This Document is the opportunity to share my solution with you.

The main libraries involved in this tutorial are:

* **Pandas** for data manipulation
* **Matplotlib** and **seaborn** for data visualization
* **Numpy** for multidimensional array computing
* **sklearn** for machine learning and predictive modeling

### **Installation procedure**

A very easy way to install these packages is to download and install the [Conda](http://conda.pydata.org/docs/install/quick.html" \l "os-x-miniconda-install) distribution that encapsulates them all. This distribution is available on all platforms (Windows, Linux and Mac OSX).

**Approach**

Data Loading -> Data completion -> feature Extraction -> Prediction

**Quality Check Performed / Error Found**

Some columns in the dataset were empty (NaN). Median of each variable was used to complete the dataset.

**Data Preprocessing**

Dummy Variable (One Hot Representation) were Created from the variables with less than 10 unique values.

Further StandardScalar from preprocessing module in scikit-learn library was used to preprocess the dataset.

**Model Used**

RandomForest was used to perform the classification in the given problem.

GridSearch was also used to find the best parameters for RandomForest.

Following were the best parameters obtained on the dataset:

Best parameters: {'max\_depth': 7, 'n\_estimators': 210, 'criterion': 'gini'}

Best score: 0.895

**Most Important Features**

SelectFromModel in featureSelection module in sklearn was used to find the most important features.

[9, 8, 29, 10, 42, 35, 0, 6, 31, 16, 13, 5, 11, 80, 33, 4, 43, 39, 37, 40, 44, 48, 38, 45, 47]

feature importance

0 PID 0.015617

1 1 0.000693

2 2 0.004645

3 3 0.000467

4 4 0.011004

5 5 0.013262

6 6 0.014860

7 7 0.001757

8 8 0.146726

9 9 0.275765

10 10 0.059885

11 11 0.013210

12 12 0.002977

13 13 0.013296

14 14 0.000140

15 15 0.000489

16 16 0.013627

17 17 0.000217

18 18 0.001411

19 19 0.000526

20 20 0.005463

21 21 0.007839

22 22 0.006186

23 23 0.008298

24 24 0.001315

25 25 0.001103

26 26 0.006992

27 27 0.002240

28 28 0.002325

29 29 0.082396

.. ... ...

53 53 0.002982

54 54 0.002883

55 55 0.002410

56 56 0.003100

57 57 0.005501

58 58 0.005401

59 59 0.004894

60 60 0.004478

61 61 0.004279

62 62 0.004068

63 63 0.002966

64 64 0.001084

65 65 0.001230

66 66 0.000576

67 67 0.000822

68 68 0.001305

69 69 0.000904

70 70 0.000557

71 71 0.000335

72 72 0.000210

73 73 0.000600

74 74 0.001031

75 75 0.002607

76 76 0.000470

77 77 0.004168

78 78 0.005685

79 79 0.002992

80 80 0.012447

81 81 0.003900

82 82 0.003007

Total Features : 83 after feature processing and generation