INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 308-316

CANCER PREDICTION AT EARLY-STAGES USING MACHINE LEARNING

Alladi Ramesh[1], Bandela Dileep[2], Adepu Naveen[3], Youthkar Manoj[4], Vartya Vijay[5]

[1]Associate. Professor, CSE Dept, ACE Engineering College, Hyderabad, India.

[2,3,4,5]Student, CSE Dept, ACE Engineering College, Hyderabad, India.

ABSTRACT

The project begins by collecting and curating extensive datasets comprising clinical records, medical imaging, and genomic information from diverse sources. These datasets are used to train and validate machine learning models that can effectively distinguish between cancer and non-cancer cases in their early stages.

Cancer remains one of the leading causes of mortality worldwide, emphasizing the critical need for early detection and intervention. This study presents a novel approach utilizing machine learning algorithms for the early prediction of cancer. Early-stage detection significantly improves treatment outcomes and patient survival rates. The proposed model harnesses the power of machine learning techniques, specifically algorithms used, RFC, KNN Support vector Machine, etc., to analyze patient data and predict the likelihood of cancer development. A comprehensive dataset comprising and utilized for training and validation.

Various features such as genetic markers, lifestyle factors, and medical history were incorporated to enhance the predictive capability of the model. The machine learning model was trained on a subset of the data and fine-tuned using cross-validation techniques to optimize performance.

Results demonstrate the efficacy of the proposed model in accurately predicting cancer at early stages. Comparative analyses with traditional methods highlight the superiority of machine learning in terms of sensitivity, specificity, and overall accuracy. Furthermore, the model's interpretability allows for the identification of key features contributing to the prediction, offering valuable insights for clinicians.

1. INTRODUCTION

Essentially, this project will be able to Detect the Cancer at Early-Stages using machine Learning. In recent years, advancements in technology have revolutionized the field of healthcare, particularly in the realm of disease detection and diagnosis.

Among these breakthroughs, the integration of machine learning algorithms in cancer detection stands out as a promising avenue for early intervention and improved patient outcomes. Cancer, a complex and multifaceted disease, often manifests silently in its early stages, making timely detection a critical factor in successful treatment.

Machine learning, a branch of artificial intelligence, empowers computer systems to learn and adapt from data without explicit programming. Leveraging vast datasets comprising patient records, medical imaging scans, genetic profiles, and more, machine learning algorithms can identify subtle patterns and anomalies indicative of cancerous growth with remarkable accuracy. By analyzing diverse sets of patient data, these algorithms can recognize unique biomarkers associated with specific types of cancer, enabling clinicians to diagnose the disease at its nascent stages when treatment options are most effective.

The integration of machine learning in cancer detection offers several key advantages, including increased efficiency in screening processes, enhanced accuracy in diagnosis, and the potential to personalize treatment plans based on individual patient profiles. Moreover, by enabling early detection, these systems hold the promise of reducing mortality rates and alleviating the burden of advanced-stage treatments on patients and healthcare systems alike.

2. OBJECTIVES

In our project there are 3 objectives. They can be listed as:

- Early Intervention
- Improved Survival Rates
- Lower Health Care Costs.

3. METHODOLOGY

Review inputs and outputs for project activities. Information will be collected and prioritized. An appropriate algorithm or framework has been selected. Several estimation algorithms will be compared and the best method will be selected. Software and hardware selection will be made according to the needs. Data will be used as a process or framework

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 308-316

4. LITERATURE SURVEY

TITLE: Early prediction of cancer based on the combination of trace element.

AUTHOR: Chao Tan, Hui Chen, Chengyun Xia..

YEAR: 2008

DESCRIPTION:

This work investigates the feasibility of a combination of Adaboost (ensemble from machining learning) using decision stumps as weak classifier and trace element analysis for predicting early lung cancer. Kennard and Stone (KS) algorithm coupled with an alternate re-sampling was used to realize sample set partitioning. The whole dataset was split into equally sized training and test set.

DISADVANTAGES:

- The main disadvantage of Adaboost is that it needs a quality dataset. Noisy data and outliers have to be avoided before adopting an Adaboost algorithm.
- KS algorithm is not a suitable method for sample partition because this method leads to different distribution of calibration and test samples.

TITLE: Decision Tree of Occupational Cancer Using Classification and Regression Analysis.

AUTHOR: Tae-Woo Kim, Dong- Hee Koh, Chung-Yill Park

YEAR: 2010

DESCRIPTION:

Determining the work-relatedness of cancer developed through occupational exposures is very difficult. Aims of the present study are to develop a decision tree of occupational cancer. The Classification and Regression Test (CART) model was used in searching for predictors of occupational cancer..

DISADVANTAGES:

- The CART model must be used sparingly in deciding the work-relatedness of lung cancer because it is not absolute.
- This decision tree must be considered as a minimal decision standard of work-relatedness for cancer.
- To make accurate decision standards for occupational cancer, additional studies for elevating validation have to be performed.

TITLE: Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients.

AUTHOR: Maciej Zięba

YEAR: 2014

DESCRIPTION:

In this paper, we present boosted SVM dedicated to solve imbalanced data problems. Proposed solution combines the benefits of using ensemble classifiers for uneven data together with cost-sensitive support vectors machines.

Finally, boosted SVM is used for medical application of predicting post-operative life expectancy in the lung cancer patients.

DISADVANTAGES:

- The main disadvantage is this method is used for post-operative life expectancy in the cancer patients.
- SVM algorithm is not suitable for large data sets.
- SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
- In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.

5. PROPOSED SYSYTEM

Machine Learning and AI:

Machine learning and artificial intelligence (AI) algorithms are increasingly used to analyze medical data and identify patterns indicative of cancer.

This major project focuses on the development of a robust and accurate cancer prediction system for early stage detection. Leveraging the power of Machine Learning and medical data analysis, this project aims to revolutionize cancer diagnosis by enabling timely interventions and treatments.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 308-316

6. HARDWARE AND SOFTWARE REQUIREMENTS

6.1 HARDWARE REQUIREMENTS:

- Processor: Min. Core i3 processor
- RAM: 2GB (Min.) or 8GB (Recommended)
- Hard Disk Space: 50GB+

6.2 SOFTWARE REQUIREMENTS:

- Programming Language: Python
- Operating System: Windows 7 or later versions
- of windows.

7. PACKAGES USED

TensorFlow- TensorFlow is a popular open-source Python machine learning toolkit for creating and training deep neural networks. It has a versatile architecture and supports a variety of platforms, including CPU, GPU, and TPU. TensorFlow simplifies the implementation of complicated algorithms and models, allowing developers to create scalable and efficient machine learning systems.

Keras - Keras is a Python-based high-level neural network API that operates on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML. It offers an easy-to-use interface for building and training deep learning models, letting users to easily experiment with alternative architectures and hyperparameters. Keras also provides pre-trained models as well as a huge collection of building blocks for developing sophisticated models.

Skimage- Scikit-image (skimage) is a Python image processing and computer vision toolkit. It includes methods and routines for image improvement, segmentation, feature extraction, and other tasks. Skimage is built on top of other well-known scientific Python libraries like NumPy, SciPy, and matplotlib, making it simple to incorporate into current Python workflows.

Scipy- Scipy is a Python package for scientific and engineering computing. It includes modules for optimization, integration, linear algebra, signal processing, and other tasks. Scipy is built on top of Numpy, another famous Python package for scientific computing, and the two combined constitute a strong data analysis and numerical calculation tool.

Numpy- NumPy is an important Python package for scientific computation. It supports huge, multidimensional arrays and matrices, as well as a diverse collection of high-level mathematical operations for these arrays. NumPy is a popular choice for numerical operations in scientific research and data analysis due to its efficient and user-friendly interface.

Pandas - Pandas is a popular open-source Python data analysis and manipulation package. It offers sophisticated data structures and tools for working with structured data, including as data frames and series, and it allows for quick data processing, cleaning, merging, and reshaping. Pandas also supports reading and writing a variety of file types, including CSV, Excel, and SQL databases.

Matplotlib- Matplotlib is a popular Python data visualization package. It includes line graphs, scatter plots, bar plots, and histograms among its 2D and 3D displays. Matplotlib is a useful tool for data exploration and communication since it is extremely customizable and supports extensive labelling, annotations, and text formatting.

OS and time- The 'os' module in Python allows you to interact with the operating system. It has functions for creating and removing folders, manipulating files, and changing environment variables. The 'time' module in Python contains methods for working with time-related actions. It has functions for obtaining the current time, postponing program execution, and converting between several time formats.

8. TECHNOLOGY DESCRIPTION

Python is an interpreted high-level programming language that is simple to learn and use. It features a basic and clear syntax that makes it suitable for both beginners and professionals. Python is utilized in many different areas, such as web development, scientific computing, data analysis, and artificial intelligence.

9. SOURCE CODE

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import cufflinks as cf
```

```
import plotly
from plotly.offline import init_notebook_mode,iplot,plot
init_notebook_mode(connected=True)
cf.go_offline()
df = pd.read_csv('C:\\Users\\dileep bandela\\Desktop\\Projects\\Major\\Cancer.csv')
df.head()
df.info()
df.drop(['Patient Id'],axis = 1,inplace=True)
df['Level']
df['Level'].replace('Medium','High',inplace=True)
df['Level'].replace('High','1',inplace=True)
df['Level'].replace('Low','0',inplace=True)
df.head()
df['Level'] = pd.to_numeric(df['Level'])
df.isnull()
df.isnull().any()
import seaborn as sns
sns.heatmap(df.isnull())
plt.figure(figsize=(10,5))
sns.countplot(x='Smoking',data=df)
plt.figure(figsize=(10,5))
sns.countplot(x='ChestPain',data = df)
plt.figure(figsize=(10,5))
sns.boxplot(x='ChestPain',y='Age',data = df)
plt.figure(figsize=(10,5))
sns.boxplot(x='Smoking',y='Age',data = df)
sorted_smokers = df.groupby('Age')['Smoking'].count().to_frame()
sorted_smokers.style.background_gradient(cmap = 'Reds')
df.style.background_gradient(cmap = 'Reds')
label = df.Age.sort_values().unique()
target = sorted_smokers.Smoking
print(label)
print(target)
import plotly.graph_objects as go
fig = go.Figure()
fig.add_trace(go.Bar(x=label,y=target))
fig.update_layout(title = 'Smokers per age',xaxis=dict(title='Age'),yaxis=dict(title='Smokers'))
fig.show()
fig = go.Figure()
fig.add_trace(go.Scatter(x=label,y=target,mode='markers+lines'))
fig.update_layout(title = 'Smokers per age',xaxis=dict(title='Age'),yaxis=dict(title='Smokers'))
fig.show()
# Machine Learning
## RandomForest Classifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.metrics import log_loss, f1_score
```

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 308-316

```
from sklearn.model_selection import cross_val_score
import numpy as np
acc_dict = {}
# create the data
X = df.drop('Level',axis = 1)
y = df['Level']
X_train, X_test, y_train, y_test = train_test_split(X,y)
from sklearn.ensemble import  RandomForestClassifier
# create model
model = RandomForestClassifier()
# fit the data in the model
model.fit(X_train,y_train)
y_pred_randomF = model.predict(X_test)
print('Accuracy score : ',accuracy_score(y_test, y_pred_randomF)*100)
acc_dict['RFC_log_loss'] = log_loss(y_test, y_pred_randomF)
acc_dict['RFC_F!1_Score'] = f1_score(y_test, y_pred_randomF,average='weighted')
# prediction visualization
plt.imshow(np.log(confusion_matrix(y_test,y_pred_randomF)),cmap = 'Blues',interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
## KNeighbourClassifier
from sklearn.neighbors import KNeighborsClassifier
# to find the best k
score = 0
scores, highscore, bestk = 0, 0, 0
for k in range(3,12):
knn = KNeighborsClassifier(n_neighbors=k)
scores = cross_val_score(knn, X_train, y_train)
score = scores.mean()
if score>highscore:
highscore = score
bestk = k
print('Best k is {} with score {}'.format(bestk, highscore))
knn = KNeighborsClassifier(n_neighbors=bestk)
knn.fit(X_train,y_train)
# prediction
y_predict = knn.predict(X_test)
print('Accuracy score : ',accuracy_score(y_test,y_predict)*100)
acc_dict['KNN_log_loss'] = log_loss(y_test, y_predict)
acc_dict['KNN_F!1_Score'] = f1_score(y_test, y_predict,average='weighted')
# prediction visualization
plt.imshow(np.log(confusion_matrix(y_test,y_predict)),cmap = 'Blues',interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
## Tree Classifier
```

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 308-316

```
from sklearn.tree import DecisionTreeClassifier
tree_ = DecisionTreeClassifier()
tree_.fit(X_train,y_train)
y_pred = tree_.predict(X_test)
print('Accuracy score : ',accuracy_score(y_test, y_pred)*100)
acc_dict['Tree_log_loss'] = log_loss(y_test,y_pred)
acc_dict['Tree_f!1_score'] = f1_score(y_test,y_pred)
# prediction visualization
plt.imshow(np.log(confusion_matrix(y_test,y_pred)),cmap = 'Blues',interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
## SVM
from sklearn.svm import SVC
model = SVC()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print('Accuracy score : ',accuracy_score(y_test, y_pred)*100)
acc_dict['svc_log_loss'] = log_loss(y_test,y_pred)
acc_dict['svc_f!1_score'] = f1_score(y_test,y_pred)
# prediction visualization
plt.imshow(np.log(confusion_matrix(y_test,y_pred)),cmap = 'Blues',interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
```

10. OUTPUT

	Patient Id	Age	Gender	AirPollution	Alcoholuse	DustAllergy	OccuPationalHazards
0	P1	33	1	2	4	5	4
1	P10	17	1	3	1	5	3
2	P100	35	1	4	5	6	5
3	P1000	37	1	7	7	7	7
4	P101	46	1	6	8	7	7

5 rows × 25 columns

Fig 10.1: Dataset Values

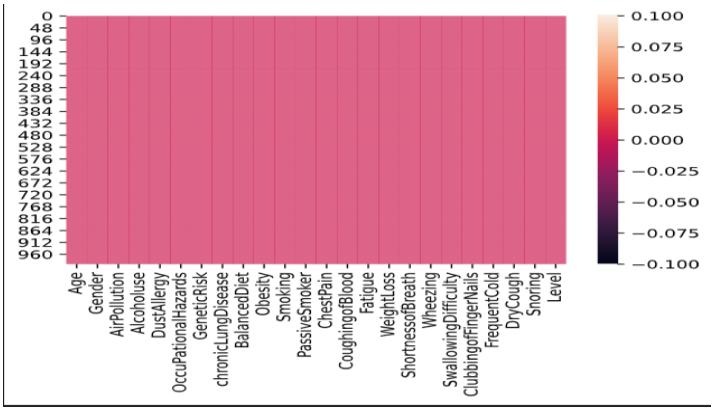

Fig 10.2: Heatmap of Dataset

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 308-316

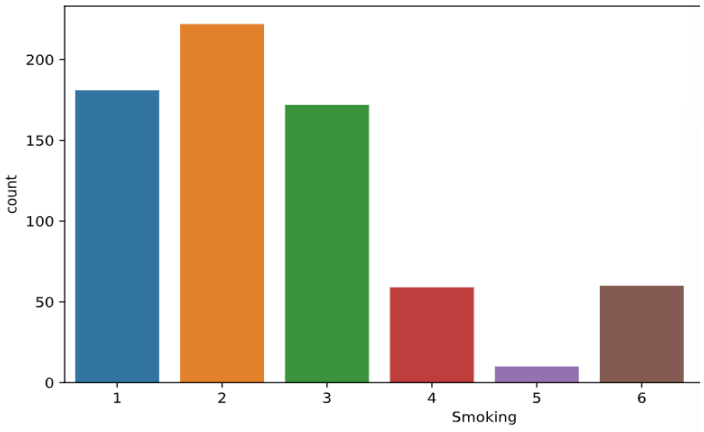

Fig 10.3: Countplot of Smoking

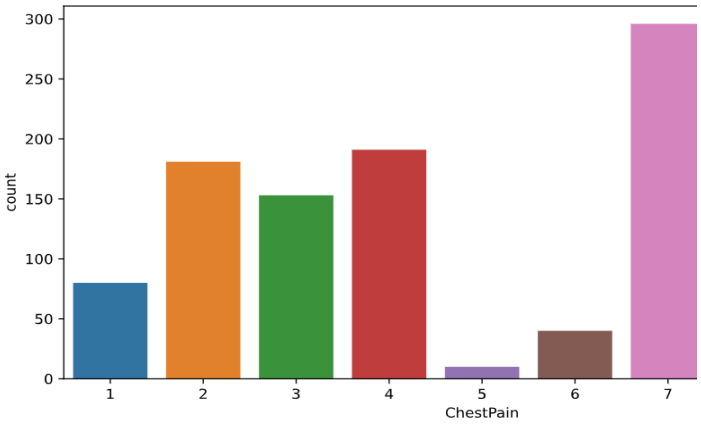

Fig 10.4: Countplot of Chestapin

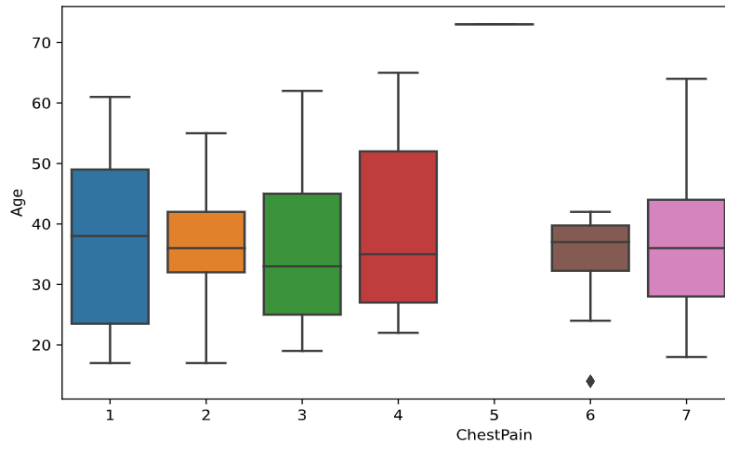

Fig 10.5: Boxplot for Chestpain according to age

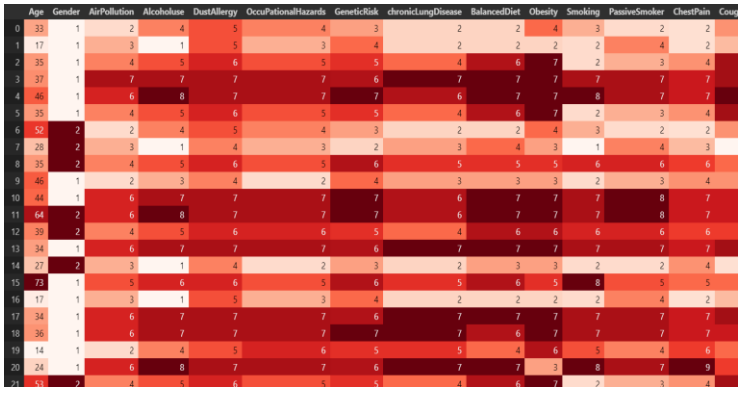

Fig 10.6 :Gradient for Whole Dataset

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 308-316

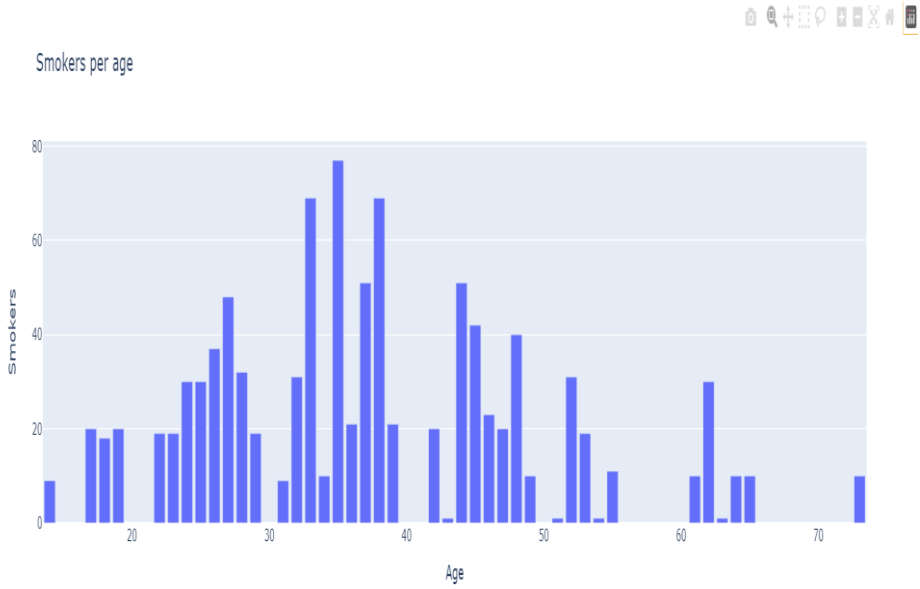

Fig 10.7: Updated Layout for Smokers per Age

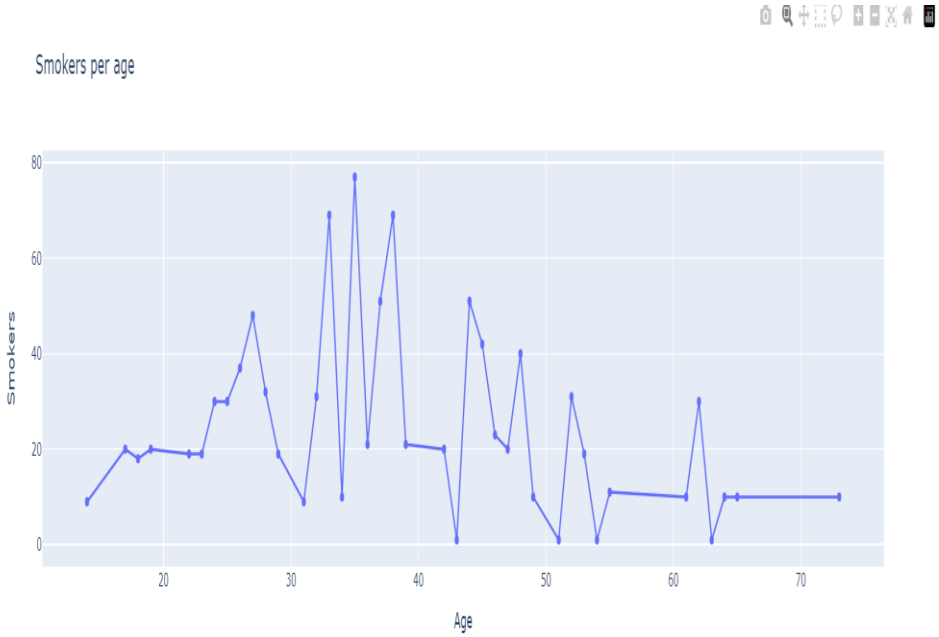

Fig 10.8: Tracing of Smokers per Age Using Markers

11. CONCLUSION

This research will assist in the detection of cancer at Early-Stages. Identifying cancer in its early stages leads to better treatment outcomes and reduces healthcare costs. Machine learning algorithms can analyze data and detect early cancer signs that may be missed by traditional techniques. In conclusion, the application of machine learning algorithms for the early prediction of cancer holds immense promise and potential in revolutionizing healthcare. Through this review, we have explored various studies and methodologies that demonstrate the effectiveness of machine learning in detecting cancer at its nascent stages. The advancements in technology, coupled with the availability of vast amounts of data, have enabled the development of robust predictive models that can assist healthcare professionals in making timely and accurate diagnoses. One of the key findings of our review is the significant impact that early-stage cancer prediction can have on patient outcomes. By identifying cancer in its early phases, treatment options can be initiated promptly, leading to higher chances of successful intervention and improved survival rates. Moreover, early detection often translates to less invasive and less costly treatments, reducing the burden on both patients and healthcare systems. Furthermore, the versatility of machine learning algorithms allows for the integration of various types of data, including genetic, imaging, and clinical data, to enhance predictive accuracy. The ability to analyze complex patterns and subtle changes in data sets enables these models to outperform traditional methods in terms of sensitivity and specificity.

12. FUTURE SCOPE

- Advancements in technology
- Continued innovation in machine learning and data analysis will further improve early cancer detection methods.
- Research opportunities
- Exploring novel biomarkers and combinations of machine learning techniques holds promise for more accurate and efficient cancer diagnosis.
- Collaborative efforts
- Collaboration between researchers, clinicians, and technology experts is crucial to unlock the full potential of machine learning in cancer detection.

13. REFERENCES

[1] Machine Learning Basics with the K-Nearest Neighbors Algorithm by Onel Harrison, [Online] Available: https://towardsdatascience.com/machine-learning-basics-with-the-knearestneighbors-algorithm-6a6e71d01761 [Accessed on July 02, 2019].

[2] Cherif, W. (2018). Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis. Procedia Computer Science, vol. 127, pp. 293–299.

[3] Rehman Amjad et al., "Lung cancer detection and classification from chest CT scans using machine learning techniques", 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA).