For the intelligent data analysis , there are two situations possible in which intelligent data analysis can help us to find solutions to certain problems or provide answers to questions that arise.

In the first case the problem at hand is by no means new ,but it is already solved as a matter of routine.

In second case a certain question arises for the first time and only little experience is available or the experience is not directly available to this new question.In such cases ,it is supposed that data from related situations may be helpful to generalize the new problem or that unknown relationship can be discovered from the data to gain insights into this unfamiliar area. Here is the second situation situation for my project.

In this project(Aim is to find a pattern so that maximum passengers might be kept alive),we are going to follow the CRISP-DM proposal which isabout what the intelligent data analysis process should look like.

CRISP-DM consist of six phases.The phase are in order as: project understanding then data understanding then data preparation then modeling then evaluation then deployment. We have graphical representation of phases in word file.

1.Project Understanding: This is often called as bussiness understanding.This event happened in 1816 a.d. which was quite threatening and happened while travelling in Titanic ship.There were many reasons behind this unwanted event.I am interested to know that what might be pertaining situations at that time which caused the deathof passengers.All told I am interested in finding a typical pattern which could lead for passengers in order to keep them alive.

So from where should I get such promising pattern? Should I stick to pattern if I get that?

can I generalize that pattern over other passengers in order to get that how many of them could have been survived and how many of them could not?

Ultimately ,because of these sudden arised questions ,I can state following:

(a)Project Objective is to find a typical pattern in the attributes of passengers so that it can be generalized over other passengers in order to predict that how manyof them could have survived.

(b)Other objective should be that which attribute is best describing the death or alive status of passengers.

2.Data Understanding: I will approach the data from a neutral point of view as one should do.Since we must not trust any data as long as we have not carried out some

simple plausibility checks. So we shall do some plausibility checks later in this phase using particular methods. Since this phase requires taking a closer look at the data that is why we follow:

(a) Attribute Understanding : So for we have considered the aim of project.But in order to proceed further we need data.Fortunately we have data.

\*\*Talking about data: I have two datasets. One the train data and other is test data. I will emphasise on train data only.So what about train data?

dim(titanic) : 891 13 So our dataset contains 891 rows and 13 columns. It's mean is that there are 12 features and obivously 1 target variable.The variables are:

"PassengerId" "Survived" "Pclass" "Name" "Sex" "Age" "SibSp" "Parch" "Ticket" "Fare" "Cabin"

"Embarked" "Ag\_grp". The target variable being "Survived".

So now we come to main topic of attribute understaing.

Class of attributes are:

PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin Embarked Ag\_grp

"integer" "integer" "integer" "factor" "factor" "numeric" "integer" "integer" "factor" "numeric" "factor" "factor" "factor"

**Possible improvements:**

**1.**PassengerId: (a)class : Integer (b)Domain: {1,2,…,891}

No violation of syntactic accuracy. But the target feature “Survived” is of course not relevant with “PassengerId” in the sense that ordering of the passengers can be changed in data without any loss of data structure and it’s quality. So we should avoid this feature for further data analysis.

2. Survived: (a)class : Factor (b) Domain: {0,1}

This is the target variable which is categorical either 0 for those who died or 1 for those who survived.

We shall be searching those features which best support this target variable.

3. Pclass: (a)class : Factor(initially it was integer) (b)Domain: {1,2,3}

It is actually an ordinal attribute in which 1 specifies first class ,2 specifies second class and 3 specifies the third class of passenger’s coach.

The aim to convert the class of Pclass from integer to factor was just to ensure that difference between 1st and 2nd class and the difference between 2nd and 3rd class are not the same.

If we consider it as integer then 2-1=3-2=1 . Which might be illusive and might not lead to a better analysis of data.

4. Name : (a)class : factor (b)domain : Each having different name.

5. Sex : (a)class : factor (b)domain : {male,female}

6.Age : (a)class :numeric (b)domain: [0.42,80] (c)Na’s 177

7.Sibsp : (a)class : integer (b)domain : {0,1,2,….,8}

8.Parch : (a) class : integer (b) domain :{0,1,2,….,6}

9.Ticket : (a) class : factor (b)domain : There are 681 individual tickets for 891 passengers.

10.Fare : (a)class : numeric (b)domain : [0,513)

11.Cabin : (a)class : factor (b)domain : there were 148 Cabins and 687 peoples were without Cabin.

12.Embarked : (a)class : factor (b)domain: Three classes of embarking.

13.Ag\_grp : (a) : class :factor (b)domain: {adult , child}

(b) Data Quality : The results of an analysis can not be better than the quality of data , so that we should be concerned about the data quality before we carry out any deeper analysis eith the data. Data quality refers to how well the data fit to their intended use.

1.Accuracy : It is defined as the closeness between the value in the data and the true value.

a. Syntactic Accuracy : A considered value might not be correct but it belongs at least to the domain of the corresponding attribute. Our data is syntactically accurate.

e.g. if there is categorical attribute say gender and some entry is ‘fmale’ then there Syntactic accuracy is violated.

b. Semantic Accuracy: A value might be in the domain of corresponding attribute , but it is not correct.

Our data is Semantically accurate.

e.g. if there is categorical data gender with domain male and female only then corresponding to the passenger name John it turns out to be female then here is not question for syntactic accuracy because it is okay with it but it is obviously wrong entry. Here is not the Semantic accuracy.

2. Completeness :

a. Completeness with respect to attributes : It referes to the missing values. There is one variable in our data set called age which contains 177 missing values.So this attribute is not complete.

b. Completeness with respect to record : Mean’s that dataset contains necessary information that is required for analysis.

3.Balanced data : in our data thre is balance between one who survived and the one’s who did not.

4.Timelines : Thre is no problem for dynamic domain as it has been already discussed at the the top while concerning the situation in which I have driven myself to analyse dataset.

(c) Data Visualisation: We shall visualize each attribute.