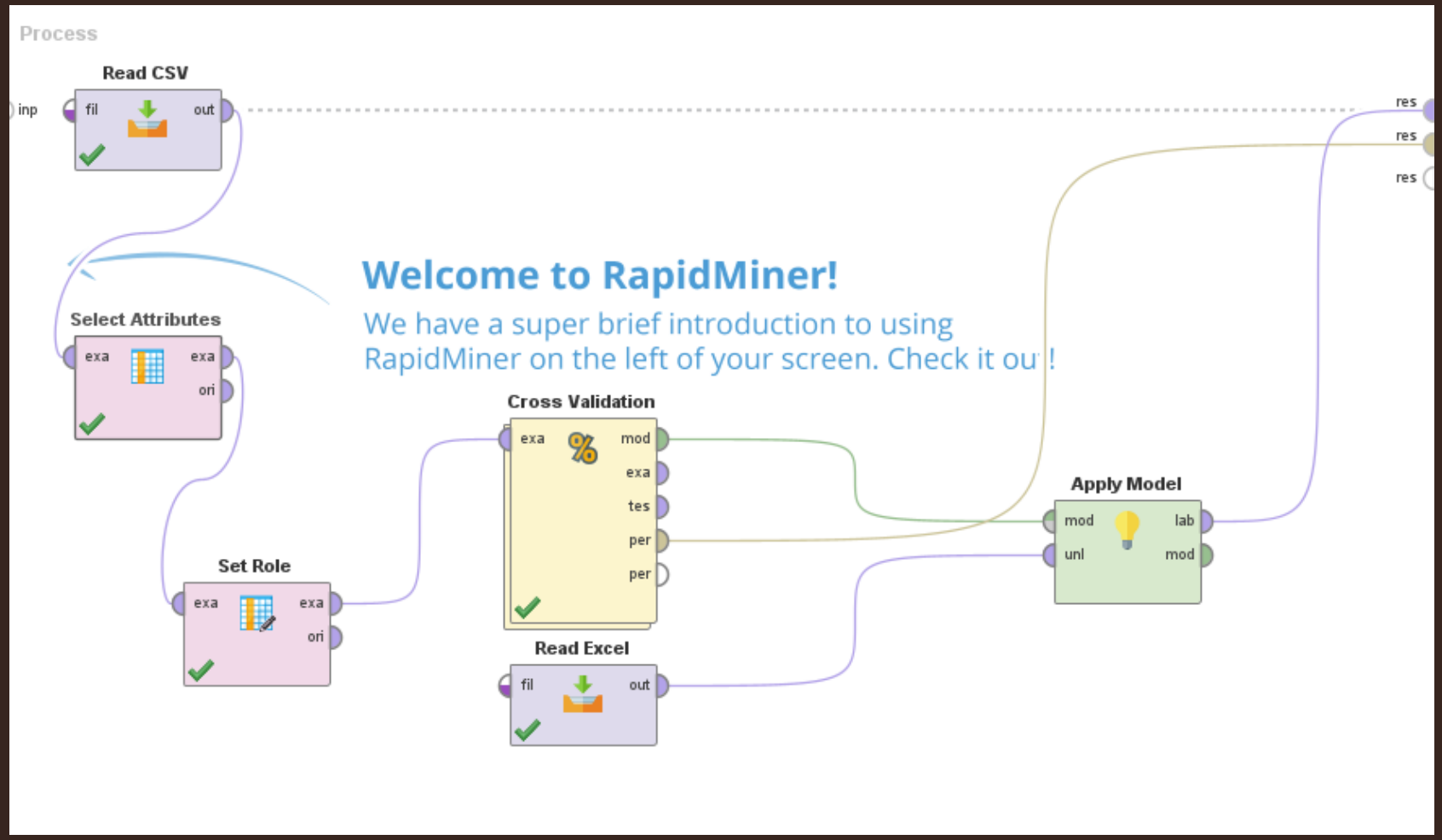
Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model

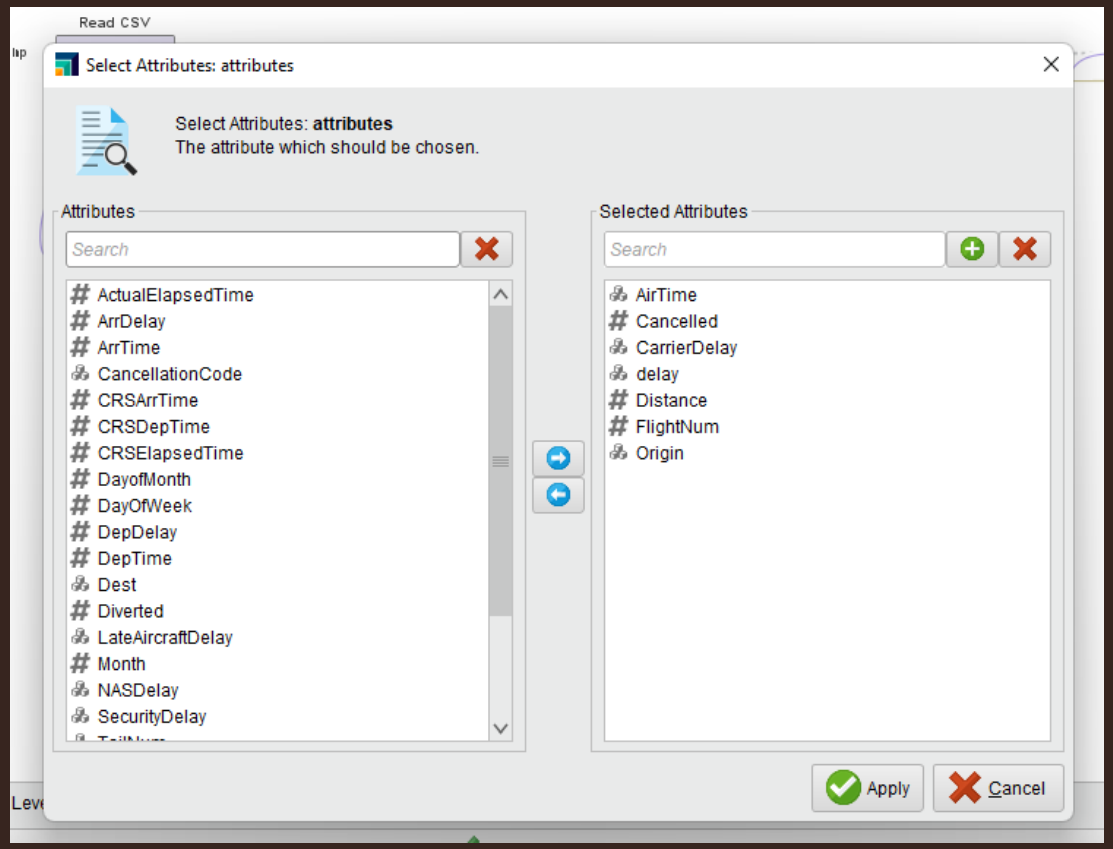
Filter (6,000 / 6,000 examples): all

Row No.	prediction(d...	confidence(Y)	confidence(N)	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime
1	Y	1	0	1987	11	17	2	738	730
2	Y	0.625	0.375	1987	12	18	5	1820	1820
3	N	0.250	0.750	1987	10	22	4	1506	1508
4	Y	0.800	0.200	1987	12	5	6	1429	1430
5	N	0.200	0.800	1987	11	29	7	800	800
6	N	0.400	0.600	1987	10	10	6	1400	1400
7	Y	0.500	0.500	1987	12	9	3	1639	1640
8	N	0.400	0.600	1987	10	14	3	830	830
9	Y	0.600	0.400	1987	11	8	7	1735	1730
10	N	0.375	0.625	1987	10	4	7	1029	1029
11	Y	0.600	0.400	1987	11	13	5	2015	2015
12	Y	1	0	1987	10	30	5	1324	1320
13	Y	0.800	0.200	1987	10	2	5	856	850
14	Y	0.625	0.375	1987	10	9	5	1436	1422
15	Y	0.625	0.375	1987	11	16	1	?	700
16	Y	0.600	0.400	1987	12	2	3	1741	1740
17	N	0.400	0.600	1987	11	25	3	2130	2115
18	Y	0.875	0.125	1987	10	30	5	1215	1215
19	Y	0.800	0.200	1987	11	10	2	853	850
20	N	0.417	0.583	1987	12	29	2	2007	1955
21	Y	0.875	0.125	1987	11	6	5	1254	1251
22	Y	0.750	0.250	1987	12	7	1	1545	1545
23	N	0.167	0.833	1987	11	22	7	2358	2359

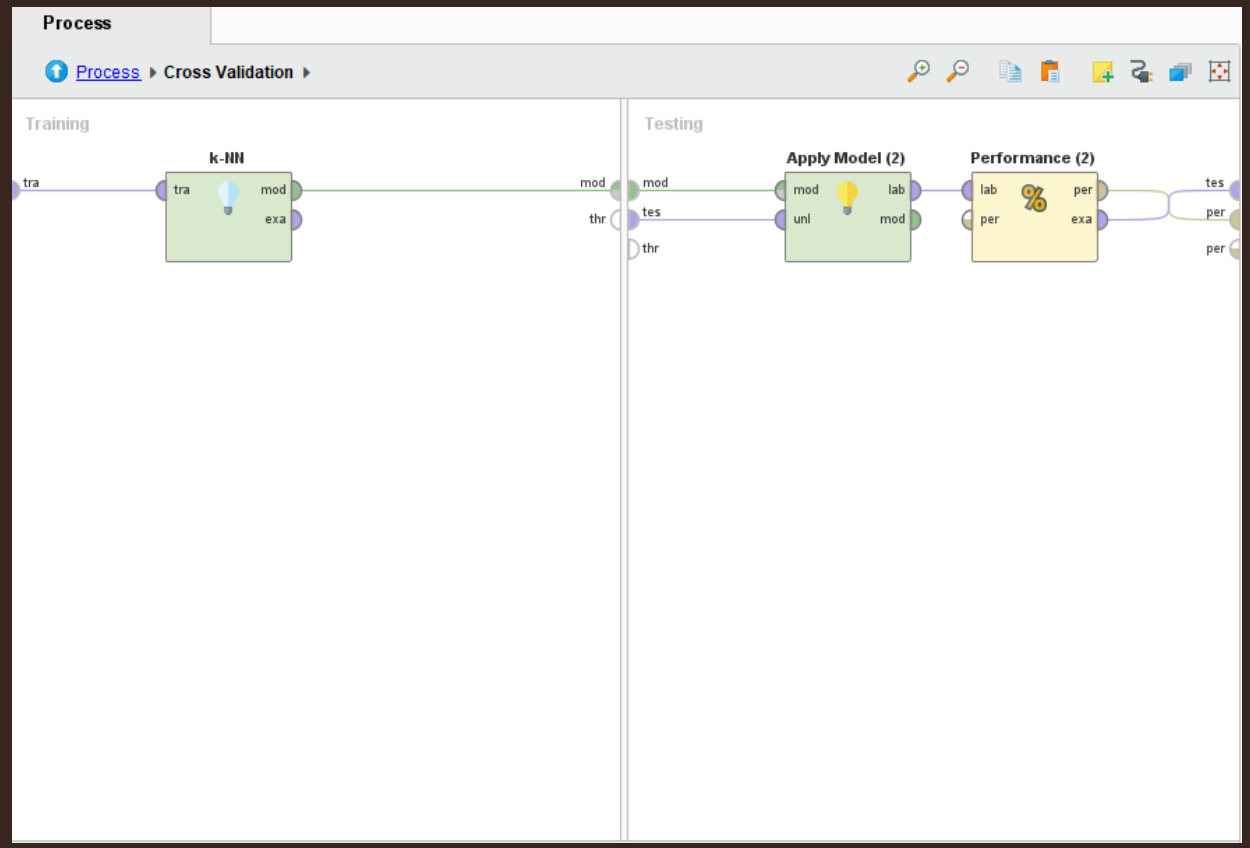
CASE FOR 12 RECORDS.



ATTRIBUTES TAKEN FOR 12 CASE RECORDS



IMPLEMENTATION OF KNN ALGORTIHM YEAR 1987



ACCURACY RESULT FOR 12 CASE RECORDS

Criterion

accuracy

weighted mean recall

weighted mean precisi...

Table View Plot View

accuracy: 57.20% +/- 1.31% (micro average: 57.20%)

	true Y	true N	class precision
pred. Y	2506	1394	64.26%
pred. N	1174	926	44.10%
class recall	68.10%	39.91%	

RECALL RESULT FOR 12 CASE RECORDS

Criterion

accuracy

weighted mean recall

weighted mean precisi...

Table View Plot View

weighted_mean_recall: 54.01% +/- 1.52% (micro average: 54.01%), weights: 1, 1

	true Y	true N	class precision
pred. Y	2506	1394	64.26%
pred. N	1174	926	44.10%
class recall	68.10%	39.91%	

PRECISION RESULT FOR 12 CASE RECORDS

Table View Plot View

weighted_mean_precision: 54.17% +/- 1.56% (micro average: 54.18%), weights: 1, 1

	true Y	true N	class precision
pred. Y	2506	1394	64.26%
pred. N	1174	926	44.10%
class recall	68.10%	39.91%	

OUTPUT FOR USING THE FOLLOWING MODEL YEAR 1987

Open in Turbo Prep Auto Model

Filter (12 / 12 examples): all

Row No.	predictio...	confidence(Y)	confidence(N)	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	A
1	Y	0.609	0.391	?	1	30	6	?	1920	?
4	Y	0.598	0.402	?	4	24	6	?	700	?
5	Y	0.561	0.439	?	5	1	7	?	1205	?
6	Y	0.799	0.201	?	6	7	2	?	945	?
8	Y	1	0	?	8	22	1	?	1750	?
9	Y	0.873	0.127	?	9	12	3	?	1500	?
10	Y	0.800	0.200	?	10	14	7	?	1500	?
11	Y	0.622	0.378	?	11	6	2	?	730	?
2	N	0.378	0.622	?	2	6	6	?	728	?
3	N	0.408	0.592	?	3	17	3	?	0	?
7	N	0.178	0.822	?	7	14	4	?	1455	?
12	N	0.221	0.779	?	12	14	5	?	2030	?