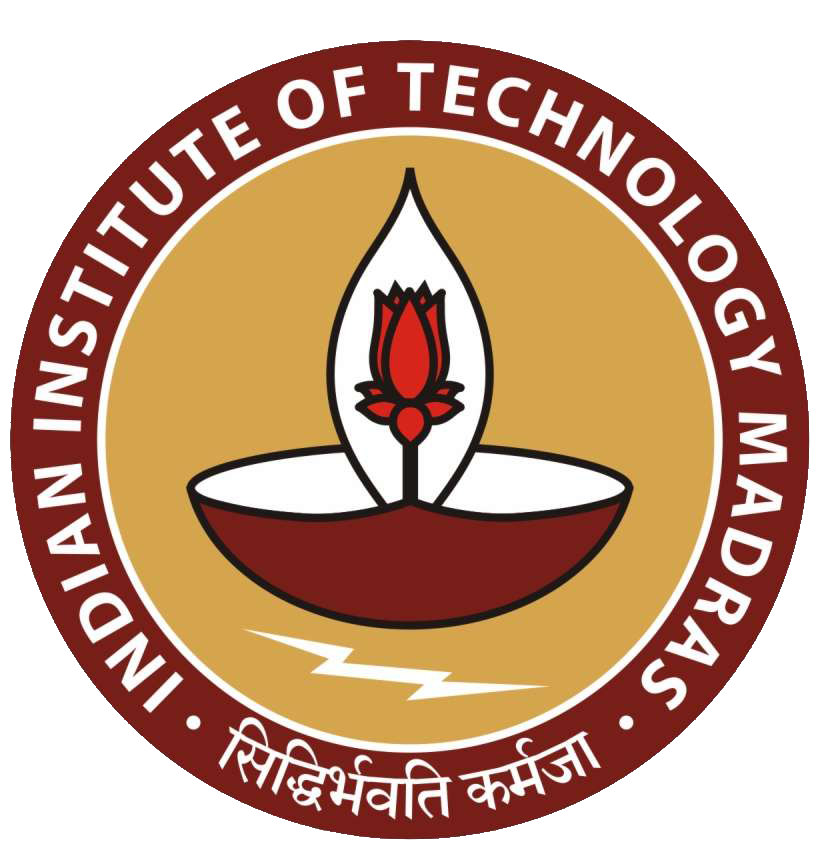
ID5002W: Industrial AI Laboratory

Assignment Report



ASSIGNMENT - 5

SOUMYA MUKHERJEE

CH24M571

### 1.0 Preface

We have been given a brain\_tumor data set with the aim of finding the SVM classifier model that fits the data . In the second question , we have the objective to predict the survival rate based on the tumor data provided .

### 2.0 Data Pre-processing

We start off by checking for null or missing values . Our data set seemed to be devoid of any such values . We dropped the patient\_id column as it seemed irrelevant to our goal .We notice there are 4 numerical columns and 14 categorical variables . SVM cannot work with categorical data , hence we adopted encoding methodologies based on the data on individual columns .

|  |  |
| --- | --- |
| Binary encoding (2 possible values ) | Gender, Radiation\_Treatment, Surgery\_Performed, Chemotherapy, Family\_History, MRI\_Result, Follow\_Up\_Required,Tumor\_Type |
| One Hot Encoding ( 2 or more values --> 4 possible values ) | Location, Histology, Symptom\_1, Symptom\_2, Symptom\_3 |
| Numerical Values ( Standard Scalar) | Age, Tumor\_Size, Survival\_Rate, Tumor\_Growth\_Rate |
| Ordinal Encoding (Order is important) | Stage |

The objective is to have different encodings for different columns . Had we used One Hot Encoding for all the columns , it would have created a lot of columns which would have contributed to longer running times for the SVM model .