TABLE OF CONTENTS

I. INTRODUCTION

This report is a summary of my 45-day internship at Hex N Bit, which was required as part of my Summer Internship program. This was a virtual internship that began on June 5, 2021 and ended on July 14, 2021. The study was focused on AIML because I am interested in Artificial Intelligence and Machine Learning.

At the beginning of the internship, I set up certain learning objectives that I want to achieve by the end of this intern.

- Reworking on my python Programming.
- Understanding various modules in python
- Data visualization
- Machine learning Techniques
- Open CV

This entire Internship project is based on Machine learning and how the machine learning techniques are assisting in different sectors by replacing the human intelligence.

So before going to dive in to the internship details here are the major skills that acquired from the internship.

Artificial Intelligence:

Artificial Intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs. It is basically understanding/ creating human intelligence.

Machine Learning:

ML is a subset of AI. ML is a science of designing and applying algorithms that are able to learn things from the data.

ML lifecycle:

1. Gathering data
2. data preparation
3. data cleaning

4. analyzing the data

5. training

6. testing

7. deployment

There are 3 different types of machine learning algorithms:

1. **Supervised Learning:**

   This is a type of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output.

   The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

   Supervised machine learning involves ----regression and classification

2. **Unsupervised Learning:**

   Unsupervised machine learning uses machine learning algorithms to analyze and cluster unlabeled datasets.

   Unsupervised machine learning involves----- clustering, association

3. **Reinforcement learning**:

   Reinforcement learning agent is able to perceive and interpret its environment, take actions and learn through trial and error.

Importance of Machine Learning:

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

Applications Of Machine Learning:

There are many sectors where machine learning is playing vital role. Here are some of the day-to-day applications:

1. Image recognition: Automatic friends tagging suggestion face detection etc.

2. Medical Diagnosis: Thyroid gland classification
3. Product recommendation: Movie suggestions in Netflix and product suggestions based on our searches in amazon or flipkart.
4. Email spam and Malware Filtering.
5. Stock Market Trading.
6. Virtual Personal assistant: Like Siri, Cortana, Google assistant
7. Self-driving cars. (Tesla)

II. OBJECTIVE OF THE INTERNSHIP

The objective of the internship is, reworking on basics of python, to prepare the data, exploring the data with visual analysis, after performing exploratory analysis on the data preparing and applying it to various machine learning Models, Training, Testing, Tuning and Evaluating the Machine Learning models. So, with that experience we can apply this Practical knowledge in real-life scenario to solve problems.

III. SKILLS ACQUIRED THROUGH THE INTERNSHIP

Week 1:

- Introduction to AIML
- Introduction to Python
- Getting started with Anaconda and Jupyter
- Python Data Types, Conditional Statements, Loops and Control Statements.
- Python Functions and Lambda functions.
- Map & Filter
- File Handling.
- Assignment 1: Basic Operations in Python (15 tasks given)

Week 2:

- NumPy Basics, Operations, Indexing and slicing.
- Pandas Series and Data Frames.
- Pandas Operations, Fixing Missing data, Merging, Group By
- Pandas File Reading and writing.
- Assignment 2: Questions related to NumPy and Pandas (22 tasks given)

Week 3:

- Introduction to Matplotlib_Pyplot API
- Matplotlib Object Oriented API
- Data Visualization
- Supervised Machine learning techniques & Evaluation Metrics.
- Linear Regression.
- Logistic Regression
- Decision Tree
- Assignment 3: Plotting sin and cos graphs, Given a dataset perform exploratory data analysis, using logistic regression model and predicting, performing metrics.

Week 4:

- Introduction to Support Vector Machine
- Unsupervised Machine Learning techniques

- K Means Theory and application
- K Means clustering
- Association Rule and Apriori Algorithm
- Assignment 4: Given a dataset, performing exploratory data analysis, plotting an elbow plot in order to implement K Means Clustering, performing K Means Clustering, Creating a scatter plot as per the clustered values

Week 5:

- Introduction to Computer Vision
- Drawing functions and basic operations in CV
- K Nearest Neighbor Theory
- Optical Character Recognition using KNN
- Arithmetic Operations in Open CV
- Color Spaces Histogram Thresholding
- Bitwise Operations and Masking
- Image processing basics
- Object Tracking using color
- Object Detection
- Assignment 5: Opening Webcam, displaying grayscale feed, detecting face, nose, eyes, smile using Open CV & Mask Detection using Open CV.
- **Final Project**

IV. OVERVIEW OF THE PROJECT/WORK CARRIED OUT DURING INTERNSHIP

Abstract:

The thyroid gland is a vascular gland that secretes two hormones which help in supervising the metabolism of the body. The two Types of Thyroid disorders are Hyper thyroidism and Hypothyroidism. When this disorder occurs in the body, they release some hormones which imbalances the body metabolism. Thyroid related (T3 Uptake test) is used to detect this disease. Machine learning plays vital role in the disease prediction. We will use Different Machine Learning algorithms like Logistic Regression, Support Vector Machine(SVM), Decision Tree, Kmeans Clustering to predict the person's risk of getting thyroid disease.

Dataset Description:

This dataset has 150 instance of normal class, 35 instance of hyperthyroidism class and 30 instance of hypothyroidism class. Class attribute, T3 resin uptake test, total T4 , total T3 , TSH , difference of TSH value after injection of 200 micrograms of thyrotropin releasing hormone

Attributes:

We have 6 columns in the dataset 5 are attributes and one label

- TSH -- Thyroid Simulating Harmons
- TSTI -- Total Serum Tri Iodothyronine
- TST -- Total Serum Thyroxin
- T3 -- T3-Resign Uptake Test
- Maximal Absolute Difference
- Class =1,2,3

Domain:

Disease Detection(Thyroid gland classification)

Classifiers:

Logistic Regression:

Logistic regression is a statistical method for predicting binary classes. predicts the probability of occurrence of a binary event utilizing a logit function.

How Logistic Regression Algorithm works:

- To separate the records, choose the best attribute using Attribute Selection Measures (ASM).
- Import train test split from skit-learn to split the data and train/test the model.
- Use LogisticRegression () to train the model and then forecast the values.
- Finally, assess the data by determining accuracy, recall, precision value, and so on.

Support Vector Machine:

A Support vector machine is a classification algorithm that can be operated for both classification as well as Regression purposes. The Main objective of SVM is finding a hyperplane that divides a dataset into two classes.

How does the Decision Tree Algorithm Works:

- Select the best attribute using ASM(Attribute Selection Measures) to split the data
- Split the data into train and test data and predict the values using SVC() function
- In most of the cases the accuracy of this model is low for the given dataset so we will try to tune the given dataset using GridSearchCv()
- Then by changing the hyperparameters we will tune the performance of model so that the accuracy can be increased to the certain point.
- So We will consider it as final accuracy the model.

Decision Tree:

A decision tree is a flowchart-like tree structure where an internal node represents feature, the branch represents a decision rule and each leaf node represents the outcome.The top most node is known as root node. It learns to partition the criteria of the attribute value. It performs and makes decision in recursive manner. Its training time is faster compared to network algorithms. It does not depend upon probability distribution assumptions So that it can handle high dimensional data with good accuracy.

How does the Decision Tree Algorithm Works:

The fundamental idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures (ASM) to split the records.
2. Make that attribute as a decision node and breaks the remaining dataset into smaller chunks.
3. Begin creating the tree by recursively repeating this approach for each kid until one of the conditions matches:

- All of the tuples have the same attribute value
- There are no more attributes to add.
- No more Instances exist.

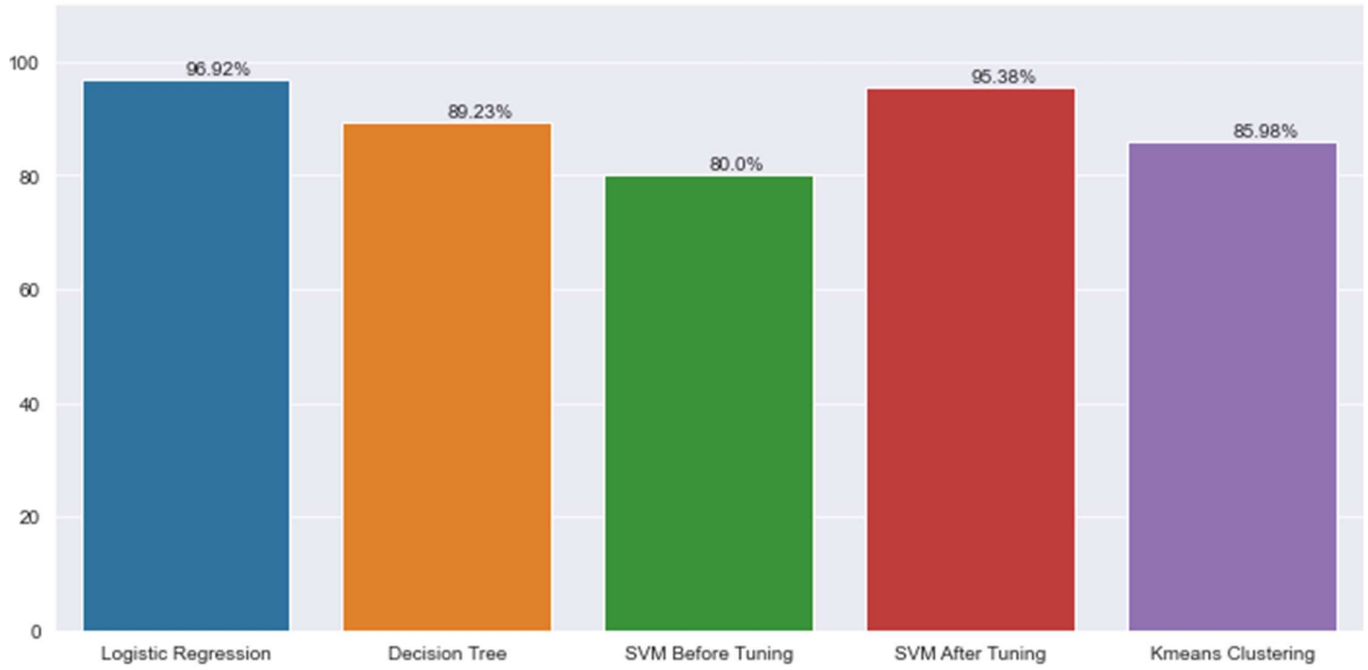
K-means Clustering:

K-means Clustering is an unsupervised learning algorithm which deals with clustering problems in machine learning. This Algorithm work on the principle of grouping the similar data into different clusters of unlabeled datasets. K specifies the number of pre-defined clusters that must be produced during the process; for example, if K=2, two clusters will be created, and if K=3, three clusters will be created, and so on. It's a centroid-based approach, which means that each cluster has its own centroid. The main goal of this technique is to reduce the sum of distances between data points and the clusters that they belong to.

The k-means clustering algorithm primarily accomplishes two goals:

- Iteratively determines the optimal value for K centre points or centroids.
- Each data point is assigned to the k-center that is closest to it. A cluster is formed by data points that are close to a specific k-center.
- As a result, each cluster contains datapoints with certain commonality and is isolated from the others.

V. RESULTS/OUTPUT



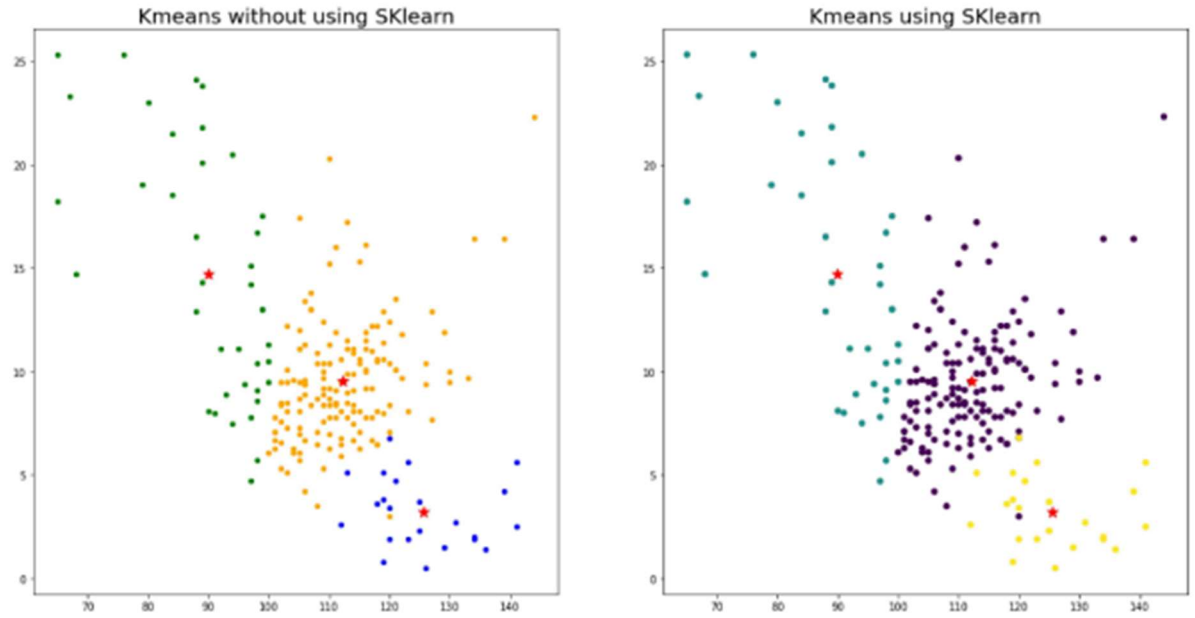
Here we got that Logistic regression algorithm and SVM gave best accuracy scores with more than 95% and there after decision tree is one such algorithm with a accuracy of 90%. So To predict the thyroid gland, we have made use of ***Logistic regression, Decision Tree***, and ***SVM Model*** (We have implemented the support vector machine model in general as well as hyper parameter tuning SVM model Here hyper parameters can be gamma, C and etc.) to train our dataset and to predict thyroid disease with more accuracy.
So, we can any one of the three Classifiers to get the desired output.

K-means Clustering without using SK learn module:

We are going to implement this kmeans clustering with the help of Euclidian distance Formulae $\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$. Here we will take k number of points as the start point of that particular cluster and measure the distance to other points. based on the distance from one point to another point will try to classify them into

different clusters, after performing two distance calculations we will try to calculate the centroid by doing to average to the clusters point. Once we get the standard center point our task will be quite easy to classify them into clusters between it will act as reference point. So, for every iteration we will calculate the distance from centroid to other points and classify them into different domains based on distance and update the centroid position by calculating the average so in this way we will keep on calculating the distance from centroid to other points and if the distance is minimum then it will be treated as their domain of the given dataset.

```
Text(0.5, 1.0, 'Kmeans using SKlearn')
```



So, the above figure represents how the data got arranged into different clusters by using the SK learn module as well as general data plotting with the help of Euclidean distance formula. The accuracy score for the general plotting is quite similar as SK learn module with an accuracy of 86%.

VI. CONCLUSION

In conclusion based on above analysis
Best Tuning Algorithm: Logistic Regression, Decision Tree Classifier and SVM
The accuracy of SVM (Before Tuning) is comparatively low when compared to SVM (After Tuning)
Correlation among the various features can be obtained easily using a heatmap.

Future Scope of Improvement in the algorithms:
This machine is trained to detect whether the person is having normal, hyper, hypo thyroidism based on the user's input. Further Development can be done by using image processing to predict thyroid nodules and cancer which cannot be recognized in blood test report.
So, by Combining both ideas and results, thyroid disease prediction can cover all thyroid related diseases.

VII. REFERENCES

https://www.javatpoint.com/machine-learning

https://www.python.org/

https://www.hexnbit.com/

https://scikit-learn.org/stable/

https://matplotlib.org/stable/gallery/index.html#

https://www.tableau.com/learn/articles/data-visualization

https://pandas.pydata.org/

https://numpy.org