A Review on Lung Cancer Detection using Machine Learning

A PROJECT REPORT

Submitted By

Vanshita Gupta
(2000290140130)
Nitin Goyal
(2000290140079)

Submitted in partial fulfillment of the Requirements for the Degree of

MASTER OF COMPUTER APPLICATIONS

Under the Supervision of
Vidushi Mishra

Submitted to

DEPARTMENT OF COMPUTER APPLICATIONS
KIET Group of Institutions, Ghaziabad
Uttar Pradesh-201206(JUNE 2022)

CERTIFICATE

Certified that **Vanshita Gupta (2000290140130), Nitin Goyal (2000290140130)** have carried out the project work having "A Review on Lung Cancer using Machine Learning" for Master of Computer Applications from Dr. A.P.J. Abdul Kalam Technical University (AKTU) (formerly UPTU), Technical University, Lucknow under my supervision. The project report embodies original work, and studies are carried out by the student himself / herself and the contents of the project report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Date:

Vanshita Gupta (2000290140130)

Nitin Goyal (2000290140079)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

Vidushi Mishra
Assistant Professor
Department of Computer Applications
KIET Group of Institutions, Ghaziabad

Signature of Internal Examiner **Signature of External Examiner**

Dr. Ajay Shrivastava
Head, Department of Computer Applications
KIET Group of Institutions, Ghaziabad

ABSTRACT

Lung cancer is a deadly disease that is frequently caused by genetic abnormalities and other pathological alterations. Lung cancer claims a significant number of lives and is particularly difficult to detect. Lung cancer is detected using a variety of procedures, including CT scan images, X-ray imaging, and others. This research aims to investigate the accuracy levels of a variety of machine learning algorithms and give a comparative study of several machine learning techniques for lung cancer detection. The numerous models employed by researchers were set out in order to determine the accuracy levels of several algorithms, and they have a few limits and drawbacks that were stated out. We discovered that some of the algorithms and models have a low degree of accuracy while others have a high level of accuracy after examining them. We came to this conclusion based on the level of accuracy and similarity of techniques used by most researchers, and we are also using a machine-learning algorithm to predict those people who are likely to develop lung cancer disease based on different parameters such as their age, gender, alcohol use, genetic risk, and smoking.

ACKNOWLEDGEMENTS

Success in life is never attained single handedly. My deepest gratitude goes to my thesis supervisor, **Vidushi Mishra** for his guidance, help and encouragement throughout my research work. Their enlightening ideas, comments, and suggestions.

Words are not enough to express my gratitude to Dr. Ajay Kumar Shrivastava, Professor and Head, Department of Computer Applications, for his insightful comments and administrative help at various occasions.

Fortunately, I have many understanding friends, who have helped me a lot on many critical conditions.

Finally, my sincere thanks go to my family members and all those who have directly and indirectly provided me moral support and other kind of help. Without their support, completion of this work would not have been possible in time. They keep my life filled with enjoyment and happiness.

Vanshita Gupta

Nitin Goyal

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Chapters

CHAPTER 1

INTRODUCTION

1. OVERVIEW

CANCER is the world's largest health problem, with an estimated 13.9 million cases and 8.29 million deaths due to early detection of blemishes. In 2018, there were 1,682,210 new cancer cell cases in the United States, with 595,690 deaths. Lung cancer cells are the most common of all cancer cells, accounting for 158,080 deaths and 224,391 new cases in 2016. The survival rate of lung cancer cells is much lower than that of other cancer cells as a result of the influence of lung cancer cells or otherwise recognition of lung cancer cells at an early stage. According to a factsheet published by the American Cancer Society, the American Lung Association, and the World Health Organization, when cancer cells are discovered at the onset, the survival rate rises to 54.4 percent from 17.7 percent, and also if cancer cells are restricted, the survival rate rises to about 16 percent.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: ***The ability to learn***. Machine learning is actively being used today, perhaps in many more places than one would expect. **Arthur Samuel**, a pioneer in the field of artificial intelligence and computer gaming, coined the term **"Machine Learning"**. He defined machine learning as – **"Field of study that gives computers the capability to learn without being explicitly programmed"**.

1.1 Machine Learning Algorithms

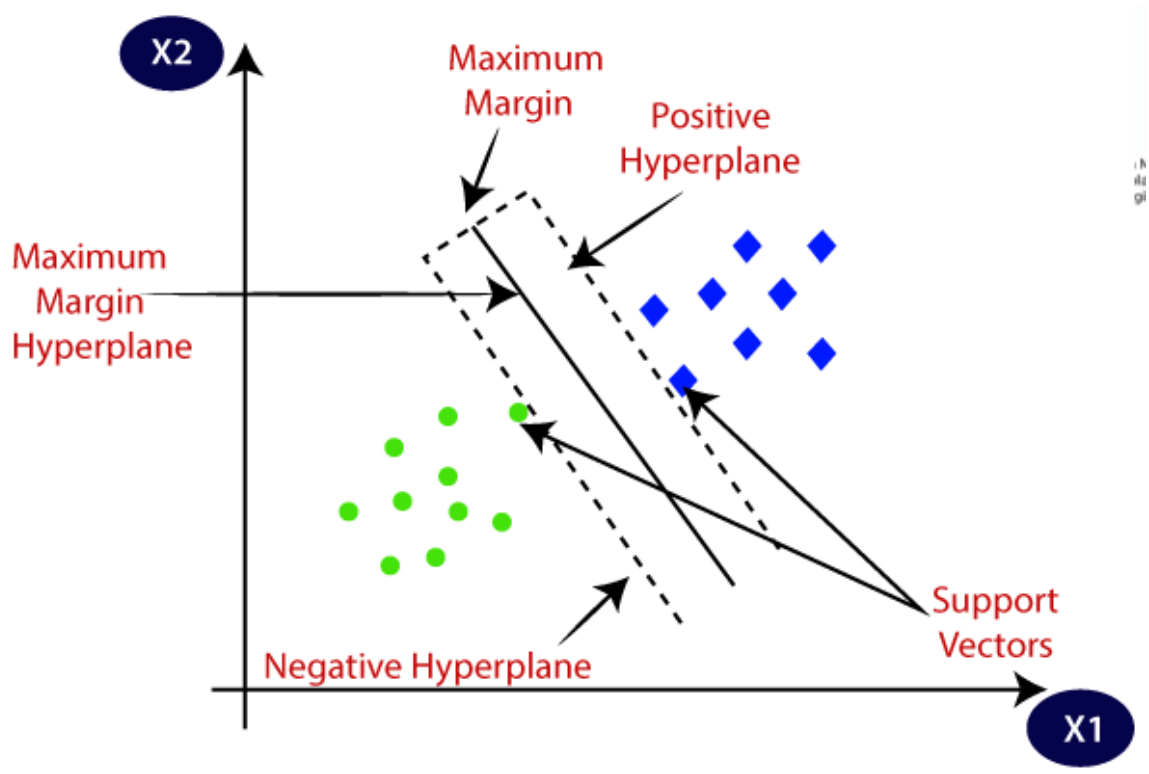
Machine Learning algorithms provides basic and advanced concepts of machine learning. Our machine learning tutorial is designed for students and working professionals.

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information**. Currently, it is being used for various tasks such as **image recognition**, **speech recognition**, **email filtering**, **Facebook auto-tagging**, **recommender system**, and many more.

This machine learning gives you an introduction to machine learning along with the wide range of machine learning techniques such as **Supervised**, **Unsupervised**, and **Reinforcement** learning. You will learn about regression and classification models, clustering methods, hidden Markov models, and various sequential models.

1.1.1 Support Vector Machine (SVM)

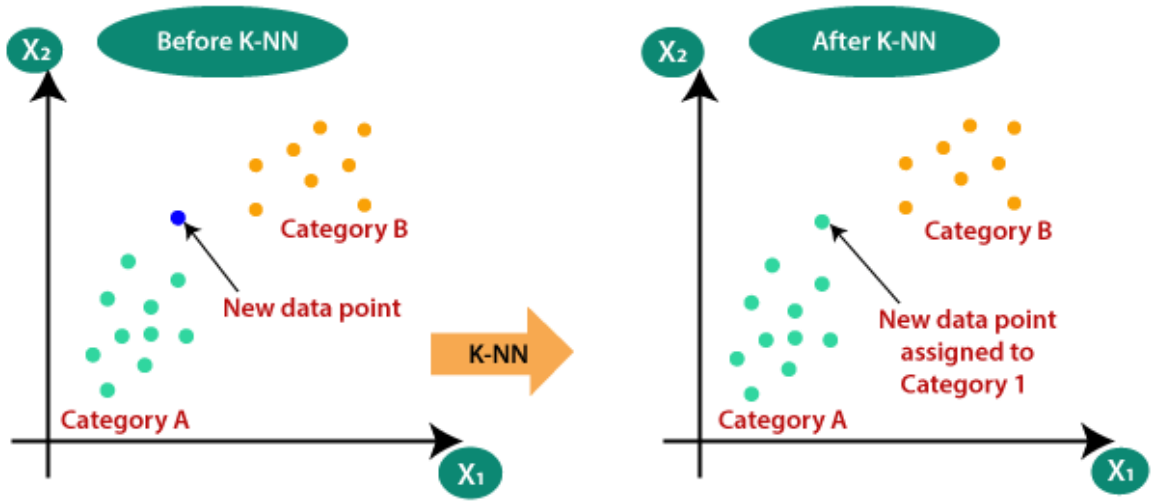
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



1.1.2 K- Nearest Neighbor (KNN)

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action

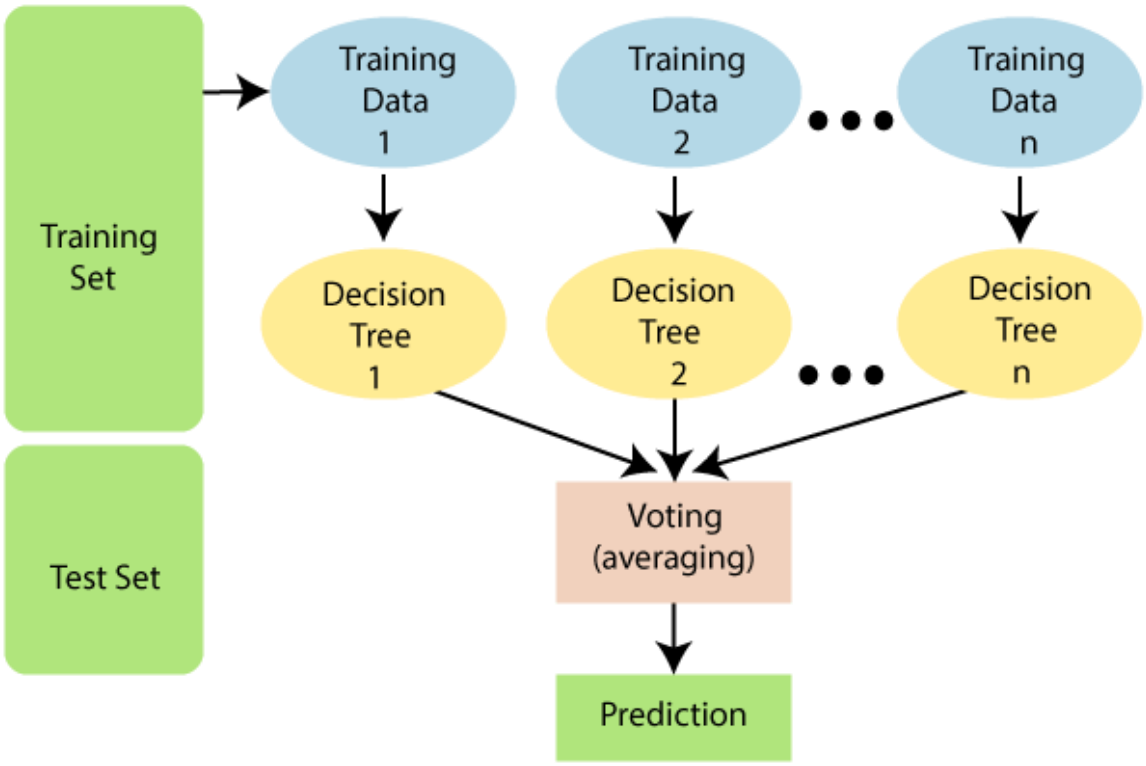
on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



1.1.2 Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.* As the name suggests, *"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."* Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:



1.2 Types of Machine Learning

Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Learning

1.2.1 Supervised Machine Learning

As its name suggests, Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based

on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset. Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the **shape & size of the tail of cat and dog, Shape of eyes, colour, height (dogs are taller, cats are smaller), etc.** After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, colour, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category. This is the process of how the machine identifies the objects in Supervised Learning. **The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).** Some real-world applications of supervised learning are **Risk Assessment, Fraud Detection, Spam filtering,** etc.

Supervised machine learning can be classified into two types of problems, which are given below:

- **Classification**
- **Regression**

Some popular Regression algorithms are given below:

- **Simple Linear Regression Algorithm**
- **Multivariate Regression Algorithm**
- **Decision Tree Algorithm**
- **Lasso Regression**

1.1.2 Unsupervised Machine Learning

Unsupervised learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision. In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision. **The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.** Machines are instructed to find the hidden patterns from the input dataset. Let's take an example to understand it more preciously; suppose there is a basket of fruit images, and we input it into the machine learning model. The images are totally unknown to the model, and the task of the machine is to find the patterns and categories of the objects. So, now the machine will discover its patterns and differences, such as colour difference, shape difference, and predict the output when it is tested with the test dataset.

Unsupervised Learning can be further classified into two types, which are given below:

- **Clustering**
- **Association**

Some of the popular clustering algorithms are given below:

- **K-Means Clustering algorithm**
- **Mean-shift algorithm**
- **DBSCAN Algorithm**
- **Principal Component Analysis**
- **Independent Component Analysis**

1.1.3 Semi-Supervised Learning

Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled

training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period. **A**lthough Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels. **To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced**. The main aim of semi-supervised learning is to effectively use all the available data, rather than only labelled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data. It is because labelled data is a comparatively more expensive acquisition than unlabeled data.

1.1.4 Reinforcement Learning

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance. Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards. In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only. The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and reward. Due to its way of working, reinforcement learning is employed in different fields such as **Game theory, Operation Research, Information theory, multi-agent systems.** A reinforcement learning problem can be formalized using **Markov Decision Process(MDP).** In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.

Reinforcement learning is categorized mainly into two types of methods/algorithms:

- **Positive Reinforcement Learning:** Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding

something. It enhances the strength of the behaviour of the agent and positively impacts it.

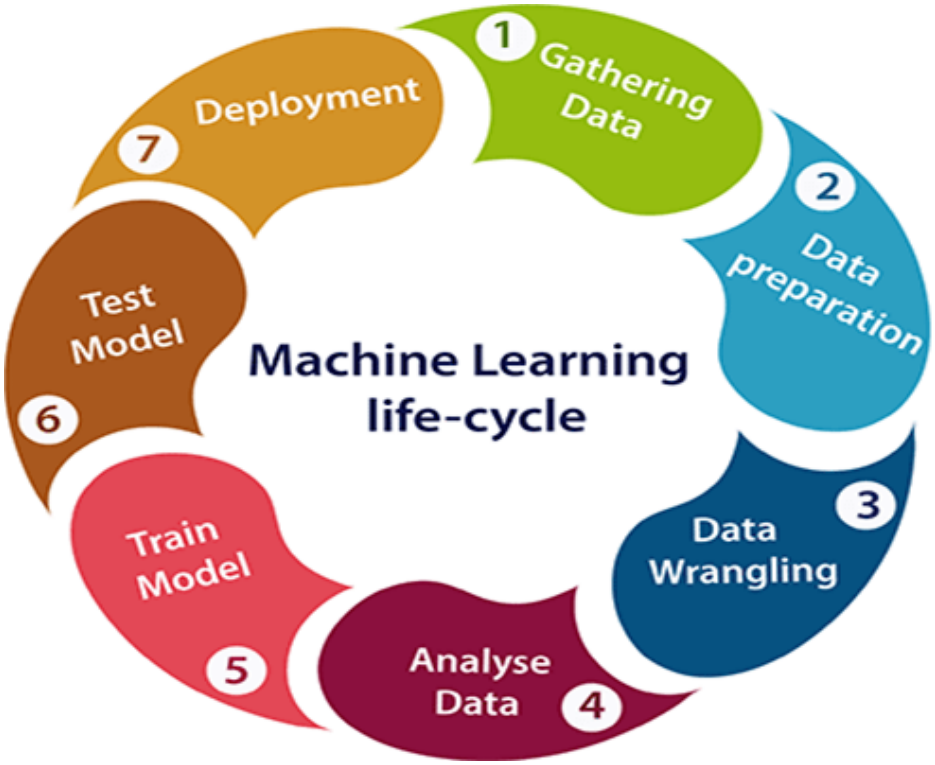
- **Negative Reinforcement Learning:** Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

1.2 Working of Machine Learning

Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed. But how does a machine learning system work? So, it can be described using the life cycle of machine learning. Machine learning life cycle is a cyclic process to build an efficient machine learning project. The main purpose of the life cycle is to find a solution to the problem or project.

Machine learning life cycle involves seven major steps, which are given below:

- **Gathering Data**
- **Data preparation**
- **Data Wrangling**
- **Analyse Data**
- **Train the model**
- **Test the model**
- **Deployment**



The most important thing in the complete process is to understand the problem and to know the purpose of the problem. Therefore, before starting the life cycle, we need to understand the problem because the good result depends on the better understanding of the problem.

1.3 Scope of Research

The Scope of this paper this is to give maximum accuracy results so that prediction of lung cancer can be done using easily

1.4 Research Methodology

By taking a review of lot of research paper , we have made a conclusion based upon maximum accuracy obtained by researchers and algorithms used by them .

CHAPTER 2

Literature Review

To detect lung cancer, Bhatia et al. (2019) employed deep residual learning with CT scans. The scientists devised a pre-processing pathway for highlighting cancer-prone lung areas and extracting characteristics using UNet and ResNet models. To forecast how probable a CT scan is to be malignant, the feature set is given to multiple categories, including XGBoost and Random Forest. The accuracy of this study was 84 percent higher than LIDC-previous IRDI's attempts.

The ability to anticipate which patients are likely to get lung cancer will benefit clinicians in determining treatment decisions. In this regard, (Faisal et al., 2018) attempted to test the discriminative power of several predictors in order to increase the efficacy of lung cancer detection by symptoms in their report. On a benchmark dataset downloaded from the UCI repository, a number of classifiers are tested, including (SVM), Multi-Layer Perceptron, Nave Bayes, C4.5 Decision tree, and Neural Network. The output is also comparable to that of well-known ensembles like Majority Voting and Random Forest. On the basis of performance assessments, Gradient-boosted Tree outperformed other individuals as well as ensemble classifiers, achieving 90% precision.

Classification: During this phase, the retrieved features are passed into a classifier to determine whether they are normal or cancerous. Multi-layer perceptron (MLP), SVM, Nave Bayes, Neural Network, Gradient Boosted Tree, Decision Tree, k-nearest neighbors, multinomial random forest classifier, nave Bayes, stochastic gradient descent, and ensemble classifier have all been utilized by researchers in the literature. (Alam et al., 2018) achieved the greatest accuracy result of around 97 percent using a multi-class SVM classifier and marker-controlled watershed-based segmentation for image segmentation.

As in other fields, researchers have effectively implemented statistical and machine learning techniques in the background of such diseases such as lung cancer, build prediction models.

The detection of lung cancer has previously been considered using techniques of image processing as the work implemented by (Abdillah et al., 2017) in initiation with deep learning and neural networks techniques, domain of medical image has been used recently. Several researchers (Chauhan and Jaiswal, 2016; Nasser and Abu-Naser, 2019; Sharif et al., 2020) have attempted to classify and detect lung cancer by using techniques of classical neural networks and machine learning. Recently, some researchers have tried to use deep learning techniques for lung cancer detection (Alakwaa et al., 2017; Shakeel et al., 2019; Bhatia et al., 2019; Shen et al., 2017; Gao et al., 2018; van and de Bruijne, 2016; Bhandary et al., 2020; Gang et al., 2018; Li et al., 2020; Ausawalaithong et al., 2018).

Using machine learning technique and chemical sensor array, (Huang, et al., 2018) developed a breath test to detect lung cancer. A prospective research to record cases of lung cancer and non-tumor controls between 2016 and 2018 is performed, and 1 alveolar air samples are analyzed using arrays of carbon nanotube sensors. For the model derivation and inner validation, subjects enrolled in 2016 and 2017 were used. In subjects recruited in 2018, the model was validated externally. Using the pathological records as the reference standard, the diagnostic accuracy was evaluated. Using the pathological records as the reference standard, the model was validated externally. then the diagnostic accuracy was evaluated. in the external validation were the areas under the receiver operating characteristic curve (AUCs) 0.91 (95 percent CI = 0.79-1.00) by linear discriminant analysis, and by the technique of the supporting vector machine 0.90 (95 percent CI = 0.80-0.99). The authors concluded that lung cancer could be identified with high precision by incorporating the sensor array technique and machine learning. Chauhan and Jaiswal (2016) suggested the automatic classification of diseases with machine learning based on a practical approach to detecting lung cancer principle. In addition, some benchmark sets showed that the proposed work model was compared to other conventional approaches. The authors have shown that the algorithm proposed is successful over other methods such as SURF and ICA. Nasser and Abu-Naser (2019) have improved an Artificial Neural Network (ANN) model to confirm detection lung cancer, Symptoms such as anxiety, yellow fingers, respiratory illness, exhaustion, allergies, wheezing, coughing, shortness of breath, chest pain and trouble swallowing have been used lung cancer diagnose. The proposed ANN model is developed, educated, and validated using a data set called "survey lung cancer" as input variables for the proposed ANN model and other information about the individual as input variables for the proposed ANN model. The model evaluation revealed that,

with 96.67 percent accuracy, the proposed ANN model was able to detect the absence or presence of lung cancer.

The prediction of lung cancer-prone patients will aid doctors in their treatment decisionmaking. In this respect, in the report, (Faisal et al., 2018) attempted to measure the discriminative capacity of multiple predictors to improve the efficacy of the diagnosis of lung cancer by their symptoms. A variety of classifiers are tested on a benchmark dataset retrieved from the UCI repository, including (SVM), Multi-Layer Perceptron, Naïve Bayes, C4.5 Decision tree, Neural Network, as seen in Figure 3. Familiar ensembles such as Majority Voting and Random Forest are also compared with the output. It is noted that Gradient-boosted Tree exceed other individuals as well as ensemble classifiers on the basis of performance assessments and achieved 90 percent precision.

Wu and Zhao (2017) suggested a new, Entropic Degradation Method (EDM) algorithm for the detection of Small Cell Lung Cancer (SCLC) for Computed Tomography (CT) images. The early diagnosis of lung cancer can be encouraged by this study. The training and test results are lung CT scans given by the National Cancer Institute in high resolution. The authors picked 12 lung CT scans from the library, six for safe lung and 6 for SCLC patients. Then, 5 random scans were taken from each party to practice their model and the remaining two scans were used for research. The suggested algorithm hit 77.8 percent accuracy. Reddy et al. (2019) suggested a structure consisting of various means such as pre handling, image securing, thresholding, binarization, extraction of attributes, division, and recognition of the neural system. The suggested strategy pursues approaches in which parallel thresholding is the initial stage, and then feature extraction, and then these highlights are used to train the fuzzy neural system for machine learning approaches and validate the neural system. From CT scan images, the proposed system accurately defines the lung condition. The proposed framework tested 150 types of pulmonary CT images and achieved the result where 96.67 percent of the framework's overall achievement rate met the framework desire. The significance of radiomics in the prediction of pathological stages of non-small cell lung cancer (NSCLC) has been studied for the first time by (Yu et al., 2019). Nine optimal image characteristics were identified as predictive and diagnostic biomarkers for NSCLC pathological phases. The prediction model has been verified using various machine learning algorithms to efficiently predict the tumor stages of NSCLC, particularly for lung adenocarcinoma (LUAD). The results not only extend the use of machine learning algorithms in the prediction of CT image features for pathological

staging, but also recognizing possibility imaging biomarkers that can be used in NSCLC for diagnosis and prediction of the pathological stage. Singh and Gupta, (2019) have demonstrated an effective approach to the classification and detection of lung cancer-linked CT scan images into malignant and benign categories. First, these images are processed using techniques of image processing in the proposed approach, and then for their classification are used supervised learning algorithms. Here, along with statistical features, texture characteristics are extracted and different extracted characteristics are supplied to classifiers. In addition, are used for seven different classifiers. Multinomial naive Bayes classifier, support vector machine classifier, k-nearest neighbors' classifier, decision tree, stochastic gradient descent classifier, multi-layer perceptron (MLP) classifier and random forest classifier. In addition, 15750 dataset clinical images are used to train and test these classifiers, consisting of both 8840 malignant and 6910 benign images related lung cancer. In the results obtained, the accuracy of the MLP classifier is higher compared to the other classifiers, with a value of 88.55 percent.

Makaju et for pathological staging, but also recognizing possibility imaging biomarkers that can be used in NSCLC for diagnosis and prediction of the pathological stage. Singh and Gupta, (2019) have demonstrated an effective approach to the classification and detection of lung cancer-linked CT scan images into malignant and benign categories. First, these images are processed using techniques of image processing in the proposed approach, and then for their classification are used supervised learning algorithms. Here, along with statistical features, texture characteristics are extracted and different extracted characteristics are supplied to classifiers. In addition, are used for seven different classifiers. Multinomial naive Bayes classifier, support vector machine classifier, k-nearest neighbors' classifier, decision tree, stochastic gradient descent classifier, multi-layer perceptron (MLP) classifier and random forest classifier. In addition, 15750 dataset clinical images are used to train and test these classifiers, consisting of both 8840 malignant and 6910 benign images related lung cancer. In the results obtained, the accuracy of the MLP classifier is higher compared to the other classifiers, with a value of 88.55 percent.

Author & Year	Technique applied	Accuracy	Pitfall
S.Bhatia, Lavika Goel, Yash Sinha & 2019	Deep Residual learning , Random Forest and XGBoost classifier.	84%	Lack of Large dataset, segmentation of lung structures due to homogeneity.
Busi Reddy Venkata Raman Reddy & 2019	Fuzzy NN, Binarization, Solid Component extraction Technique.	96.67%	Can't find bosom growth, Generally utilizes part of tumor only.
P.W.C Prasad, A.K Singh, Suren Makaju & 2018	Water Shed Segmentation for detection and SVM for Classification of nodule e as Malignant or benign.	92%	Not classifies the Stage like I, II, III, and IV of Cancer.
Faisal et al & 2018	A number of classifiers including MLP, Neural Network, Decision Tree, Naïve Bayes, Gradient Boosted Tree, and SVM are assessed.	90%	pre-processing for data cleaning.
Janee Alam, Sabrina Alam, Alamgir Hosan & 2018	multi-class SVM (Support Vector Machine) classifier, Threshold and marker-controlled watershed segmentation.	97 %	Requires large image set.
Waffa Alakwa, Mohd. Naseef & 2017	Deep learning algorithms, CNN, DBNs, SDAE and CAD System.	86%	only determine whether or not the patient has cancer, not the location.
Manikandan Thyagranjan & 2016	A kernel-based SVM classifier, FACMM Algorithm.	93%	No significant scope in measuring very small dot-like tissue clusters, time for segmentation is more.

Wafaa Alakwaa, Mohammad Nassef, Amr Badr & 2017	3D Convolutional Neural Networks (CNNs).	The coarseness/size of the scan differs from scan to scan which hurt the performance of the model and the average error increases with iteration.	86.6%.

CHAPTER 3

Technology used

There are many different types of technology, and each one has unique functions. The technology which we used are mentioned below:-

- Python
- Anaconda
- Numpy
- Panda
- Seaborn
- Matplotlib

3.1 PYTHON

Python is a widely-used, interpreted, object-oriented, and high-level programming language with dynamic semantics, used for general-purpose programming and you may know the python as a large snake, the name of the Python programming language comes from an old BBC television comedy sketch series called **Monty Python's Flying Circus**. Python was created by **Guido van Rossum**, born in 1956 in Haarlem, the Netherlands.

- Python isn't a young language anymore. It is **mature and trustworthy**. It's not a one-hit wonder. It's a bright star in the programming firmament, and time spent learning Pythonit's **easy to learn** - the time needed to learn Python is shorter than for many other languages; this means that it's possible to start the actual programming faster;
- it's **easy to teach** - the teaching workload is smaller than that needed by other languages; this means that the teacher can put more emphasis on general (language-independent) programming techniques, not wasting energy on exotic tricks, strange exceptions and incomprehensible rules;
- it's **easy to use** for writing new software - it's often possible to write code faster when using Python;
- it's **easy to understand** - it's also often easier to understand someone else's code faster if it is written in Python;
- it's **easy to obtain, install and deploy** - Python is free, open and multiplatform; not all languages can boast that.

Of course, Python has its drawbacks, too:

- it's not a speed demon - Python does not deliver exceptional performance;
- In some cases it may be resistant to some simpler testing techniques - this may mean that debugging Python's code can be more difficult than with other languages; fortunately, making mistakes is always harder in Python.

Where can we see Python in action?

We see it every day and almost everywhere. It's used extensively to implement complex Internet services like search engines, cloud storage and tools, social media and so on. Whenever you use any of these services, you are actually very close to Python, although you wouldn't know it.

Many developing tools are implemented in Python. More and more everyday use applications are being written in Python. Lots of scientists have abandoned expensive proprietary tools and switched to Python. Lots of IT project testers have started using Python to carry out repeatable test procedures. The list is long.

Python has two direct competitors, with comparable properties and predispositions. These are:

- **Perl** - a scripting language originally authored by Larry Wall;
- **Ruby** - a scripting language originally authored by Yukihiro Matsumoto.

Python Versions

There are two main kinds of Python, called Python 2 and Python 3.

Python 2 is an older version of the original Python.

Python 3 is the newer version of the language.

If you're going to start a new Python project, **you should use Python 3.**

In addition to Python 2 and Python 3, there is more than one version of each.

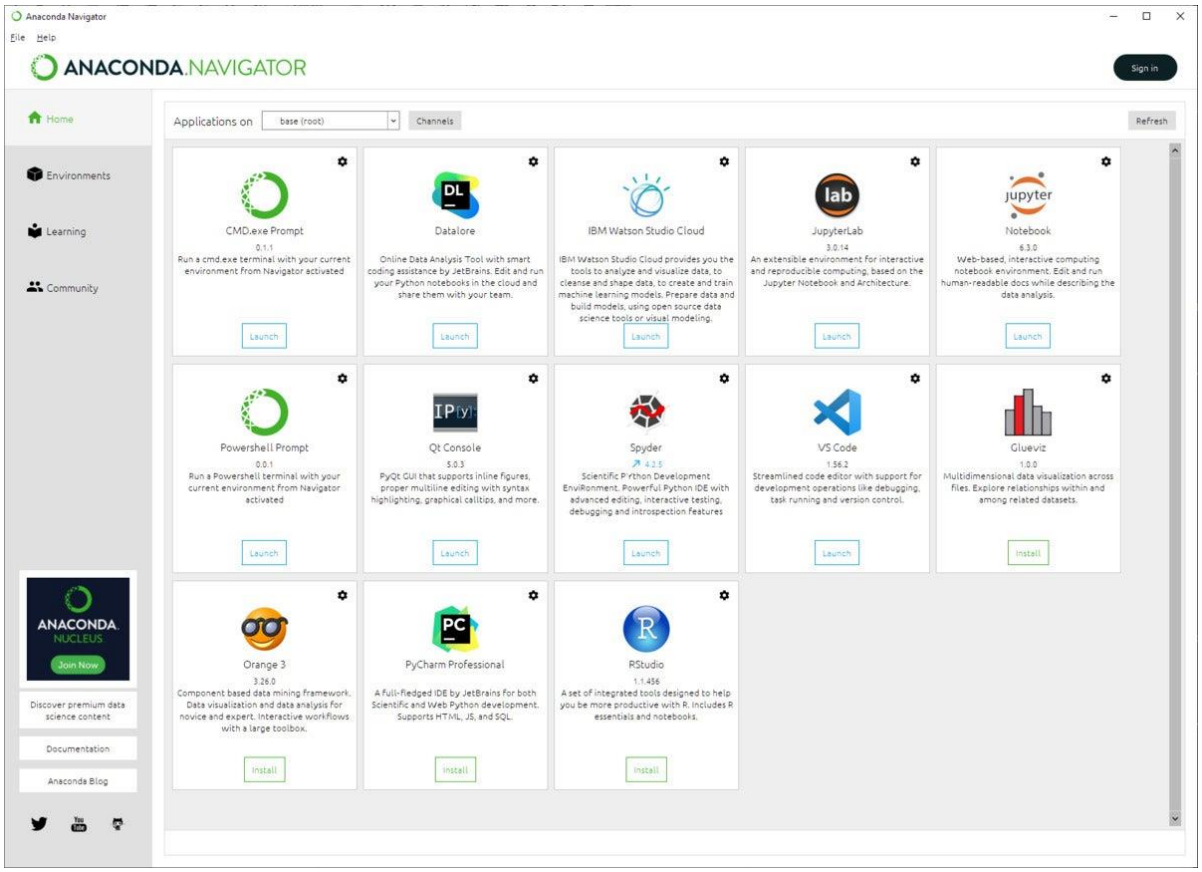
Another Python family member is **Cython**.

Cython automatically translate the Python code (clean and clear) into "C" code (complicated and talkative).

3.2 ANACONDA

Anaconda is an open-source distribution for python and R. It is used for **data science, machine learning, deep learning, etc**. With the availability of more than 300 libraries for data science, it becomes fairly optimal for any programmer to work on anaconda for data science.
Anaconda helps in simplified package management and deployment.
Anaconda comes with a wide variety of tools to easily collect data from various sources using various machine learning and AI algorithms. It helps in getting an easily manageable environment setup which can deploy any project with the click of a single button.
To install anaconda go to https://www.anaconda.com/distribution/
Choose a version suitable for you and click on download. Once you complete the download, open the setup.
Follow the instructions in the setup. Don’t forget to click on add anaconda to my path environment variable. After the installation is complete, you will get a window.

After finishing the installation, open anaconda prompt and type jupyter notebook and check libraries are installed or not.



3.2.1 Numpy

NumPy is a Python library used for working with arrays.

It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.

NumPy stands for Numerical Python.

In Python we have lists that serve the purpose of arrays, but they are slow to process.

NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.
The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.

Arrays are very frequently used in data science, where speed and resources are very important

NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.

Once NumPy is installed, import it in your applications by adding the import keyword and NumPy is usually imported under the np alias.

3.2.2 Pandas

Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy , which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions like Active State Active Python.

Pandas makes it simple to do many of the time consuming, repetitive tasks associated with working with data, including:

- Data cleansing
- Data fill
- Data normalization
- Merges and joins
- Data visualization
- Statistical analysis
- Data inspection
- Loading and saving data

3.2.3 Seaborn

Seaborn is one of an amazing library for visualization of the graphical statistical plotting in Python. Seaborn provides many color palettes and defaults beautiful styles to make the creation of many statistical plots in Python more attractive. Seaborn library aims to make a more

attractive visualization of the central part of understanding and exploring data. It is built on the core of the matplotlib library and also provides dataset-oriented APIs.

Seaborn is also closely integrated with the Panda's data structures, and with this, we can easily jump between the various different visual representations for a given variable to better understand the provided dataset.
Plots are generally used to make visualization of the relationships between the given variables. These variables can either be a category like a group, division, or class or can be completely numerical variables. There are various different categories of plots that we can create using the seaborn library.

- **Distribution plots:** This type of plot is used for examining both types of distributions, i.e., univariate and bivariate distribution.
- **Relational plots:** This type of plot is used to understand the relation between the two given variables.
- **Regression plots:** Regression plots in the seaborn library are primarily intended to add an additional visual guide that will help to emphasize dataset patterns during the analysis of exploratory data.
- **Categorical plots:** The categorical plots are used to deals with categories of variables and how we can visualize them.
- **Multi-plot grids:** The multi-plot grids are also a type of plot that is a useful approach is to draw multiple instances for the same plot with different subsets of a single dataset.
- **Matrix plots:** The matrix plots are a type of arrays of the scatterplots.

3.2.4 Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

CHAPTER 4

Dataset Description

4.1 DATASET Attributes-

(https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data)

The different attributes of dataset are mentioned below:-

1- **AIR POLLUTION**

The World Health Organization (WHO) has classified air pollution as a human carcinogen (like tobacco smoke, asbestos and arsenic) and is calling it a leading cause of cancer deaths globally specially Lung Cancer. Air pollution is a broad term used to refer to a complex mix of particulates in the air. Air pollution can be caused by transportation, power generation, industrial or agricultural emissions and residential heating and cooking. Air pollutants include by products of fossil fuel combustion, exhaust from motor vehicles and diesel engines, and emissions from power plants and industrial centres.Air pollution causes roughly 1 in 10 cases of lung cancer.

2- **ALCOHOL USE**

Alcohol is a carcinogen suspected of increasing lung cancer risk. Therefore, we prospectively evaluated the relationship between alcohol consumption and lung carcinoma Alcohol has been ranked alongside smoking as a group 1 carcinogen by the International Agency for Research on Cancer.

There are three main ways alcohol can cause cancer:

- **Damage to cells.** When we drink alcohol, our bodies turn it into a chemical called acetaldehyde. Acetaldehyde can cause damage to our cells and can also stop the cells from repairing this damage.

- **Changes to hormones.** Alcohol can increase the levels of some hormones such as oestrogen and insulin. Hormones are chemical messengers and higher levels can make cells divide more often, which raises the chance that cancer cells will develop.
- **Changes to cells in the mouth and throat**. Alcohol can make cells in the mouth and throat more likely to absorb harmful chemicals. This makes it easier for cancer-causing substances (like those found in cigarette smoke) to get into the cell and cause damage.

3. Dust Allergy

4. Occupation

There are occupation in the lives of people which can increase the chance of getting Lung Cancer such as Occupational agents/activities that are associated with increased risk for lung cancer are: **Mining and usage of asbestos in industry or manufacture** (asbestos cement products, thermal and electrical insulation in construction and shipyard work, brakes, textile industry)

Exposure to beryllium and beryllium oxide (nuclear technology, X-ray and radiation technology, dental applications and as beryllium-copper alloys in the electronics, aerospace technology, automotive)

Industrial use of cadmium [nickel-cadmium (Ni-Cd) batteries is its major use, pigments, coatings and plating in the form of cadmium-alloys, stabilizers for plastics

Nickel-producing industries (mining, milling, smelting, and refining) as well as nickel-using industries (alloys and stainless steel manufacture is its major use, electroplating, welding, grinding and cutting) Workers in the former industries are exposed to insoluble nickel whereas soluble nickel is the predominant exposure in the later;

Coke-ovens workers (coke production) are mainly exposed to polycyclic aromatic hydrocarbons. Increased risk for lung cancer has been proved by some but not all studies.

5. Genetic Risk

Lung cancer have provided evidence for hereditary transmission of lung cancer from one generation to the next generation. Approximately 8% of lung cancers are inherited or occur as

a result of a genetic predisposition genetic factors have also been associated with the risk of lung cancer, particularly among women Indoor cooking is still practiced in some rural areas this practice can be considered an important risk factor for lung cancer among women. It was demonstrated that the risk of lung cancer development was significantly elevated among the first-degree relatives of a lung cancer patient.

6. Chronic Lung Disease

There is increasing evidence linking the two diseases beyond a common etiology. COPD is an independent risk factor for lung carcinoma, particularly for squamous cell carcinoma and lung cancer is up to five times more likely to occur in smokers with airflow obstruction than those with normal lung function . Even excluding factors such as over diagnosis COPD patients still have twice the risk of lung cancer development The high prevalence of lung cancer in COPD suggests that there may be common mechanisms, such as premature aging in the lungs, genetic predispositions to either disease or common pathogenic factors, such as growth factors, activation of intracellular pathways or epigenetics.

7. Balanced diet

A diet rich in fruit and vegetables reduces the incidence of lung cancer by approximately 25%.

8.Obesity

Overweight and obesity can cause changes in the body that help lead to cancer. These changes can include long-lasting inflammation and higher than normal levels of insulin, insulin-like growth factor, and sex hormones. The risk of cancer increases with the more excess weight a person gains and the longer a person is overweight. You can achieve a healthy weight by making healthy lifestyle choices. You can reduce your risk of obesity-related cancer by following healthy eating plan and getting regular physical activity.

9.Smoking

Smoking can cause cancer and then block your body from fighting it:[4]

- Poisons in cigarette smoke can weaken the body’s immune system, making it harder to kill cancer cells. When this happens, cancer cells keep growing without being stopped.

- Poisons in tobacco smoke can damage or change a cell's DNA. DNA is the cell's "instruction manual" that controls a cell's normal growth and function. When DNA is damaged, a cell can begin growing out of control and create a cancer tumor.

- When you breathe in tobacco smoke, thousands of chemicals enter your lungs. Many of these chemicals have the potential to damage the DNA in your lung cells.
- Your body will work to repair the damage that's done by these chemicals, but over time, smoking can cause more damage than your body can heal. Eventually this may lead to the formation of cancer cells.
- Inhaling tobacco smoke can also damage the tiny air sacs called alveoli in your lungs. These microscopic air sacs are the center of your respiratory system's gas exchange. They move oxygen into your blood, and expel carbon dioxide when you exhale.
- Over time, the damage to the alveoli in your lungs can lead to chronic disease

10. Passive smoker

When you breathe it in, the smoke and its toxic chemicals go deep into your lungs and damage the lining. Smoke irritates your airways every time you breathe it in. And it starts working as soon as you inhale.

At first, your body may be able to repair the damage. But if you're around secondhand smoke for a long time, your body may not be able to fix it any longer. Cells in your lung tissue start to act in an abnormal way, and then they form tumors, or cancer.

The longer you're around it, the more likely you are to get cancer. It's hard to say exactly how much. Many things, like the mix of chemicals you breathe in and how strong they are, play a role. But if you live with a smoker, your odds for lung cancer go up 20%-30%.

11. Chest Pain

When lung cancer causes chest pain, the discomfort may result from enlarged lymph nodes or metastatis to the chest wall, the lining around the lungs (called pleura), or the ribs.

Lung cancer that has spread to your bones may produce pain in your back or in other areas of your body. Bone Pain is often worse at night and increases with movement.

12.Coughing

However, some characteristics of a cough are linked to the risk of lung cancer, includingTrusted Source:

- a persistent cough that does not go away, or that gets worse over time
- coughing up blood or brown or red mucous
- cough in a person with a history of smoking
- frequent respiratory infections such as pneumonia and bronchitis
- cough associated with wheezing or shortness of breath
- a cough that producesTrusted Source lots of thin mucous

13.Fatigue

In lung cancer specifically, reports vary from 37% to 78% of patients undergoing treatment who experience fatigue, and **lung cancer is linked to fatigue lasting more than 6 months**. Fatigue can also continue for months or years after treatment is concluded.

14.Weight Loss

Unexplained rapid weight loss can be the sign of cancer or other health problems. It is recommends that you see your doctor if you lose more than 5 percent of your total body weight in six months to a year. To put this into perspective: If you weigh 160 pounds, 5 percent of your body weight is 8 pounds.An unexplained weight loss of 10 pounds or more could be the first sign of cancer. The types of cancer often identified with this type of weight loss include cancers of the lung

`15.Shortness

Shortness of breath happens when you are not taking in enough oxygen and your lungs try to draw in more air to make up for it.

Difficulty breathing is called dyspnoea (pronounced dis-nee-a).

Between 5 and 7 out of every 10 cancer patients (50% to 70%) have this symptom at some time during their illness. This figure rises to 9 out of 10 (90%) for people who have advanced lung cancer.

You are more likely to have breathing problems if you have:

- lung cancer
- mesothelioma
- cancer that has spread to the lung

Other types of cancer can also cause breathing difficulties.

Being short of breath can be very uncomfortable and frightening. It can make you feel very anxious and panicky, which often makes it even harder to catch your breath.

People with cancer can become breathless for many different reasons. Once your doctor finds out the cause of your breathing problems there is usually a type of treatment that will help you.

16. Wheezing

Wheezing (whistling noise while breathing) is one of the most common symptoms experienced in lung cancer patients, particularly those with late stage, or advanced, disease. Estimates range that 40-85% of patients with lung cancer experience respiratory symptoms such as coughing, shortness-shortness of breath, wheezing, and coughing up blood.

Wheezing may be caused by pulmonary (related to the lungs) or cardiac (related to the heart) factors. Wheezing is a common cause of dyspnea, which is shortness of breath or difficulty breathing.

17. Swallowing Difficulty

Dysphagia is difficulty swallowing — taking more time and effort to move food or liquid from your mouth to your stomach. Dysphagia can be painful. In some cases, swallowing is impossible.

Occasional difficulty swallowing, such as when you eat too fast or don't chew your food well enough, usually isn't cause for concern. But persistent dysphagia can be a serious medical condition requiring treatment.

18.Dry coughing

A dry or unproductive cough doesn’t produce mucus. A tickling sensation in the throat can make you have a dry cough. Dry coughs can come on after a cold or flu or if you have COVID-19. Other conditions like GERD, heart failure and lung cancer can cause chronic dry coughs. You may also have chest tightness with a dry cough.

19.Snoring

Snoring is the hoarse or harsh sound that occurs when air flows past relaxed tissues in your throat, causing the tissues to vibrate as you breathe. Nearly everyone snores now and then, but for some people it can be a chronic problem. Sometimes it may also indicate a serious health condition. In addition, snoring can be a nuisance to your partner.

Lifestyle changes, such as losing weight, avoiding alcohol close to bedtime or sleeping on your side, can help stop snoring.

In addition, medical devices and surgery are available that may reduce disruptive snoring. However, these aren't suitable or necessary for everyone who snores.

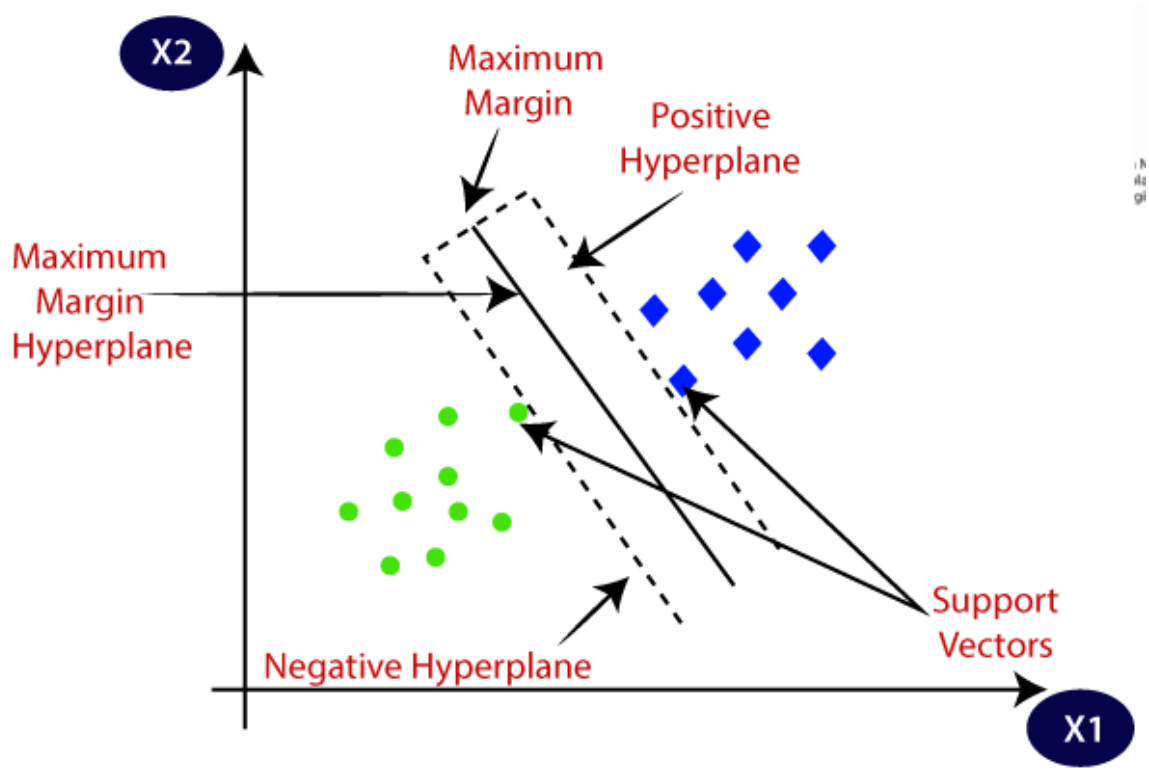
CHAPTER 5

Support Vector Machine

5.1 Support Vector Machine

Support Vector Machines are the supervised machine learning models which are used both for the Classification and Regression problems. The support vector machines are way different from the models like Logistic Regression, Naive Bayes etc. The key idea that SVM uses is to find the hyperplane which maximizes the margin and keep positive and negative class points as wide as possible.

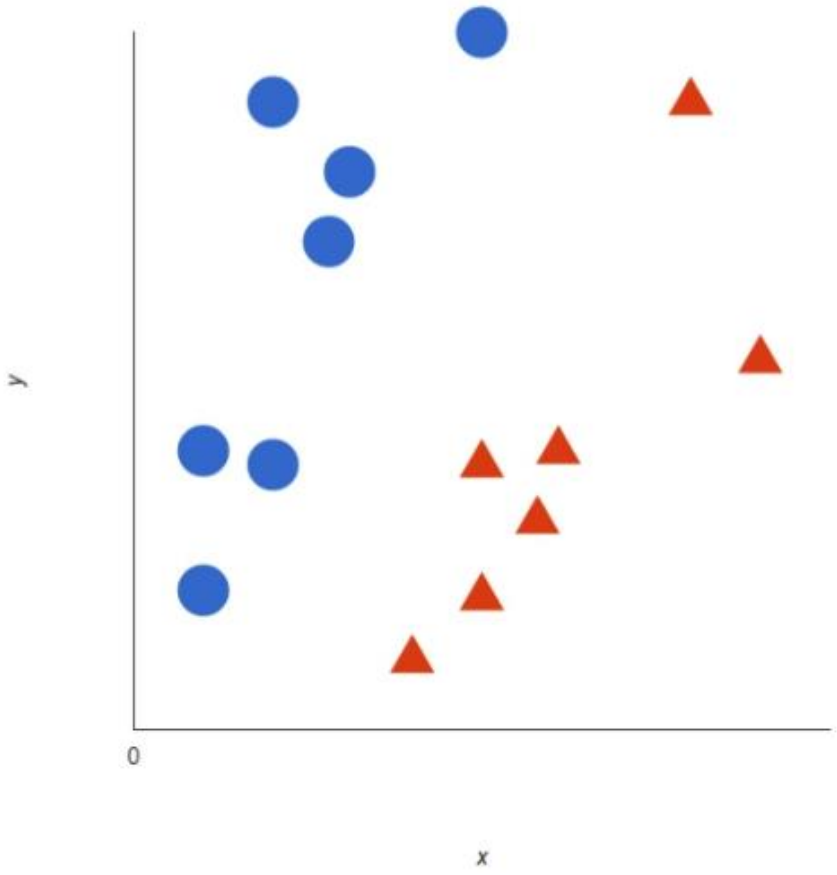
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



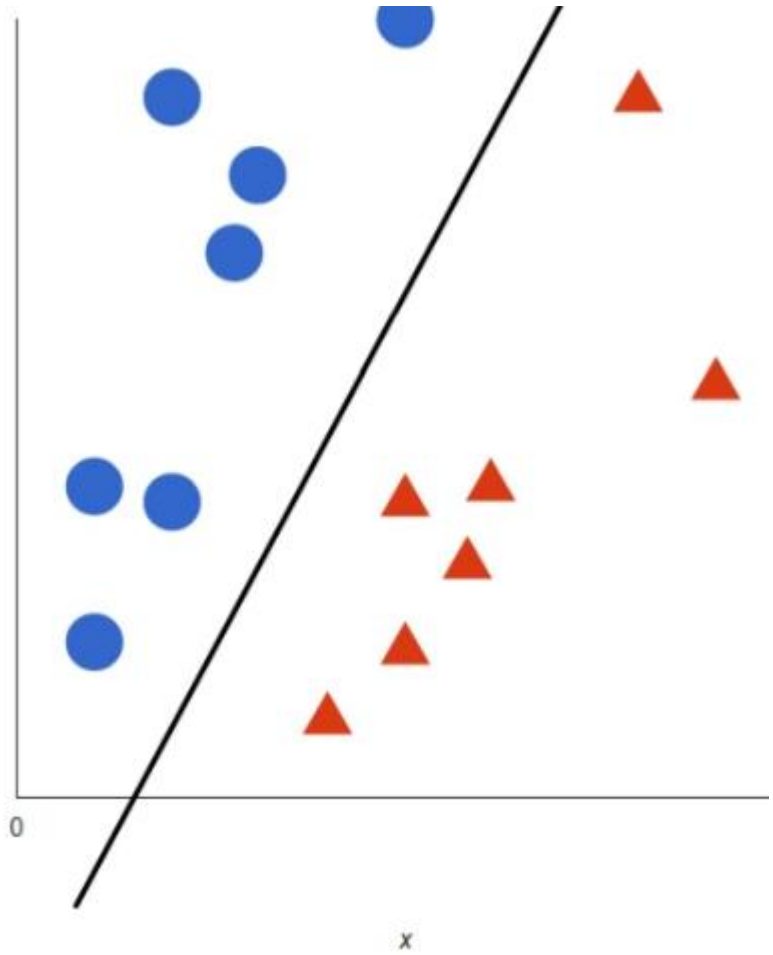
Example: SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat.

5.1 Working of SVM

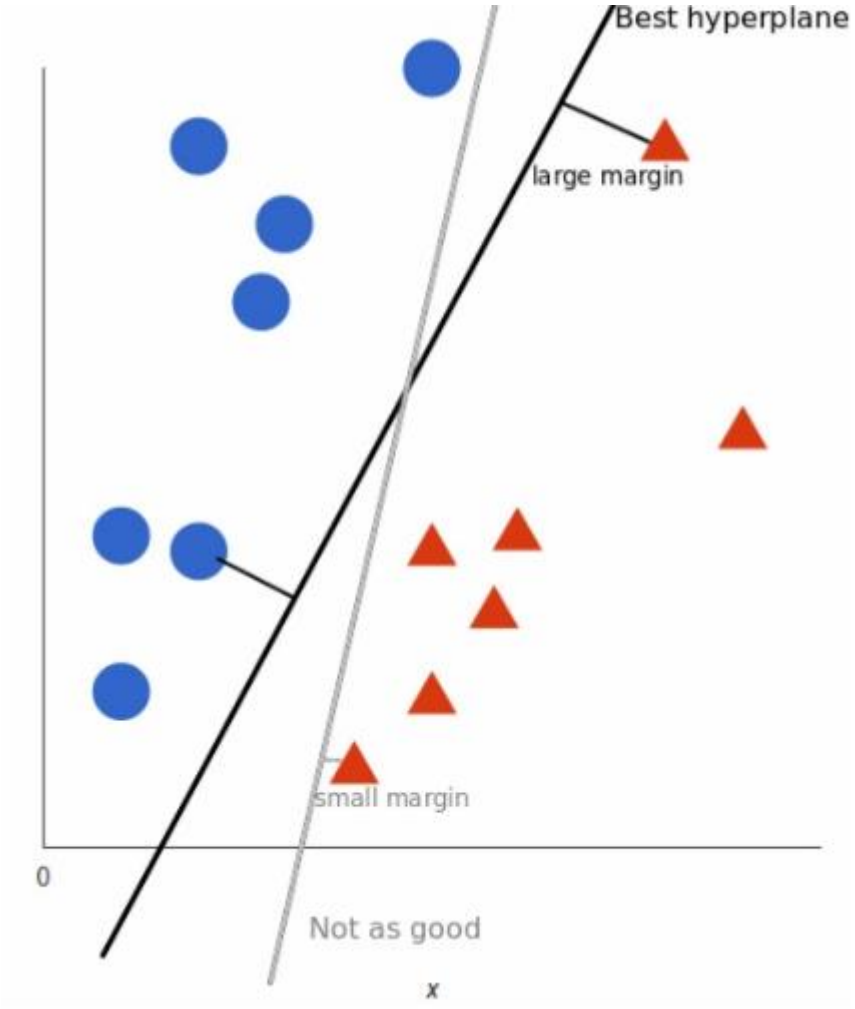
The basics of Support Vector Machines and how it works are best understood with a simple example. Let's imagine we have two tags: *red* and *blue*, and our data has two features: *x* and *y*. We want a classifier that, given a pair of *(x,y)* coordinates, outputs if it's either *red* or *blue*. We plot our already labeled training data on a plane:



A support vector machine takes these data points and outputs the hyperplane (which in two dimensions it's simply a line) that best separates the tags. This line is the **decision boundary**: anything that falls to one side of it we will classify as *blue*, and anything that falls to the other as *red*.



But what exactly is *the best* hyperplane? For SVM, it’s the one that maximizes the margins from both tags. In other words: the hyperplane remember it's a line in this case whose distance to the nearest element of each tag is the largest.



5.2 Types of SVM

Linear SVM

Linear SVM is used for data that are linearly separable i.e. for a dataset that can be categorized into two categories by utilizing a single straight line. Such data points are termed as linearly separable data, and the classifier is used described as a Linear SVM classifier.

Non-linear SVM

Non-Linear SVM is used for data that are non-linearly separable data i.e a straight line cannot be used to classify the dataset. For this, we use something known as a kernel trick that sets data points in a higher dimension where they can be separated using planes or other mathematical

functions. Such data points are termed as non-linear data, and the classifier used is termed as a Non-linear SVM classifier.

5.3 HYPERPLANE

A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts.

For example let's assume a line to be our one dimensional Euclidean space (i.e. let's say our datasets lie on a line). Now pick a point on the line, this point divides the line into two parts. The line has 1 dimension, while the point has 0 dimensions. So a point is a hyperplane of the line.

For two dimensions we saw that the separating line was the hyperplane. Similarly, for three dimensions a plane with two dimensions divides the 3d space into two parts and thus act as a hyperplane. Thus for a space of n dimensions we have a hyperplane of n-1 dimensions separating it into two parts.

5.4 SUPPORT VECTOR

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

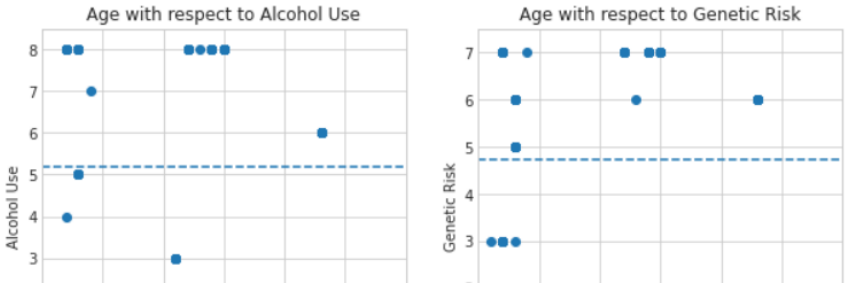
CHAPTER 6

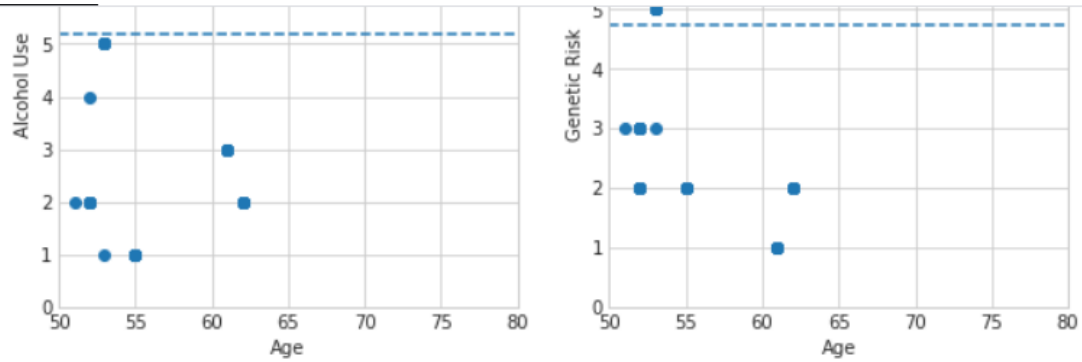
Coding

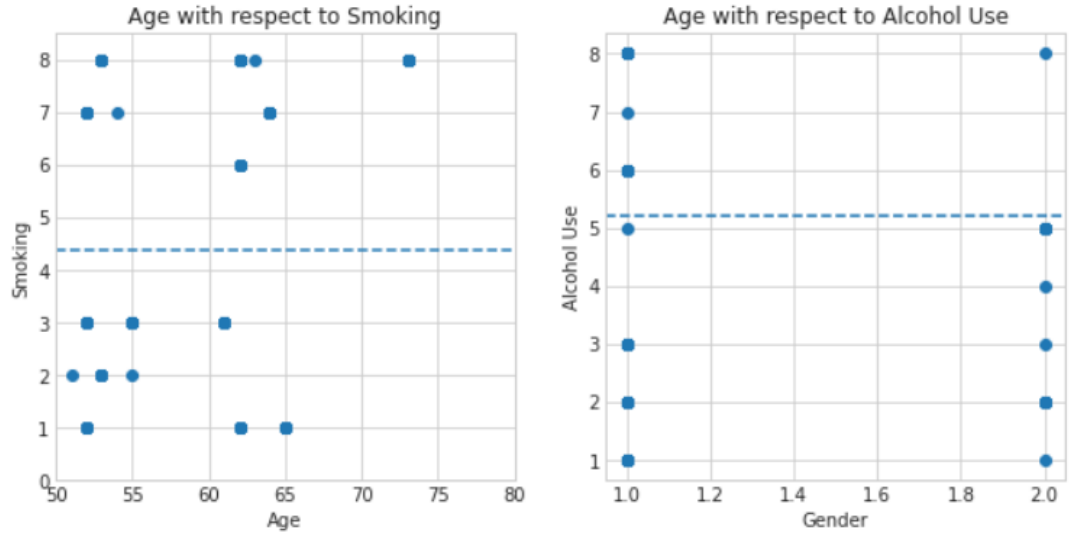
In [11]:

```
# Making Subplots
fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(nrows = 2, ncols= 2, figsize=(10, 10))

# Adding Data to the plot
scatter = ax1.scatter(x = cancer_over50["Age"], y = cancer_over50["Alcohol use"], cmap = "wi
nter")

# For Plot ax1
ax1.set(title = "Age with respect to Alcohol Use",
        xlabel = "Age",
        ylabel = "Alcohol Use")
ax1.axhline(cancer_over50["Alcohol use"].mean(),
           linestyle = "--");
ax1.set_xlim([50, 80])
ax1.set_ylim([0, 8.5])

# For Plot ax2
scatter = ax2.scatter(x = cancer_over50["Age"], y = cancer_over50["Genetic Risk"])
ax2.set(title = "Age with respect to Genetic Risk", xlabel = "Age", ylabel = "Genetic Risk")
ax2.axhline(cancer_over50["Genetic Risk"].mean(),
           linestyle = "--");
```

```
# For Plot ax3
scatter = ax3.scatter(x = cancer_over50["Age"], y = cancer_over50["Smoking"])
ax3.set(title = "Age with respect to Smoking", xlabel = "Age", ylabel = "Smoking")
ax3.axhline(cancer_over50["Smoking"].mean(),
           linestyle = "--");
ax3.set_xlim([50, 80])
ax3.set_ylim([0, 8.5])

# For Plot ax4
scatter = ax4.scatter(x = cancer_over50["Gender"], y = cancer_over50["Alcohol use"])
ax4.set(title = "Age with respect to Alcohol Use", xlabel = "Gender", ylabel = "Alcohol Us
e")
ax4.axhline(cancer_over50["Alcohol use"].mean(),
           linestyle = "--");
```

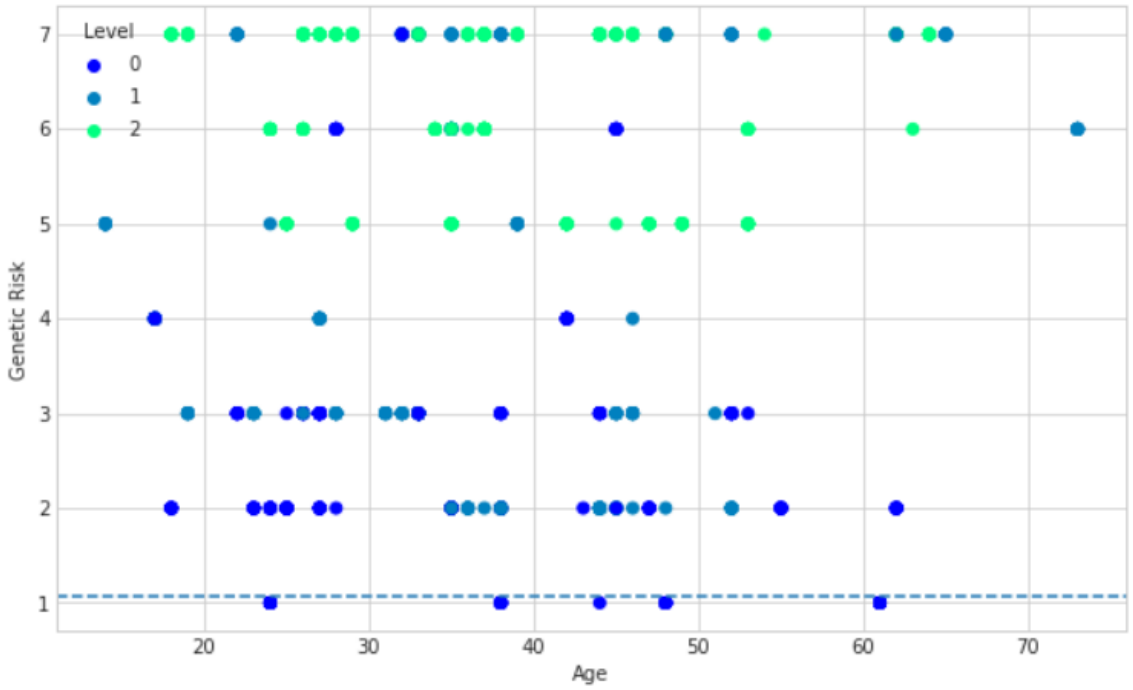






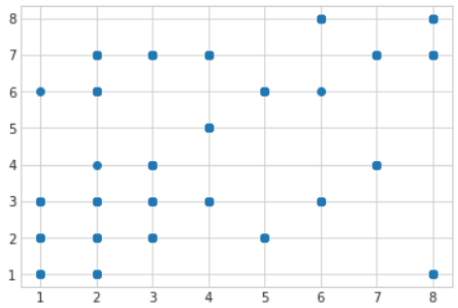
In [20]:

```
fig, ax = plt.subplots(figsize = (10, 6));

scatter = ax.scatter(x = cancer_patient["Age"],
                     y = cancer_patient["Genetic Risk"],
                     c = cancer_patient["Level"],
                     cmap = "winter")

ax.set(xlabel = "Age",
       ylabel = "Genetic Risk");

ax.legend(*scatter.legend_elements(), title = "Level");

ax.axhline(cancer_patient["Level"].mean(),
           linestyle = "--");
```



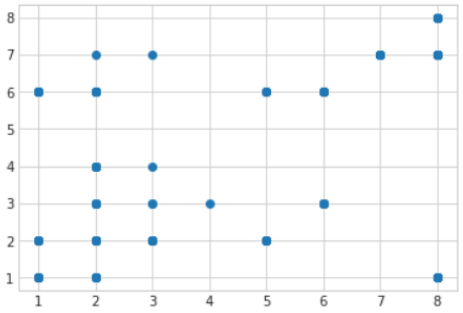
In [30]:

```
fig, ax = plt.subplots()
scatter = ax.scatter(x = cancer_patient_male["Alcohol use"], y = cancer_patient_male["Smokin
g"])
# cancer_patient_male.plot(x = cancer_patient_male["Alcohol use"], y = cancer_patient_male["Ag
e"], kind = "scatter");
```



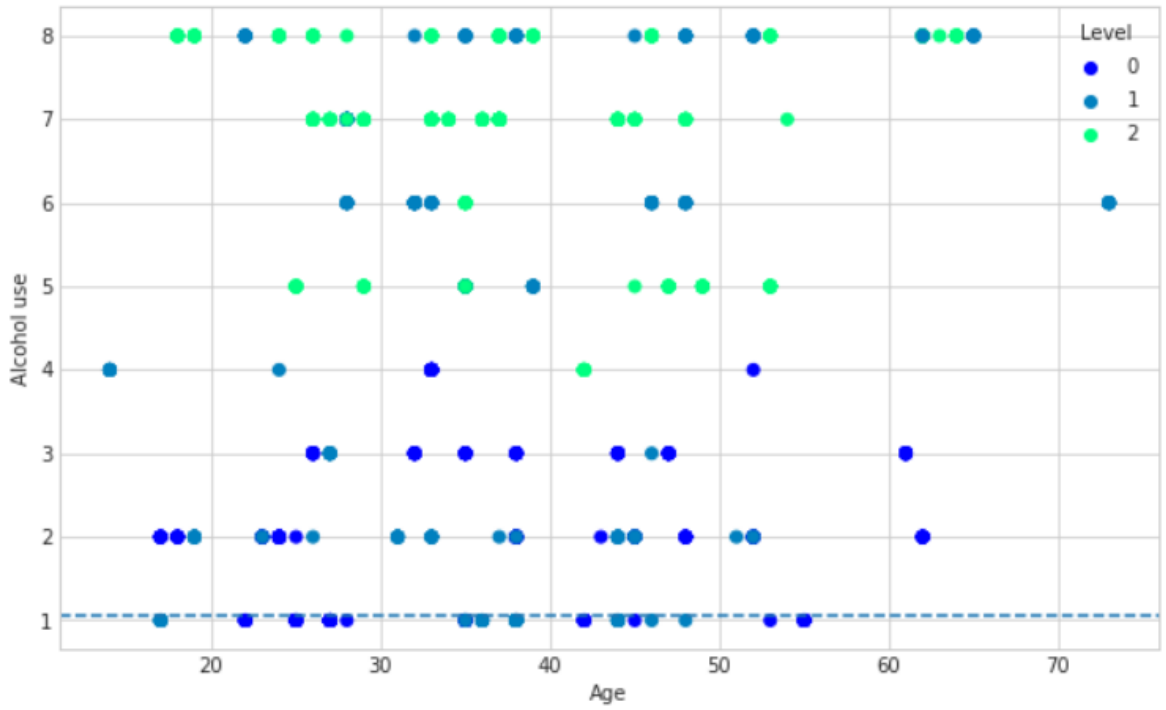
In [31]:

```
fig, ax = plt.subplots()
scatter = ax.scatter(x = cancer_patient_female["Alcohol use"], y = cancer_patient_female["Sm
oking"]);
```



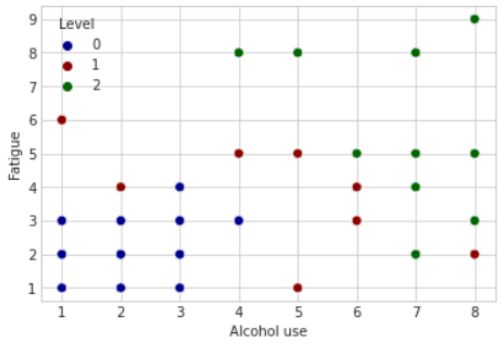
In [36]:

```
fig, ax = plt.subplots(figsize = (10, 6))
scatter = ax.scatter(x = cancer_patient["Age"],
                     y = cancer_patient["Alcohol use"],
                     c = cancer_patient["Level"],
                     cmap = "winter")

ax.set(xlabel = "Age",
       ylabel = "Alcohol use");

ax.legend(*scatter.legend_elements(), title = "Level");

ax.axhline(cancer_patient["Level"].mean(),
           linestyle = "--");
```



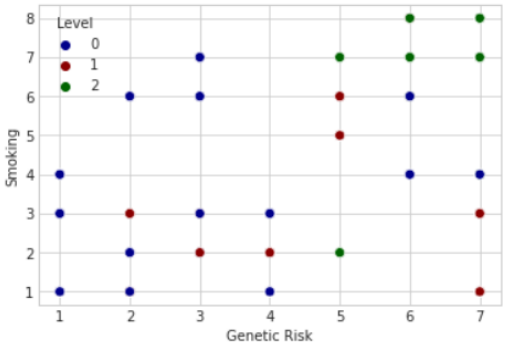
In [37]:

```
fig, ax=plt.subplots()#Required outside of function. This needs to be activated first when pl
otting in every code block
plot=sns.scatterplot(data=cancer_patient,
                     x='Alcohol use',
                     y='Fatigue',
                     hue='Level',
                     palette=['darkblue','darkred','darkgreen'],
                     s=50,
                     marker='o')#Count plot
```



In [38]:

```
fig, ax=plt.subplots()#Required outside of function. This needs to be activated first when pl
otting in every code block
plot=sns.scatterplot(data=cancer_patient,
                     x='Genetic Risk',
                     y='Smoking',
                     hue='Level',
                     palette=['darkblue','darkred','darkgreen'],
                     s=50,
                     marker='o')#Count plot
```



First we use Support Vector Machine Estimator

In [42]:
```
from sklearn import svm
from sklearn.model_selection import train_test_split

X = cancer_patient.drop(["Level"], axis = 1)
y = cancer_patient["Level"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)

sv = svm.SVC()
sv.fit(X_train, y_train)
sv.score(X_test, y_test)
```

Out[42]:
```
0.96
```

Checking accuracy with other model

In [47]:
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split

X = cancer_patient.drop(["Level"], axis = 1)
y = cancer_patient["Level"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)

knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
knn.score(X_test, y_test)
```

Out[47]:
```
1.0
```

Lastly we use RandomForestClassifier

In [48]:

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split

X = cancer_patient.drop(["Level"], axis = 1)
y = cancer_patient["Level"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)

rfr = RandomForestRegressor()
rfr.fit(X_train, y_train)
rfr.score(X_test, y_test)
```

Out[48]:

```
1.0
```

Cross val for all the above algorithms to make sure for scores accuracy

In [49]:

```
from sklearn.model_selection import cross_val_score

crossVal_sv = cross_val_score(sv, X, y)
crossVal_knn = cross_val_score(knn, X, y)
crossVal_rfr = cross_val_score(rfr, X, y)

print(f"For SupportVectorMachine: {crossVal_sv}, \nFor KNeighborClassifier: {crossVal_knn},
\nFor RandomForestRegressor: {crossVal_rfr}")
```

```
For SupportVectorMachine: [0.98  0.975 0.985 0.97  0.97 ],
For KNeighborClassifier: [0.995 1.    1.    1.    0.995],
For RandomForestRegressor: [1. 1. 1. 1. 1.]
```

CHAPTER 7

Conclusion

It is a big advantage to detect lung cancer at an early stage since therapy can then be started to prevent the disease from becoming detrimental. As a result, this work offers a comprehensive review of various machine learning algorithms for classifying lung cancers using CT scans or X-ray pictures. Many classifiers, such as MLP, SVM, Nave Bayes, Neural Network, Gradient Boosted Tree, Decision Tree, k-nearest neighbors, multinomial random forest classifier naive Bayes, stochastic gradient descent, and ensemble classifier, have been utilized by researchers in the literature. As a result, and based on the broad survey conducted in this work, it can be determined that methods utilizing MLImage processing and multi-class SVM (Support Vector Machine) classifier, Threshold and marker controlled watershed segmentation produced higher accuracy results with 97% accuracy.

References

Bhatia, S., Sinha, Y., & Goel, L. (2019). Lung cancer detection: A deep learning approach. In Soft Computing for Problem Solving (pp. 699-705). Springer, Singapore.

Manikandan, T., & Bharathi, N. (2016). Lung cancer detection using fuzzy auto-seed cluster means morphological segmentation and SVM classifier. Journal of medical systems, 40(7), 181.

Alakwaa, W., Nassef, M., & Badr, A. (2017). Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). Lung Cancer, 8(8), 409

Reddy, U., Reddy, B., & Reddy, B. (2019). Recognition of Lung Cancer Using Machine Learning Mechanisms with Fuzzy Neural Networks. Traitement du Signal, 36(1), 87-91.

Makaju, S., Prasad, P. W. C., Alsadoon, A., Singh, A. K., & Elchouemi, A. (2018). Lung cancer detection using CT scan images. Procedia Computer Science, 125, 107-114.

Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018, December). An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST) (pp. 1-4). IEEE.

Alakwaa, W., Nassef, M., & Badr, A. (2017). Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). Lung Cancer, 8(8), 409.

Yu, K. H., Lee, T. L. M., Yen, M. H., Kou, S. C., Rosen, B., Chiang, J. H., & Kohane, I. S. (2020). Reproducible Machine Learning Methods for Lung Cancer Detection Using Computed Tomography Images: Algorithm Development and Validation. Journal of medical Internet research, 22(8), e16709.

Alakwaa, W., Nassef, M., & Badr, A. (2017). Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). Lung Cancer, 8(8), 409.

Jothilakshmi, R., & SV, R. G. (2020) Early Lung Cancer Detection Using Machine Learning And Image Processing.

Bhatia, S., Sinha, Y., & Goel, L. (2019). Lung cancer detection: A deep learning approach. In Soft Computing for Problem Solving (pp. 699-705). Springer, Singapore.

Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2016, August). A review of machine learning techniques using decision tree and support vector machine. In 2016 International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-7). IEEE.

Nasser, I. M., & Abu-Naser, S. S. (2019). Lung Cancer Detection Using Artificial Neural Network. International Journal of Engineering and Information Systems (IJEAIS), 3(3), 17-23.

Ausawalaithong, W., Thirach, A., Marukatat, S., & Wilaiprasitporn, T. (2018, November). Automatic lung cancer prediction from chest X-ray images using the deep learning approach. In 2018 11th Biomedical Engineering International Conference (BMEiCON) (pp. 1-5). IEEE.

Li, X., Shen, L., Xie, X., Huang, S., Xie, Z., Hong, X., & Yu, J. (2020). Multi-resolution convolutional networks for chest X-ray radiograph-based lung nodule detection. Artificial intelligence in medicine, 103, 101744.

Gang, P., Zhen, W., Zeng, W., Gordienko, Y., Kochura, Y., Alienin, O., ... & Stirenko, S. (2018, March). Dimensionality reduction in deep learning for chest X-ray analysis of lung cancer. In 2018 tenth international conference on advanced computational intelligence (ICACI) (pp. 878-883). IEEE.

Bhandary, A., Prabhu, G. A., Rajinikanth, V., Thanaraj, K. P., Satapathy, S. C., Robbins, D. E., ... & Raja, N. S. M. (2020). Deep-learning framework to detect lung abnormality–A study with chest X-Ray and lung CT scan images. Pattern Recognition Letters, 129, 271-278.

Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I. (2020). Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier. Neural Computing and Applications, 1-14.

Tripathi, P., Tyagi, S., & Nath, M. (2019). A comparative analysis of segmentation techniques for lung cancer detection. Pattern Recognition and Image Analysis, 29(1), 167-173.

Abdillah, B., Bustamam, A., & Sarwinda, D. (2017, October). Image processing based detection of lung cancer on CT scan images. In Journal of Physics: Conference Series (Vol. 893, No. 1, p. 012063). IOP Publishing.

Manikandan, T., & Bharathi, N. (2016). Lung cancer detection using fuzzy auto-seed cluster means morphological segmentation and SVM classifier. Journal of medical systems, 40(7), 181.

van Tulder, G., & de Bruijne, M. (2016). Combining generative and discriminative representation learning for lung CT analysis with convolutional restricted Boltzmann machines. IEEE transactions on medical imaging, 35(5), 1262-1272.

Al-jaboriy, S. S., Sjarif, N. N. A., Chuprat, S., & Abduallah, W. M. (2019). Acute lymphoblastic leukemia segmentation using local pixel information. Pattern Recognition Letters, 125, 85-90. Huang, C. H., Zeng, C., Wang, Y. C., Peng, H. Y., Lin, C. S., Chang, C. J., & Yang, H. Y. (2018).

A study of diagnostic accuracy using a chemical sensor array and a machine learning technique to detect lung cancer. Sensors, 18(9), 2845. Chauhan, D., & Jaiswal, V. (2016, October).

An efficient data mining classification approach for detecting lung cancer disease. In 2016 International Conference on Communication and Electronics Systems (ICCES) (pp. 1-8). IEEE.

Nasser, I. M., & Abu-Naser, S. S. (2019). Lung Cancer Detection Using Artificial Neural Network.

International Journal of Engineering and Information Systems (IJEAIS), 3(3), 17-23. Sharif, M. I., Li, J. P., Naz, J., & Rashid, I. (2020).

A comprehensive review on multi-organs tumor detection based on machine learning. Pattern Recognition Letters, 131, 30-37. Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018, December).

An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer.

In 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST) (pp. 1-4). IEEE.

Wu, Q., & Zhao, W. (2017, October). Small-cell lung cancer detection using a supervised machine learning algorithm.

In 2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC) (pp. 88-91). IEEE. Reddy, U., Reddy, B., & Reddy, B. (2019). Recognition of Lung Cancer Using Machine Learning Mechanisms with Fuzzy Neural Networks.

Traitement du Signal, 36(1), 87-91. Yu, L., Tao, G., Zhu, L., Wang, G., Li, Z., Ye, J., & Chen, Q. (2019). Prediction of pathologic stage in nonsmall cell lung cancer using machine learning algorithm based on CT image feature analysis.

BMC cancer, 19(1), 1-12. Singh, G. A. P., & Gupta, P. K. (2019). Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Computing and Applications, 31(10), 6863-6877.

Alam, J., Alam, S., & Hossan, A. (2018, February). Multi-stage lung cancer detection and prediction using multi-class svm classifie.

In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

Makaju, S., Prasad, P. W. C., Alsadoon, A., Singh, A. K., & Elchouemi, A. (2018). Lung cancer detection using CT scan images. Procedia Computer Science, 125, 107-114. Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I. (2019).

Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. Measurement, 145, 702-712.