**Developing a Flight Delay Prediction Model using Machine Learning**

**TEAM LEADER** : SWATHI K

**TEAM MEMBER 1** : VINOTHINI S

**TEAM MEMBER 2** : ABITHA A

**TEAM MEMBER 3** : SUMITHRA S

**MODELLING**

def map\_labels(delays):

if delays > 45:

return 1

else:

return 0

df['delayed'] = ((df['dep\_delay'].map(map\_labels) + df['arr\_delay'].map(map\_labels)) != 0).astype(int)

df['delayed'].value\_counts(normalize = True)

0 0.87084

1 0.12916

Name: delayed, dtype: float64

df.head(20)

year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time arr\_delay carrier flight tailnum origin dest air\_time distance hour minute time\_hour delayed

0 2013 1 1 517 515 2 830 819 11 UA 1545 N14228 EWR IAH 227.0 1400 5 15 1/1/2013 5:00 0

1 2013 1 1 533 529 4 850 830 20 UA 1714 N24211 LGA IAH 227.0 1416 5 29 1/1/2013 5:00 0

df.tail(5)

# feature omission

columns\_to\_remove = ['dep\_time', 'sched\_dep\_time', 'dep\_delay', 'arr\_time', 'sched\_arr\_time', 'arr\_delay', 'flight', 'tailnum', 'air\_time', 'distance', 'hour', 'minute', 'time\_hour']

df.drop(columns\_to\_remove, axis = 1, inplace = True)

df['delayed'].value\_counts().to\_frame()

delayed

0 285066

1 42280

df['delayed'].value\_counts().to\_frame()

df['delayed'].value\_counts().to\_frame()

dest

ATL 16837

ORD 16566

LAX 16026

BOS 15022

MCO 13967

CLT 13674

SFO 13173

FLL 11897

MIA 11593

DCA 9111

DTW 9031

DFW 8388

RDU 7770

TPA 7390

DEN 7169

IAH 7085

MSP 6929

PBI 6487

BNA 6084

LAS 5952

SJU 5773

IAD 5383

PHX 4606

BUF 4570

CLE 4394

STL 4142

MDW 4025

SEA 3885

CVG 3725

MSY 3715

RSW 3502

CMH 3326

CHS 2759

PIT 2746

MKE 2709

SAN 2709

JAX 2623

BTV 2510

SLC 2451

AUS 2411

ROC 2358

RIC 2346

PWM 2288

HOU 2083

IND 1981

MCI 1885

SYR 1707

BWI 1687

MEM 1686

PHL 1541

GSO 1492

ORF 1434

DAY 1399

PDX 1342

SRQ 1201

SDF 1104

XNA 992

MHT 932

BQN 888

CAK 842

OMA 817

SNA 812

GSP 790

SAV 749

GRR 728

HNL 701

LGB 661

SAT 659

TYS 578

MSN 556

DSM 523

STT 518

ALB 418

BDL 412

BUR 370

PVD 358

PSE 358

BGR 358

SJC 328

OKC 315

OAK 309

TUL 294

SMF 282

BHM 269

ACK 264

AVL 261

ABQ 254

MVY 210

EGE 207

CRW 134

ILM 107

CAE 106

TVC 95

MYR 58

CHO 46

BZN 35

JAC 21

PSP 18

EYW 17

MTJ 14

HDN 14

SBN 10

ANC 8

LEX 1

#df\_filtered = df[df['dest'].isin(["LEX","TVC","MYR","CHO","BZN","JAC","PSP","EYW","HDN","MTJ","SBN","ANC"])]

#print(df\_filtered.head(15))

#df.drop(df[df['dest'].isin(["LEX","TVC","MYR","CHO","BZN","JAC","PSP","EYW","HDN","MTJ","SBN","ANC"])].index, inplace = True , axis = 0)

#print(df.shape)

#df['delayed'].value\_counts().to\_frame()

saving\_data = df.to\_csv("Data/Processed\_data45.csv", index = False)