AN INDUSTRIAL TRAINING REPORT

**on**

**BRAIN TUMOR ANALYSIS**

**Submitted by**

**KARTIK SRIVASTAVA**

**171500156**

Department of Computer Engineering & Applications

**Institute of Engineering & Technology**



**GLA University**

**Mathura- 281406, INDIA**

**August, 2019**

Declaration

I hereby declare that the work which is being presented in the Training Project Report **“BRAIN TUMOR ANALYSIS MODEL USING MACHINE ALGORITHM** in partial fulfillment of the requirements for Industrial Project is an authentic record of my own work carried under the supervision of **Prof**. **Amey** **Karkare, Co-Chief Investigator, E & ICT Academy,IIT KANPUR.**

Signature of Candidate:

Name of Candidate: Kartik Srivastava

Roll. No. : 171500156

Course: B.Tech

Year: 3rd

Semester: 5th

**Certificate**

**A screenshot of a cell phone

Description automatically generated**

**SYNOPSIS**

**Student Information:**

|  |  |
| --- | --- |
| Name: Kartik Srivastava | University Roll. No.171500156 |
| Mobile:9565030271 | Email: Kartik.srivastava\_cs17@gla.ac.in |

**Information about Industry/Organization:**

|  |  |
| --- | --- |
| Industry/Organization Name with full Address | Electronics & ICT Academy,IIT Kanpur, Kalyanpur,  Uttar Pradesh - 208016 |
| Contact Person | Name & Designation : Prof. Amey Karkare  Email : karkare@cse.iitk.ac.in |

**Project Information:**

|  |  |
| --- | --- |
| Title Of Project/Training/Task | Brain Tumor Analysis Model Using Machine Learning Algorithms |
| Role & Responsibility |  |
| Technical Details | Hardware Requirements: Laptop  Software Requirements: Google Colaboratory |
| Training Implementation Details | Fully Implemented |
| Training Period | Start Date: 11th June,2019  End Date: 9th July,2019  Duration Of Training (In Weeks):4 weeks |
| **Summary of the Training Work:**  In the training I had learned about Machine Learning and Data Analytics, and Basic Python Operations. It was 4 weeks training and everyday implementation of ML algorithms was taught and further used them in project.  Using the Brain Tumor Diagnostic Database, we can create a classifier that can help diagnose patients and predict the likelihood of breast cancer. A few machine learning techniques will be explored.  Data were obtained from survey questions completed by the radiologist during his observation of the patients. The results using logistic regression cross tabulation was to obtain the significant values between the breast cancer factors. | |

**Acknowledgement**

It gives me a great sense of pleasure to present the report of the Industrial Training undertaken before B. Tech. Third Year. This training in itself is an acknowledgement to the inspiration, drive and technical assistance contributed to it by many individuals. This project would never have seen the light of the day without the help and guidance that we have received.

Our heartiest thanks to **Prof. Amey Karkare,  Co-Chief Investigator, Electronics & ICT Academy, IIT Kanpur** for providing me with an encouraging platform to develop this project, which thus helped me in shaping our abilities towards a constructive goal.

We owe special debt of gratitude **Prof. Amey Karkare,** **Co-Chief Investigator, Electronics & ICT Academy, IIT Kanpur,**  for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for me. He has showered me with all his extensively experienced ideas and insightful comments at virtually all stages of the project & has also taught us about the latest industry-oriented technologies.

I also do not like to miss the opportunity to acknowledge the contribution of all authorities of the organization for their kind guidance and cooperation during the development of our project.

**ABSTRACT**

In this study, the diagnosis of brain tumor by MRI is complemented by using logistic regression. The radiologists can use the results to make a proper judgment as to the presence of brain tumor. Data were obtained from survey questions completed by the radiologist during his observation of the patients. The results using logistic regression cross tabulation was to obtain the significant values between the brain tumor factors. The classification table from 130 samples shows the occurrence from prediction and observation sample, producing percentage of correct classification for MRI results is 91.5%. The accuracy is compared with validated samples which are 46 samples and the percentage of correct classification is 67.4%. The analysis for MRI screening using parameter estimation is to identify all the factors that were available in the survey. The presence of headaches, seizures, nausea or vomiting, Memory problems had high odds of getting brain tumor.

**Contents**

• Declaration 2

• Certificate 3

• Synopsis 4

• Acknowledgment 5

• Abstract 6

1. Introduction 9

1.1 Objective / Problem 11

2. Model Design 12

2.1 DFD 12

3. Methods/Phases 13

3.1 Data Preparation 13

3.2 Data Exploration 15

3.3 Model Selection 19

4. Result 22

5. References/Bibliography 24

Building a Simple Machine Learning Model on Brain Tumor Data



**Introduction**

Brain Tumor (BT) is one of the most common cancers among people worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society.

Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

**Recommended Screening Guidelines:**

**Magnetic Resonance Imaging (MRI).** The most important screening test for brain tumor is the MRI. A MRI is an X-ray of the tumor. It is one of the most efficient techniques in recognizing tumor

**CT Scan**

**Some Risk Factors for Brain Tumor**

Most of the time, the cause of a brain tumor is unknown, but the following factors may raise a person’s risk of developing a brain tumor:

* **Age.** Brain tumors are more common in children and older adults, although people of any age can develop a brain tumor.
* **Gender.** In general, men are more likely than women to develop a brain tumor. However, some specific types of brain tumors, such as meningioma, are more common in women.
* **Home and work exposures.**Exposure to solvents, pesticides, oil products, rubber, or vinyl chloride may increase the risk of developing a brain tumor. However, there is not yet scientific evidence that supports this possible link.
* **Family history.**About 5% of brain tumors may be linked to hereditary genetic factors or conditions, including [**Li-Fraumeni syndrome**](https://www.cancer.net/node/19133), [**neurofibromatosis**](https://www.cancer.net/node/19450), **[nevoid basal cell carcinoma syndrome](https://www.cancer.net/node/19452)**, [**tuberous sclerosis**](https://www.cancer.net/node/19686), [**Turcot syndrome**](https://www.cancer.net/node/18852), and [**von Hippel-Lindau disease**](https://www.cancer.net/node/19322). Scientists have also found “clusters” of brain tumors within some families without a link to these known hereditary conditions. Studies are underway to try to find a cause for these clusters.
* **Exposure to infections, viruses, and allergens.** Infection with the Epstein**-**Barr virus (EBV) increases the risk of CNS lymphoma. EBV is more commonly known as the virus that causes mononucleosis or “mono”. In other research, high levels of a common virus called cytomegalovirus (CMV) have been found in brain tumor tissue. The meaning of this finding is being researched. Several types of other viruses have been shown to cause brain tumors in research on animals. More data are needed to find out if exposure to infections, other viruses, or allergens increase the risk of a brain tumor in people. Of note, studies have shown that patients with a history of allergies or skin conditions have a lower risk of glioma.
* **Electromagnetic fields.**Most studies evaluating the role of electromagnetic fields, such as energy from power lines or from cell phone use, show no link to an increased risk of developing a brain tumor in adults. Because of conflicting information regarding risk in children, the World Health Organization (WHO) recommends limiting cell phone use and promotes the use of a hands**-**free headset for both adults and children.
* **Race and** **ethnicity.** In the United States, white people are more likely to develop gliomas but less likely to develop meningioma than black people. Also, people from northern Europe are more than twice as likely to develop a braiain tumor as people in Japan.
* **Ionizing radiation.** Previous treatment to the brain or head with ionizing radiation, including x**-**rays, has been shown to be a risk factor for a brain tumor.
* **Head injury and seizures.**Serious head trauma has long been studied for its relationship to brain tumors. Some studies have shown a link between head trauma and meningioma, but not between head trauma and glioma. A history of seizures has also been linked with brain tumors, but because a brain tumor can cause seizures, it is not known if seizures increase the risk of brain tumors, if seizures occur because of the tumor, or if anti-seizure medication increases the risk.
* **N-nitroso compounds.**Some studies of diet and vitamin supplementation seem to indicate that dietary N-nitroso compounds may raise the risk of both childhood and adult brain tumors. Dietary N-nitroso compounds are formed in the body from nitrites or nitrates found in some cured meats, cigarette smoke, and cosmetics. However, additional research is necessary before a definitive link can be established.

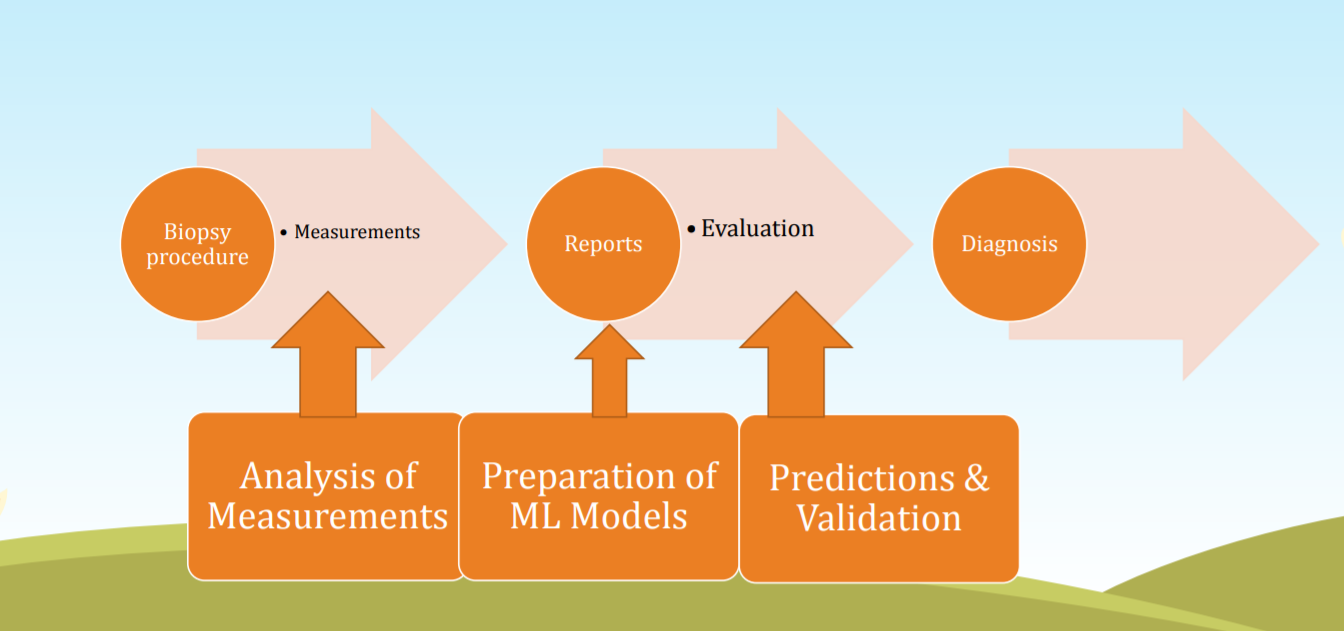
**Objectives**

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the brain tumor is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

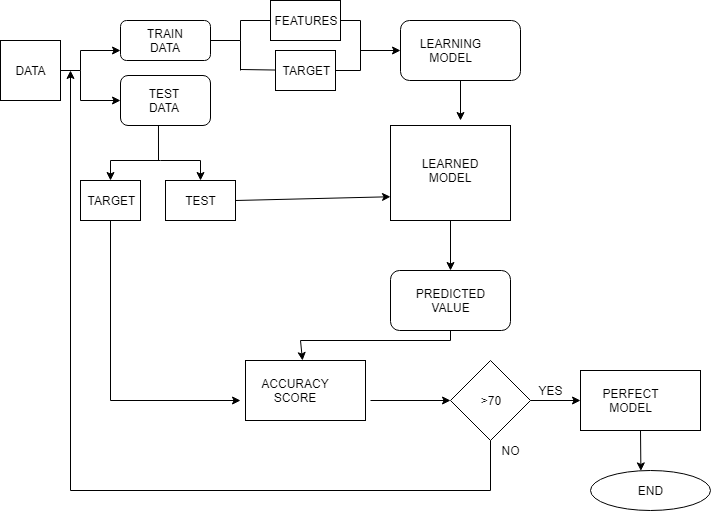
**Problem**

The early diagnosis of BT can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BT and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BT datasets, machine learning (ML) is widely recognized as the methodology of choice in BT pattern classification and forecast modelling.

**MODEL DESIGN**

****

**Data Flow Diagram**

****

**Phase 0 — Data Preparation**

* <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>
* The data used for this POC is from University of Wisconsin.
* This brain cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg

**Attribute Information:**

1. ID number 2) Diagnosis (M = malignant, B = benign) 3–32)

Ten real-valued features are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area — 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (“coastline approximation” — 1)

The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

**Phase 1 — Data Exploration**

We will be using ***Google Colab***  to work on this dataset. We will first go with importing the necessary libraries and import our dataset to Google Colab :

#importing the libraries  
  
  
import pandas as pd#importing our cancer dataset  
import io

data = pd.read\_csv( io.BytesIO(uploaded['BrainTumorData.csv']) , index\_col = False)

We can examine the data set using the pandas’ **head()** method.

dataset.head()

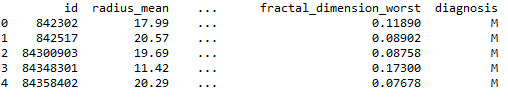


Fig : top 5 data of our dataset

We can find the dimensions of the data set using the panda dataset ‘shape’ attribute.

print("Cancer data set dimensions : {}".format(dataset.shape))Cancer data set dimensions : (569, 32)

We can observe that the data set contain 569 rows and 32 columns. ‘*Diagnosis*’ is the column which we are going to predict , which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as 0 and 212 as 0.

A screenshot of a cell phone

Description automatically generated

Visualization of data is an imperative aspect of data science. It helps to understand data and also to explain the data to another person. Python has several interesting visualization libraries such as Matplotlib, Seaborn etc.

In this tutorial we will use pandas’ visualization which is built on top of matplotlib, to find the data distribution of the features.

**A screenshot of a cell phone

Description automatically generated** ……………A close up of a logo

Description automatically generated

Fig : Visualization of Dataset

**Missing or Null Data points**

We can find any missing or null data points of the data set (if there is any) using the following pandas function.

dataset.isnull().sum()  
dataset.isna().sum()

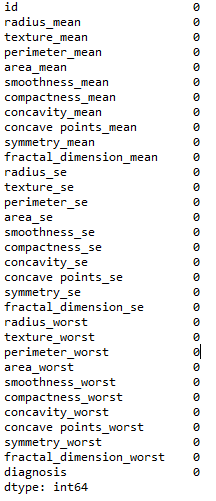


Fig : Observe missing data

**Splitting the dataset**

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model’s prediction on this subset.

We will do this using SciKit-Learn library in Python using the train\_test\_split method.

# Splitting the dataset into the Training set and Test set from sklearn.model\_selection import train\_test\_split  
X\_train,X\_test,Y\_train,Y\_test = train\_test\_split(X,Y,test\_size = 0.33)

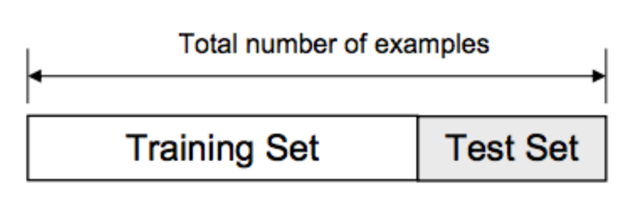


Fig: Training and test set

**Phase 2 — Model Selection**

This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for Predicting the best results.

Usually Data Scientists use different kinds of Machine Learning algorithms to the large data sets. But, at high level all those different algorithms can be classified in two groups : supervised learning and unsupervised learning.

Without wasting much time, I would just give a brief overview about these two types of learnings.

Supervised learning : Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. Supervised learning problems can be further grouped into **Regression** and **Classification** problems.

A **regression** problem is when the output variable is a real or continuous value, such as “salary” or “weight”.

A **classification** problem is when the output variable is a category like filtering emails “spam” or “not spam”

Unsupervised Learning : Unsupervised learning is the [algorithm](https://whatis.techtarget.com/definition/algorithm) using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

In our dataset we have the outcome variable or Dependent variable i.e Y having only two set of values, either M (Malign) or B(Benign). So we will use Classification algorithm of supervised learning.

We have different types of classification algorithms in Machine Learning :-

1. Logistic Regression

2. Nearest Neighbor

3. Support Vector Machines

Lets start applying the algorithms :

We will use sklearn library to import all the methods of classification algorithms.

We will use LogisticRegression method of model selection to use Logistic Regression Algorithm,

#Using **Logistic Regression** Algorithm to the Training Setfrom sklearn.linear\_model

import warnings

warnings.filterwarnings(action="ignore")

from sklearn import linear\_model

model = linear\_model.LogisticRegression()

X\_train,X\_test,Y\_train,Y\_test = train\_test\_split(X,Y,test\_size = 0.33)

model.fit(X\_train, Y\_train)

#Using **KNeighborsClassifier** Method of neighbors class to use Nearest Neighbor algorithm*from sklearn.neighbors import KNeighborsClassifier  
classifier = KNeighborsClassifier(n\_neighbors = 5, metric = 'minkowski', p = 2)  
classifier.fit(X\_train, Y\_train)*

#Using **SVC method** of svm class to use Kernel SVM Algorithm  
*from sklearn.svm import SVC  
classifier = SVC(kernel = 'rbf', random\_state = 0)  
classifier.fit(X\_train, Y\_train)*

We will now predict the test set results and check the accuracy with each of our model:

predictedY = model.predict(X\_test)

**Result**

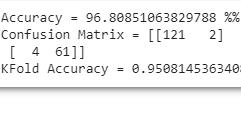
To check the accuracy we need to import confusion\_matrix method of metrics class. The confusion matrix is a way of tabulating the number of mis-classifications, i.e., the number of predicted classes which ended up in a wrong classification bin based on the true classes.

from sklearn.metrics import confusion\_matrix  
cm = confusion\_matrix(Y\_test, Y\_pred)

We will use Classification Accuracy method to find the accuracy of our models. Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

https://miro.medium.com/max/466/1*AQ3X4c8Ot2FDbyyCcOA7nA.png

Fig: Accuracy



To check the correct prediction we have to check confusion matrix object and add the predicted results diagonally which will be number of correct prediction and then divide by total number of predictions.

After applying the different classification models, we have got below accuracies with different models:

1. Logistic Regression — 96.8%

2. Nearest Neighbor — 95.1%

3. Support Vector Machines — 97.2%

So finally we have built our classification model and we can see that Support Vector Machines Classification algorithm gives the best results for our dataset. Well its not always applicable to every dataset. To choose our model we always need to analyze our dataset and then apply our machine learning model.

This is a basic application of Machine Learning Model to any dataset Remember to always keep in mind the problem of [overfitting and underfitting](https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76) as well.

**Bibliography**

Site Referred:-

⚫ <http://scikit-learn.org/stable/index.html>

⚫

⚫ <http://medium.com>

⚫ <http://pythonprogramming.net/>

⚫ <http://pythonprogramming.net/introduction-to-python-programming>