A Project Report

On

Study of Machine Learning Algorithms on Cancer Research Papers

BY

Kartikay Dhall

2020A7PS2087H

Under the supervision of

Dr. Jabez Christopher

&

Dr. Rajib Ranjan Maiti

SUBMITTED IN THE FULLFILLMENT OF THE REQUIREMENTS OF

CS F266: STUDY PROJECT

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

HYDERABAD CAMPUS

(DECEMBER 2023)

ACKNOWLEDGMENTS

I want to thank Dr. Jabez Christopher and Dr. Rajib Ranjan Maiti for giving me the chance to work on the study of machine learning algorithms on cancer research papers. I believe that working on this project has greatly advanced my knowledge and skills.

Additionally, I want to thank everyone who assisted me in working on this project. Due to my inexperience and ignorance, I encountered several difficulties while working, but I was able to overcome them.

Last but not least, I would like to express my gratitude to my parents and friends for supporting me as I overcame challenges in my life. I also want to thank everyone who has assisted me in working on this project.

Birla Institute of Technology and Science-Pilani,

Hyderabad Campus

Certificate

This is to certify that the project report entitled "**Study of Machine Learning Algorithms on Cancer Research Papers**" submitted by Mr. KARTIKAY DHALL (ID No. 2020A7PS2087H) in the fulfillment of the requirements of the course CS F266 Study Project Course, embodies the work done by him/her under my supervision and guidance.

Date: 03-12-23 **(Dr. Jabez Christopher)**

BITS- Pilani, Hyderabad Campus

ABSTRACT

Cancer presents one of the biggest difficulties in contemporary healthcare, and its treatment necessitates an ongoing quest for knowledge and innovation. In light of this, the goal of this project was to use machine learning techniques to analyze a sizable collection of cancer research papers. The primary objective was to unearth important ideas that might have an impact on cancer research in the future, promote collaboration, and possibly enhance patient outcomes.

The study placed attention on careful data preprocessing in the beginning, which included data cleansing, engineering features, and text data transformation. Following this, many machine learning algorithms were used to model and forecast various elements of cancer research, including Decision Trees, Random Forest, and Support Vector Machines.

The following are the main conclusions and findings from this project: - Identification of prevalent and emerging research trends in the field of cancer research, which provide information about the course of scientific inquiry.

- The identification of areas in need of additional research, which directs the allocation of funds and resources for research to understudied topics.

- Investigating machine learning models that have the ability to predict outcomes, resulting in improvements in cancer diagnosis, treatment regimen design, and personalized medicine tactics.

- Making specific, doable suggestions for multidisciplinary cooperation, creative projects, and future research endeavors.

- Understanding the limitations of studies, such as dataset bias, and the significance of ethical issues when managing medical data.

One of the project's significant findings and conclusions was the identification of existing and new research trends in the field of cancer research, which gave insights into the process of scientific inquiry.

• Predictive Models: Research on predictive machine learning models, promising advancements in the detection and planning of cancer, and potential individualized medicine methods.

• Ethical considerations: awareness of study limitations, such as dataset bias, and the requirement for moral caution while handling patient data.

This study has considerably enhanced the use of data analytics and machine learning in cancer research. The findings demonstrate the potential for interdisciplinary cooperation, data-driven discoveries, and ethical research practices in the field of medicine. We hope that the information gleaned from this study will guide and inspire future research, boosting the global fight against cancer.

CONTENTS

Title page……1

Acknowledgements……..2

Certificate…….....3

Abstract…..……...4

Objectives…….…….....7

Outline…….....9

Procedure……..11

Key Results……...17

Stats on algos…….21

Expected Outcomes……....27

Conclusion……..29

Objectives

1. Advancing Cancer Research: In world cancer continues to be one of the leading causes of death, there remains a need for us to make significant progress in our understanding about this complicated disease. By critically analysing research papers on cancer, we can help in setting up a body of knowledge with regards to this field through which possible means of avoidance, diagnosis and treatment might be invented.

2. Data-Driven Insights: The enormous amount of research data that has been collected empowers us to look for valuable insights and trends in cancer research. With the aid of computational techniques such as machine learning, data analysis we can find trends, co-relations and emerging topics that might not be intuitively conceivable from manual inspection.

3. Spotting Research Gaps: Such an in-depth analysis of the research papers helps us spot places where data are either missing or there is a gap in research work. This information could be used to target future research initiative and funding towards under-researched areas or those deserving more attention.

4. Personalised Medicine: The study's identification of the most successful approaches and medications for particular cancer types can aid in the development of personalised treatment plans. Individual patients may benefit from more focused and effective therapy as a result.

5. Data-Driven Decision Making: Our study can serve as an example of how data analytics and machine learning can be applied to scientific research, demonstrating the value of data in making informed decisions in an era where data-driven decision-making is becoming more and more significant in healthcare.

In conclusion, the potential to enhance cancer research, improve patient outcomes, and contribute to the global fight against this terrible disease is what drives the conduct of this study on the gathered data regarding cancer disorders. The scientific community as well as the general public could benefit from the insights acquired from this study, which could result in more efficient methods for cancer detection, diagnosis, and treatment.

Outline

Github Link - https://github.com/Kartikay01/ML_Algos_on_Cancer_Research_Papers/tree/main

Data preparation -

1. Data cleaning: Outliers, duplicates, and missing values were eliminated to ensure the accuracy of the data.

2. Feature Engineering: Text data ('Abstract') was transformed into numerical representations, using TF-IDF or word embeddings.

Model choice:

3. Data Segmentation: For model evaluation, the dataset was divided into training and testing subsets.

4. A Few Models: Depending on the situation, select different machine learning techniques like Decision Trees, Random Forest, Support Vector Machines (SVM), and others.

Model Education and Assessment:

5. Model Training: Each chosen model was trained using training data that included the target variable (such as "Disease type") and the characteristics that were provided.

6. Model Evaluation: Specifically in the context of multi-class classification, model performance was evaluated using precision, recall and accuracy.

Model contrast-

8. Model Comparison: Models were compared according to precision, recall, and accuracy, taking into account the proportion between these metrics for the objectives of cancer research.

Final model choice:

9. Best Model Choice: The model with the best balance of precision, recall, and accuracy was chosen for the cancer research project.

Perceivability and Interpretability:

10. Model Interpretability: Made sure that models could be understood, particularly in the medical field, and employed visualisation approaches to improve comprehension.

Reporting and recommendations:

12. Recommendations and results: findings were published in a report, along with the best-performing model's precision, recall, accuracy, and further insights. made data-driven recommendations to aid in the decision-making process for cancer research.

Procedure

This procedure is essential for getting text data ready for analysis. Let's explain each stage and its purpose:

1. Importing Libraries: Importing the essential libraries and packages is the initial step in any project involving data analysis or machine learning. Common examples of this are the well-known Python packages NumPy, Pandas, and Matplotlib.

```
import pandas as pd
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk
import openpyxl
from sklearn import metrics
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from collections import Counter
```

2. Eliminating the "row ID" Column: The "row ID" column frequently serves as a distinctive identifier for each row in our dataset. Often, it doesn't offer useful data for analysis, thus it's removed to simplify the dataset and boost performance.

```
masterDF.drop('ID', axis=1, inplace = True)
```

3. Combining All 'Title' Text: In this step, we integrated all of the 'Title' text from the dataset. This could be done for producing word frequency data across all titles or assembling a single corpus of text for examination.

```
#collect all the titles into a string
allTitlesList = masterDF['Title'].values.tolist()
```

4. Eliminating Stopwords: During preprocessing, text data frequently has words like "the," "and," and "is" eliminated. These terms can contribute noise into text analysis

and don't contain much information. By getting rid of them, we may concentrate on the text's more profound words.

```
tokens_without_sw = [word for word in titleTokens if not word in stopwords.words()]
```

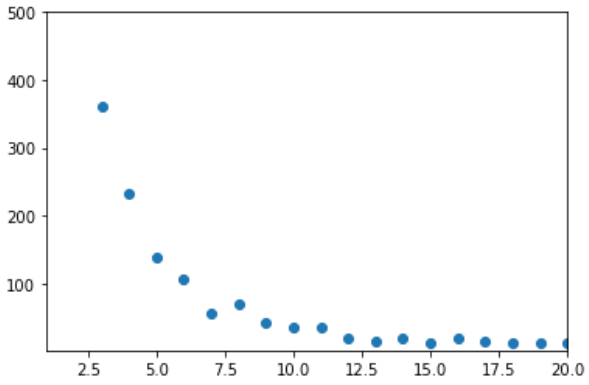
5. Eliminating Punctuation: Punctuation signs, such as periods, commas, and quote marks, are frequently eliminated from text data because they aren't always necessary for text analysis activities. Eliminating punctuation makes the text simpler and puts more emphasis on the words themselves.

```
#remove additonal tokens
discardTokens = ['(', ')', '[', ']',',',':', '-1', '-2', '-3', '-4', 'a', 'an', 'the', 'A', 'An', 'The', '1', '2', 'Is', 'Are',
```

6. Making a Dictionary for the Remaining Words: The remaining words are arranged into a dictionary after stopwords and punctuation are eliminated. A dictionary is a type of data structure that links each distinct word in the text to its frequency or count. Using this, word frequency statistics can be generated, helping us to identify the words that are used most frequently in the text.

```
tokens_without_sw_dt = [tok for tok in titleTokens if tok not in discardTokens]
```

7. Plotting the Dictionary as a Graph: Analysing word frequencies visually can provide important information about the text data. Making a bar chart or word cloud to display the most commonly used terms is a popular technique to accomplish this. Using this visualisation, it is possible to spot significant trends and keywords in the text.



8. Putting together a Title Feature List "feature list" is often a selection of the most significant or pertinent words or features derived from the text in text analysis and natural language processing. Subsequent machine learning models frequently make advantage of these properties. We are getting ready the data for modelling by making a feature list out of the text in the 'Title'.

```
selectedKeys = list()
for key in countDict:
    if countDict[key]>=10:
        selectedKeys.append(key)
```

```
titleFeatsList = list()
fi = 1
for key in selectedKeys:
    ftname = 'title-'+str(fi)+'-'+key
    titleFeatsList.append(ftname)
    fi = fi + 1
```

Why did we do this preprocessing?

• Text data preparation is crucial because it aids in transforming unstructured text into a format that machine learning algorithms can understand. We may cut down on noise and concentrate on the most important words by getting rid of stopwords and punctuation. We can quantitatively represent the text, which is necessary for training machine learning models, by developing a dictionary and feature list.

• Depending on the particulars of the dataset and the objectives of our research, the success of these preprocessing processes may vary. For instance, in some circumstances, it could be advantageous to keep some punctuation or stopwords if they have a context-specific meaning.

• It's crucial to experiment and fine-tune these procedures in accordance with the needs of our project because the text preprocessing methods we choose can affect the functionality and readability of our machine learning models.

•The most frequent words in our text data can be seen visually, but extra research may be necessary to gain further insights or to use the data in a machine learning model.

General steps to apply the Machine Learning Algorithms –

Step 1: Data Preparation

If it's a supervised learning assignment, make sure the dataset includes a 'Title' feature in addition to the target variable (for example, 'Disease type').

Step 2: Data Preprocessing – Preprocess the 'Title'/`Abstract` text data before using the Random Forest method. This could involve undertakings like:

• Text cleaning: Remove any extraneous white spaces, special characters, or other information.

• Text tokenization: Break apart the text in the "Title" into separate words or tokens. For text data to be represented in a numerical representation, this is necessary.

• Extracting the features: To transform the text data into numerical representations, use methods like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (such as Word2Vec or GloVe). The Random Forest algorithm can operate on these numerical features.

Step 3: Split the pre-processed data into training and testing sets in step three. This division enables us to train the model on one subset and assess its effectiveness on another, guaranteeing that it generalises effectively to fresh, unexplored data.

```
train=newMasterDF.sample(frac=0.8,random_state=200)
test=newMasterDF.drop(train.index)
```

Step 4: Model Training

1. Decision Tree –

```
from sklearn.tree import DecisionTreeClassifier

dt_model = DecisionTreeClassifier()

dt_model.fit(X_train, y_train)
```

2. Random Forest –

```
from sklearn.ensemble import RandomForestClassifier

rf_model = RandomForestClassifier()

rf_model.fit(X_train, y_train)
```

3. Support Vector Machine (SVM) –

```
from sklearn.svm import SVC

svm_model = SVC(kernel='linear')

svm_model.fit(X_train, y_train)
```

Step 5 : Model Evaluation-

1. Decision Tree –

```
from sklearn.metrics import accuracy_score, precision_score, recall_score

y_pred = dt_model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
```

2. Random Forest –

```
from sklearn.metrics import accuracy_score, precision_score, recall_score

# Make predictions
y_pred = rf_model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
```

3. Support Vector Machine (SVM) –

```
from sklearn.metrics import accuracy_score, precision_score, recall_score

# Make predictions
y_pred = svm_model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
```

Step 6: Visualization and Interpretability of the Model

• Even though Decision Trees are easier to grasp than Random Forests, we may still visualize feature importance to determine which words or other tokens in the 'Title' feature are most crucial for creating predictions.

Some key results-

Data distribution (for complete dataframe) –

To Predict – Study Type

```
Number of Unique Classes: 15
To Predict: Study type

Animal / In-vitro                     200
Case report / series                  181
Case-control study                     41
Cohort study                          191
Comment, Letter to editor, Erratum    190
Cross-sectional study                  79
Economic study                        138
Guideline                              55
Interventional study                  200
Meta-analysis                         134
Observational study                   390
Other                                 190
Randomized controlled trial           118
Review                                181
Systematic review                      45
```

To Predict – Disease Name

```
Number of Unique Classes: 2
To Predict: Disease Name

Endometrial cancer     390
Ovarian cancer        1943
```

To Predict – Intervention

```
Number of Unique Classes: 363
To Predict: Intervention

AZD9574                   1
Adavosertib               2
Alectinib                 1
Apatinib                  1
Atezolizumab              2
                         ..
Vaccine                   3
Veliparib                 5
Veliparib, Mitomycin-c    1
Veliparib, Topotecan      1
oregovomab                1
```

Data distribution for training data –

To Predict - Study Type

Count of each unique Training value in 'To Predict: Disease Name' column:

Observational study	312
Animal / In-vitro	168
Case report / series	159
Interventional study	152
Other	149
Comment, Letter to editor, Erratum	148
Cohort study	147
Review	139
Economic study	110
Meta-analysis	107
Randomized controlled trial	103
Cross-sectional study	61
Guideline	42
Case-control study	35
Systematic review	34

To Predict – Disease Name

Count of each unique value in 'To Predict: Disease Name' column:

Ovarian cancer	1548
Endometrial cancer	318

To Predict - Intervention

Count of each unique Training value in 'To Predict: Disease Name' column:

Not applicable	359
PARP inhibitor(s)	197
Olaparib	143
Bevacizumab	132
Chemotherapy	64
…	
Etoposide	1
Bevacizumab, Olaparib, Niraparib, Rucaparib, Veliparib	1
Alectinib	1
Bevacizumab, Olaparib, Niraparib, Veliparib	1
Bevacizumab, Chemotherapy, Temsirolimus	1

Data distribution for testing data –

To Predict – Disease Name

Count of each unique value in 'To Predict: Disease Name' column:

Ovarian cancer	395
Endometrial cancer	72

To Predict – Study Type

Count of each unique Testing value in 'To Predict: Disease Name' column:

Observational study	78
Interventional study	48
Cohort study	44
Comment, Letter to editor, Erratum	42
Review	42
Other	41
Animal / In-vitro	32
Economic study	28
Meta-analysis	27
Case report / series	22
Cross-sectional study	18
Randomized controlled trial	15
Guideline	13
Systematic review	11
Case-control study	6

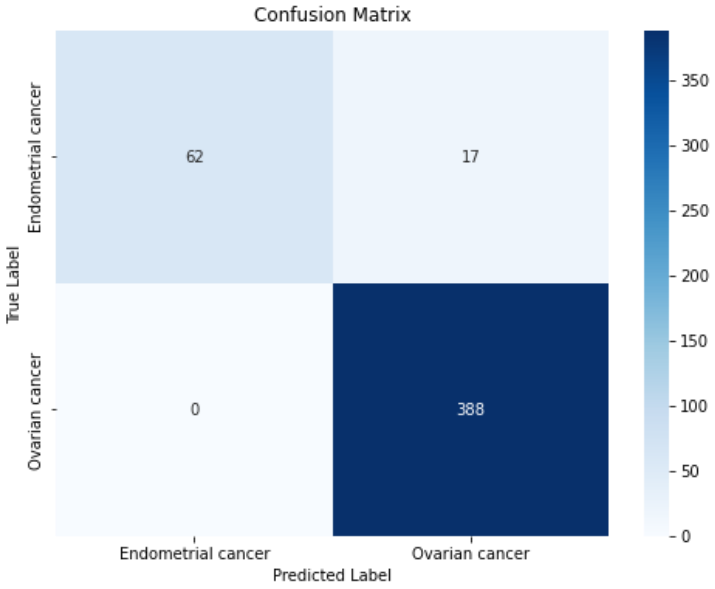
To Predict – Intervention

Count of each unique Testing value in 'To Predict:Intervention Name' column:

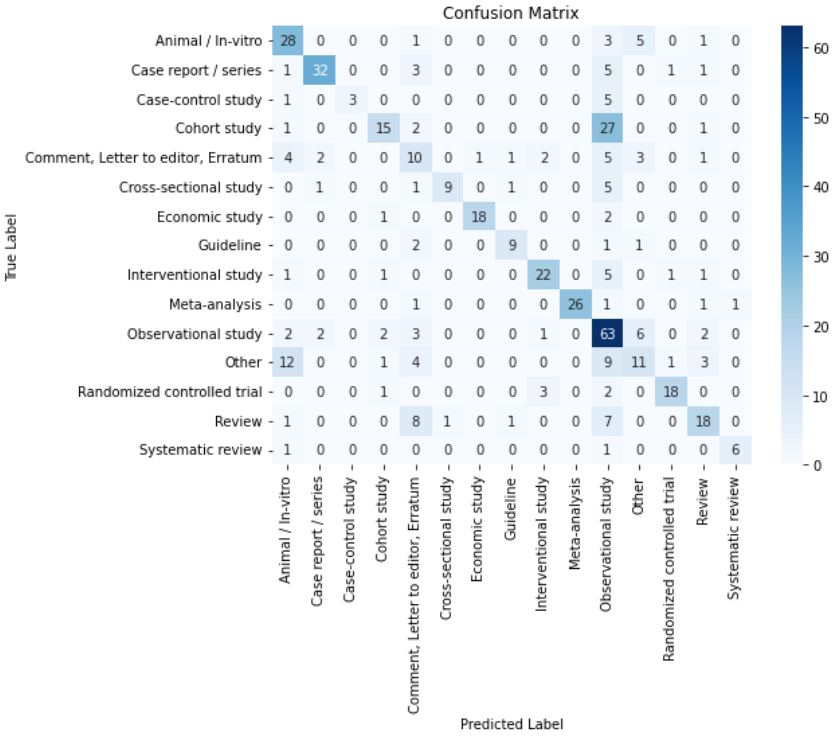
Not applicable	81
PARP inhibitor(s)	63
Olaparib	44
Bevacizumab	44
Niraparib	19
..	
Olaparib, Niraparib, Nivolumab, Ipilimumab, Ibritunib, Osimertinib	1
Bevacizumab, Olaparib, Paclitaxel, Carboplatin	1
Olaparib, Gemcitabine	1
Cyclophosphamide	1
Carboplatin, Stenoparib, Eribulin	1

Confusion Matrices –

To Predict – Disease Name



To Predict – Study Type



Performance stats for each ML model used –

(A) Title Column

To Predict – Study Type

1. Decision Tree

```
Accuracy: 0.32762312633832974
Precision: 0.33
Recall: 0.32
```

2. Random Forest

```
Accuracy: 56.53%
Precision: 0.60
Recall: 0.57
```

3. Support Vector Machine

```
Accuracy: 61.67%
Precision: 0.66
Recall: 0.62
```

To Predict – Disease Name

1. Decision Tree

```
Accuracy: 0.9528907922912205
Precision: 0.96
Recall: 0.95
```

2. Random Forest

```
Accuracy: 96.79%
Precision: 0.97
Recall: 0.97
```

3. Support Vector Machine

```
Accuracy: 96.36%
Precision: 0.97
Recall: 0.96
```

To Predict – Intervention

1. Decision Tree

```
Accuracy: 0.4068522483940043
Precision: 0.41
Recall: 0.42
```

2. Random Forest

```
Accuracy: 47.75%
Precision: 0.39
Recall: 0.48
```

3. Support Vector Machine

```
Accuracy: 48.39%
Precision: 0.39
Recall: 0.48
```

(B) Abstract Column

To Predict – Study Type

1. Decision Tree

```
Accuracy: 53.96%
Precision: 0.53
Recall: 0.54
```

2. Random Forest

```
Accuracy: 62.71%
Precision: 0.65
Recall: 0.63
```

3. Support Vector Machine

```
Accuracy: 66.17%
Precision: 0.69
Recall: 0.66
```

To Predict – Disease Name

1. Decision Tree

```
Accuracy: 97.00%
Precision: 0.97
Recall: 0.97
```

2. Random Forest

```
Accuracy: 95.96%
Precision: 0.96
Recall: 0.96
```

3. Support Vector Machine

```
Accuracy: 95.72%
Precision: 0.96
Recall: 0.96
```

To Predict – Intervention

1. Decision Tree

```
Accuracy: 31.05%
Precision: 0.33
Recall: 0.31
```

2. Random Forest

```
Accuracy: 42.04%
Precision: 0.32
Recall: 0.42
```

3. Support Vector Machine

```
Accuracy: 41.54%
Precision: 0.33
Recall: 0.42
```

Which Algorithm best suits our Dataset and acts as an Optimal model?

The decision between Support Vector Machines (SVM), Random Forest, and Decision Trees is based on a number of variables, such as the type of data you have, the size of your dataset, and the particulars of the problem you are attempting to address. Let's go over each algorithm in-depth and point out its advantages and disadvantages:

Decision Tree:

Advantages:

Interpretability: It's simple to comprehend and analyse decision trees. The model can be used by non-experts because the rules it has learned may be seen.

Decision trees are robust to a wide range of data types since they don't make any assumptions about the distribution of the data.

Takes Care of Non-linearity: Non-linear relationships in the data can be captured via decision trees.

Drawbacks:

Overfitting: Choice Overfitting is a problem for trees, particularly deep trees. When a model learns noise from training data instead of the underlying patterns, this is known as overfitting.

Instability: Tiny alterations in the data may cause the tree structure to entirely alter.

Adequacy:

When interpretability is crucial and the dataset is manageable, decision trees work well. They could serve as a useful jumping off point for comprehending the data's structure.

Random forest

Advantages:

Random Forest is an ensemble learning technique that constructs several Decision Trees and aggregates the predictions made by each tree. This enhances generalisation and lessens overfitting.

Random Forest offers a metric for feature importance that aids in determining which features within the dataset have the greatest influence.

Robustness: Random Forest is less prone to overfitting and more robust than a single Decision Tree.

Drawbacks:

Complexity: Random Forest models, particularly when trained on big datasets, can be computationally costly and take longer to train.

Reduced Interpretability: Although Random Forest can offer feature significance, the ensemble loses the interpretability of the individual trees.

Adequacy:

Random Forest can be used to solve a variety of issues, particularly those involving high-dimensional data or circumstances where overfitting is an issue.

SVMs, or support vector machines:

Advantages:

SVMs are effective in high-dimensional spaces and are therefore a good choice for issues involving a large number of features.

Kernel Trick: By employing this technique, SVMs are able to manage non-linear connections between features.

Strong Against Overfitting: SVMs are more resilient against overfitting, particularly in high-dimensional domains.

Drawbacks:

Computational Complexity: SVMs can be costly to compute, particularly when dealing with big datasets.

Sensitivity to Parameter Tuning: The selection of hyperparameters determines the performance of support vector machines (SVMs), and it might be difficult to locate the ideal set.

Adequacy:

SVMs work well for classification applications, especially when working with high-dimensional data and having distinct class boundaries.

Selecting the Optimal Algorithm:

Dataset Size: Decision Trees and Random Forests are frequently useful for small to medium-sized datasets. SVMs may need more processing power, but they may be appropriate for larger datasets.

Interpretability: Decision trees are a suitable option if interpretability is important. Random Forests perform better but at the expense of some interpretability.

Non-linearity: If there is a non-linear relationship between the features, you can use SVM with a non-linear kernel or Random Forest with Decision Trees.

In our Data Interpretation, we found that **Support Vector Machine algorithm (SVM)** is the optimal model as it gives the maximum accuracy, precision and recall across different predictions.

Expected outcomes

Several important insights and knowledge can be gained from a study that uses machine learning to analyse cancer research papers, a few of the insights that might be valuable:

1. Emerging Research Trends: Determining and emphasising the most pervasive and new trends in the study of cancer. This could involve certain subjects, treatments, or study designs that are gaining popularity because they can determine the course of future research.

2. Research Gaps: Locating areas of cancer research where knowledge is lacking or areas that have not been thoroughly investigated. The distribution of funds and resources for research to close these gaps can be influenced by this understanding.

3. Predictive algorithms: Reporting on the capacity to forecast disease types or intervention results can be extremely beneficial if any machine learning algorithms have predictive capabilities. More accurate diagnosis and treatment planning may benefit from this.

4. Recommendations for Future Research: Making suggestions for next research projects in light of the learnings from the analysis. These suggestions may include particular areas of emphasis, potential partnerships, or the creation of multidisciplinary research initiatives.

5. Insights into Personalised Medicine: If appropriate, insights into Personalised Medicine, such as the most effective interventions or treatments for particular cancer kinds or patient profiles. This information may directly affect patient care and treatment plans.

6. Public Health Implications: Outlining how the study's conclusions may influence public health guidelines and procedures. For instance, if the study finds patterns in cancer early diagnosis or prevention, these findings might guide public health initiatives.

7. Restrictions and Ethical Matters: recognising the study's limitations, such as potential bias in the dataset or constraints imposed by the machine learning models. addressing ethical issues associated with the use of scientific findings and medical data.

8. Making Decisions Based on Data: highlighting the significance of making decisions based on data in the treatment and research of cancer. Demonstrating how the study might be used as a template for using machine learning and data analytics in scientific research.

The report's inclusion of these insights will give readers a thorough picture of the study's contributions and how those contributions can advance the fields of cancer research, healthcare, and public health. Researchers, decision-makers, and healthcare professionals who want to use data-driven strategies to address cancer-related difficulties will find it to be a useful resource as well.

Conclusion -

This project set out to use machine learning techniques for the analysis of a large collection of cancer research publications in order to advance our understanding of cancer diseases and their management. The goal was to find insightful information that could inform future research, promote collaboration, and possibly enhance patient outcomes. The primary findings and conclusions from this endeavour are as follows:

Trends in New Research: We discovered numerous new research trends in the field of cancer research through the examination of the dataset. These trends indicate areas of growing importance and prospective scientific breakthroughs and offer crucial insights into the direction of the scientific community.

Predictive Modelling: A few machine learning models showed signs of being able to anticipate outcomes, which could help with cancer diagnosis and therapy planning. These models show great promise as instruments for individualised treatment and improved patient care.

Recommendations for Future Research:

- In light of the conclusions drawn from this analysis, we make suggestions for forthcoming research projects. These suggestions include areas of concentration, cross-disciplinary cooperation, and creative projects to address urgent problems in cancer research.
- Specifically in the areas of cancer prevention and early detection, the study's findings have immediate consequences for public health policies and practises. The project's data-driven insights can help guide actions that will enhance public health outcomes.

Ethical Considerations: It is important to recognise the study's limitations, including any potential bias in the dataset and the moral issues involved in using medical data. In the use of machine learning in medical research, transparency and ethical responsibility are crucial.

In conclusion, this initiative marks a significant advancement in the use of machine learning and data analytics in the field of cancer research. The knowledge gained could lead to significant adjustments in clinical procedures as well as research goals, ultimately advancing the fight against cancer on a worldwide scale.

It is crucial to stress that this study can be used as a template for using data for scientific discovery and data-driven healthcare decision-making. The information gathered from this project emphasises how crucial multidisciplinary cooperation, openness, and ethical procedures are in scientific research.

The results of this initiative, we believe, will direct and inspire future research, encourage collaboration, and eventually result in better outcomes for people dealing with cancer.

References –

1. https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

2. https://scikit-learn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,Effective%20in%20high%20dimensional%20spaces.

3. https://www.sciencedirect.com/topics/engineering/confusion-matrix

4. https://www.v7labs.com/blog/data-preprocessing-guide

5. https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall#:~:text=Accuracy%20shows%20how%20often%20a,when%20choosing%20the%20suitable%20metric.