Lung Cancer Prediction Model and High-Risk Group Prediction Using Machine Learning

Mariya Mushtaq

Capstone Project Report: Computational Science (Bryn Mawr College)

December 2023

ABSTRACT

This paper aims to make a meaningful contribution to the evolving landscape of lung cancer research by exploring the Lung Cancer Prediction Dataset through a comprehensive and nuanced approach rooted in machine learning methodologies. The primary objective is the construction and analysis of a Lung Cancer Prediction Model, which encompasses implementations of logistic regression and K-nearest neighbors (KNN) models. The second main aim of the project shifts focus to evaluating potential avenues of predicting high-risk patient subgroups, achieved through implementation of the K-Means clustering algorithm. By revealing traits that patients have in common, this clustering method could help provide a more comprehensive knowledge of the heterogeneity of lung cancer. Through the integration of sophisticated machine learning methods with analysis of their applications, the results could have the potential to guide clinical judgment and pave the way for further studies focused on customized therapies and enhanced patient care in the context of lung cancer.

TABLE OF CONTENTS

INTRODUCTION

1.1 Background

Lung cancer, a formidable global health challenge, remains a pervasive threat, standing as the primary cause of cancer-related deaths worldwide and leading cause of morbidity and mortality. The severity of the issue is extremely apparent, particularly considering its prevalence in the US, where it is the primary cause of cancer-related deaths and is diagnosed almost every two minutes [1]. This alarming statistic underscores the urgent need for innovative approaches to comprehend and combat the complexity of this disease. With estimated 2.2 million new cases and 1.8 million related deaths in 2020 alone, the global burden of lung cancer reached unprecedented levels, accounting for 11.4% of all cancer cases and 18.0% of all cancer deaths worldwide [2].

1.2 Goal

This project's main objective is to evaluate the predictive power of logistic regression and K-nearest neighbors (KNN) in estimating a person's risk of lung cancer by using relevant clinical and demographic characteristics as predictor variables. Furthermore, by using the K-Means technique to apply unsupervised learning, the study investigates the accuracy and viability of identifying unique patient subgroups within a lung cancer dataset based on shared characteristics. The project aims to address some/all the following questions:

- Can logistic regression and/or K-nearest neighbors (KNN) accurately predict individual lung cancer risk using clinical and demographic features?
- To what extent can these models, operating on a dataset rich in clinical and demographic features, reliably and accurately gauge an individual's risk of developing lung cancer?
- Does the K-Means algorithm, applied to a lung cancer dataset, have the capacity to identify distinct patient subgroups based on shared characteristics?
- To what degree can the K-Means algorithm accurately categorize patients into different groups, and what are the implications of this accuracy for understanding lung cancer case heterogeneity?

LITERATURE REVIEW

Important Python libraries are loaded into the project to help with a variety of data processing, modeling, and assessment tasks. The NumPy library is used for array operations and mathematical functions. The Pandas library is utilized for efficient data processing. Tools for dataset splitting, label encoding, logistic regression modeling, and performance metric computations are available in the scikit-learn toolkit. The project also uses the Seaborn and Matplotlib packages for data visualizations, which improves the results' interpretability.

Additionally, a few different machine learning algorithms can be found implemented in this project. Classification algorithms are used to categorize unseen data whereas Clustering algorithms are used to group data into clusters to examine for greater similarities. Regression algorithms are used to find patterns and build models. This section provides an overview of the key concepts and techniques essential to the implementation and understanding of the models utilized in this project. All of these subsections together provide the theoretical framework for the later application of these machine learning ideas in estimating the risk of lung cancer and identifying hidden patterns in the dataset.

4.1 Supervised and Unsupervised Learning

Supervised learning: "Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values.", [3].

Unsupervised learning: "Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.", [4].

4.2 Logistic Regression

“Logistic regression estimates the probability of an event occurring, based on a given dataset of independent variables. It is used to estimate the relationship between a dependent variable and one or more independent variables, and also to make a prediction about a categorical variable versus a continuous one.”, [5]. Specifically, multinomial logistic regression has been utilized for this project, where the dependent variable has three or more possible outcomes.

4.3 K-nearest Neighbors

“The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.”, [6].

4.4 Dimensionality Reduction and Feature Selection

“Feature selection reduces the dimensionality of data by selecting only a subset of measured features (predictor variables) to create a model. Feature selection algorithms search for a subset of predictors that optimally models measured responses, subject to constraints such as required or excluded features and the size of the subset.”, [7]

It mainly improves prediction performance of a model by selecting a subset of the most significant features from a dataset.

“Dimensionality reduction is a data preparation technique performed on data prior to modeling. It refers to techniques for reducing the number of input variables in training data.”[8]. The aim with dimensionality reduction is to lower the number of dimensions (or features) in a dataset while maintaining as much information as feasible.

EXPLORATORY DATA ANALYSIS

Plot 1: Genetic Risk vs Average Risk Level

The attached bar plot illustrates the categorically encoded levels of genetic risk for patients, ranging from 1 (minimum risk) to 7 (maximum risk) concerning the development of lung cancer. Upon examination, a discernible pattern emerges within the dataset, highlighting a substantial distinction in lung cancer risk levels corresponding to the increasing presence of genetic risk. Interestingly, the average risk of lung cancer is, for those falling into the low to medium hereditary risk categories, on the lower end of the spectrum. In contrast, as the genetic risk level surpasses the medium threshold and extends to higher values, the average level of lung cancer risk proportionally elevates. This finding is consistent with the intuitive theory that people who have higher genetic susceptibilities to cancer are more likely to be diagnosed with lung cancer than people who have lower genetic contributions.

In essence, the bar plot supports the logical expectation that higher genetic risk corresponds to higher probabilities of receiving a corresponding lung cancer risk classification by effectively illustrating the correlation between genetic risk levels and the average risk of developing lung cancer

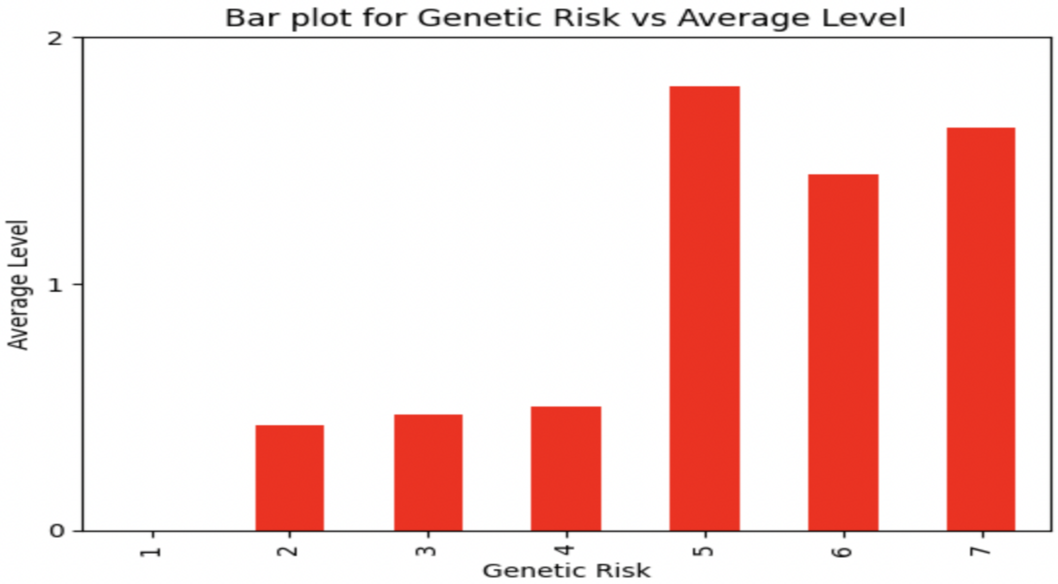

Figure: Bar plot for Genetic Risk vs Average Level

Plot 2: Occupational Hazards vs Average Risk Level

The attached bar plot shows the relationship between Occupational Hazards, numerically encoded from 1 (minimum) to 8 (maximum), and the corresponding Average Risk Level for cancer. Upon examination, a pattern becomes apparent in the dataset that indicates a significant difference in lung cancer risk levels in relation to the growing number of occupational hazard factors. There is a trend indicating an escalation in the risk of developing cancer to levels beyond the moderate range. The occupational hazards that are represented in this category have an effect on the average risk level. There is a distinct trend towards increased cancer risk when the occupational hazard level is above 50%. This finding is consistent with the pattern in which workers who are subjected to high levels of occupational risks are more likely to fall into higher cancer risk groups. Asbestos Exposure (high) and Office Ergonomics (low) are two instances of potential risks.

In conclusion, the bar plot clearly illustrates the relationship between the average risk level for cancer and the degree of occupational hazards. The pattern is apparent and highlights the significant impact that occupational hazards—especially those with increasing intensity—have on raising the risk of cancer to values above the moderate range.

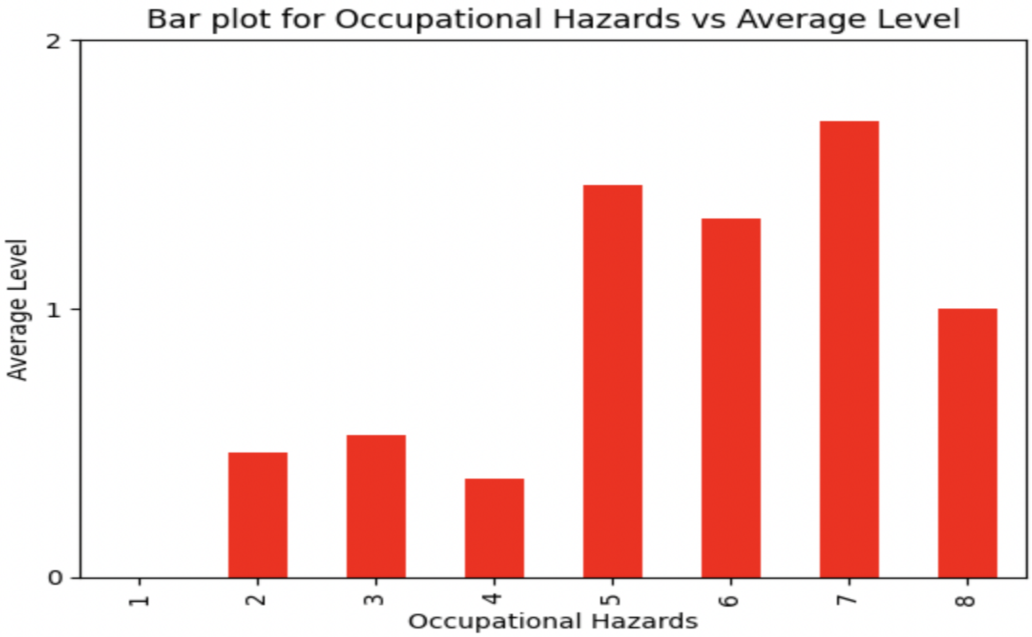

Figure: Bar plot for Occupational Hazards vs Average Level

METHODOLOGY

6.1 Data Processing

The project utilizes data sourced from the "Lung Cancer Prediction" Kaggle dataset which offers comprehensive insights into individuals diagnosed with lung cancer. The dataset contains a selection of health-related characteristics in addition to basic demographic data like age and gender. Notable variables include exposure levels to factors like air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease etcetera. These variables, a blend of numeric and categorical attributes, constitute the foundation for subsequent machine learning model applications. There were no null values or outliers in the original dataset. The original dataset was 1000 rows x 26 columns in size accounting for 25 features each over a total of 1000 patients.

In the preprocessing phase, categorical values representing the "level of cancer" were encoded to numerical equivalents: 'low' as 0, 'medium' as 1, and 'high' as 2, ensuring compatibility with machine learning models. Three unnecessary features were removed from the original dataset, including the index of the dataset, patient ID, and the 'level of cancer' attribute, as the latter served as the target feature for the predictive model. The target variable, representing the 'level of cancer,' was segregated for both training and test sets. The dataset was split into training and test sets using a 70:30 ratio, with the training data comprising 700 instances and 23 features. Simultaneously, the test data was constructed with 300 instances and the same 23 features.

6.2 Logistic Regression Model

The logistic regression model implementation relies on the LogisticRegression class from the sklearn.linear_model module. The hyperparameters are used to tailor the logistic regression model to optimize its performance. A regularization strength of C=0.1 is employed to mitigate the risk of overfitting, ensuring a balanced and generalizable model. The maximum number of iterations during optimization is set to 100,000, providing ample opportunity for the model to converge to an optimal solution. The choice of the 'liblinear' solver algorithm is driven by its effectiveness in optimization tasks. Adopting a 'one-vs-rest' (ovr) multi-class strategy complements the selected solver algorithm, enhancing the model's ability to handle multiple classes. The training data is then fed into the model through the fit method, facilitating the iterative optimization process. Once trained, the model is capable of making predictions on new data

6.3 K-nearest Neighbors Model

The K-nearest neighbors (KNN) model is implemented using the KNeighborsClassifier class from the sklearn.neighbors module. The model is configured to consider the 30 nearest neighbors (n_neighbors=30) when making predictions. It is important to note that selecting a smaller value for n, such as 5, could lead to overlap between decision boundaries, potentially resulting in incorrect predictions. The parameter 'p=2' signifies the use of the Euclidean distance metric, denoted as "minkowski," for proximity calculations (∑i=1n|Xi−Yi|p)^(1/p). The 'uniform' weight assignment ensures that each neighbor exerts equal influence in the decision-making process. Following model instantiation, the KNN model is trained on the training dataset using the fit method. Subsequently, predictions are generated for the test dataset using the predict method. The foundation for using the KNN algorithm to forecast the "level of cancer" based on the clinical and demographic characteristics in the dataset is laid by this methodical procedure.

6.4 Principal Component Analysis

Principal Component Analysis (PCA) is implemented utilizing the PCA class from the sklearn.decomposition module, while data standardization is accomplished with the StandardScaler class from the sklearn.preprocessing module. This standardization process ensures that the data is transformed to have a zero mean and unit variance using the StandardScaler. PCA is then applied with the number of components specified as 5 (n_components=5). The loadings, indicating the contributions of each feature to the principal components, are then extracted. To visually assess these contributions, a bar plot is generated to illustrate the loadings for the first five principal components. This visualization provides valuable insights into the relative importance of each feature in capturing the variance within the dataset.

RESULTS AND DISCUSSIONS

7.1 Logistic Regression Model Results

The implemented multiclass logistic regression model achieved an accuracy of 94%, indicating robust predictive performance. It was applied to both the original training set and a PCA-transformed version, revealing minimal differences in accuracy and confusion matrix analyses (93%). This could suggest that, *for lung cancer prediction models, emphasizing specific disease-related features does not significantly affect predictive performance. Considering all dataset features may be preferable to reduce potential inaccuracies in predictions*.

Prediction analysis:

- For low-risk cancer predictions (level 0), the model made 82 correct predictions and 2 incorrect predictions. No predictions were made for high-risk levels, avoiding overestimation. Incorrect predictions occurred at the boundaries between low and medium risk levels.
- For medium-risk cancer predictions (level 1), the model achieved 87 correct predictions and 10 incorrect predictions. Similar to low-risk predictions, no predictions were made for high-risk levels, preventing overestimation. Inaccuracies were noted at the interfaces between low and medium risk levels.
- For high-risk cancer predictions (level 2), the model made 114 correct predictions with 5 incorrect predictions. The absence of predictions for low-risk levels indicated avoidance of underestimation, while inaccuracies occurred at the boundaries between high and medium risk levels.

Prediction analysis (PCA transformed data):

- For low-risk cancer (level 0) comprised 89 correct predictions and 9 incorrect predictions. Consistently, no predictions were made for high-risk levels, mitigating the risk of overestimation. Inaccuracies persisted at the boundaries between low and medium risk levels.
- For medium-risk cancer predictions (level 1), the model demonstrated 76 correct predictions and 3 incorrect predictions. Similar to low-risk predictions, no predictions

were made for high-risk levels, preventing overestimation. The model exhibited inaccuracies at the interfaces between low and medium risk levels.

- For high-risk cancer predictions (level 2), the model achieved 114 correct predictions with 9 incorrect predictions. The absence of predictions for low-risk levels suggested a prudent avoidance of underestimation, while inaccuracies were observed at the boundaries between high and medium risk levels.

In all scenarios, the model consistently misclassified instances only at the interfaces between consecutive risk levels (except once), without exhibiting extreme deviations

7.2 K-nearest Neighbors Model Results

The reason another predictive model was also generated was to observe if the type of model employed would make any difference in the overall predictive accuracy. This model too was applied to both original and PCA transformed training datasets. On the original dataset, the implemented KNN model again achieved an accuracy of 93%, indicating similar predictive performance. However, for the PCA transformed training dataset, the model's accuracy increased from 93% to 97%.

Prediction analysis:

- For low-risk cancer predictions (level 0), the model made 87 correct predictions and 7 incorrect predictions. No predictions were made for high-risk levels, avoiding overestimation. Incorrect predictions occurred at the boundaries between low and medium risk levels.
- For medium-risk cancer predictions (level 1), the model achieved 79 correct predictions and 5 incorrect predictions. Similar to low-risk predictions, no predictions were made for high-risk levels, preventing overestimation. Inaccuracies were noted at the interfaces between low and medium risk levels.
- For high-risk cancer predictions (level 2), the model made 114 correct predictions with 18 incorrect predictions. The absence of predictions for low-risk levels indicated

avoidance of underestimation, while inaccuracies occurred at the boundaries between high and medium risk levels.

Prediction analysis (PCA transformed data):

- For low-risk cancer (level 0) comprised 87 correct predictions and 0 incorrect predictions. No inaccuracies were observed.
- For medium-risk cancer predictions (level 1), the model demonstrated 92 correct predictions and 5 incorrect predictions (level 0) and 1 incorrect prediction (level 2). Here, the model exhibited inaccuracies at the interfaces between low, medium risk levels and high risk levels.
- For high-risk cancer predictions (level 2), the model achieved 113 correct predictions with 2 incorrect predictions. The absence of predictions for low-risk levels suggested a prudent avoidance of underestimation, while inaccuracies were observed at the boundaries between high and medium risk levels.

7.3 K Means Clustering Algorithm Results

The secondary objective of incorporating the K Means Clustering algorithm was to assess its capability to identify discernible patient subgroups defined by shared characteristics. However, when put into practice, the observed clusters were unable to provide definitive insights that would have helped develop a more nuanced knowledge of the connections between the elements of the dataset, especially in predicting high-risk patient subgroups. There are several possible reasons for this result. One explanation is that the technique used in conjunction with the dataset's intrinsic complexity may have made it difficult to identify significant patterns. Alternatively, the characteristics defining high-risk patient subgroups might be more subtle and intricate than what the clustering algorithm could effectively capture.

When interpreting the findings of clustering analysis, it is crucial to take into account the interaction between dataset features and method selection. In this case, the inability of the K Means Clustering algorithm to identify significant patient subgroups points to the necessity of a more focused strategy or the investigation of alternate clustering methods that are more in line with the intricate nature of the dataset and the complex nature of high-risk patient characteristics.

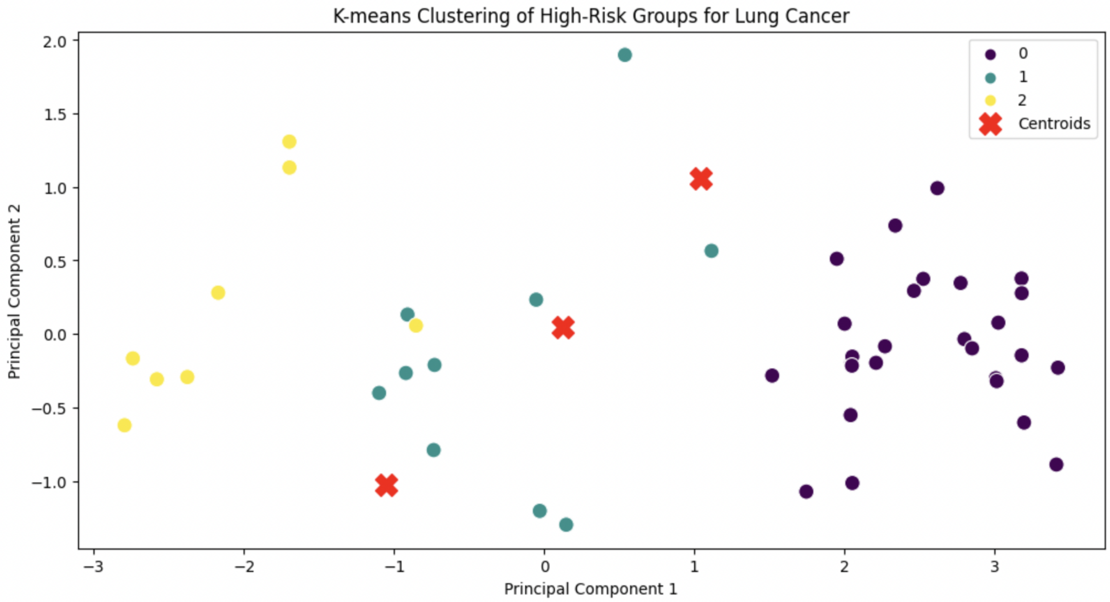

Figure: K-means Clustering of High Risk Group for Lung Cancer

7.4 Interpretations of results, Conclusion and Future Direction

The decision to generate an alternative predictive model aimed to explore potential variations in predictive performance. The comparison between the first model implemented and the KNN model indicates that the *predictive model selection may have an impact on overall predicted accuracy*.

For the original dataset, both models— the Logistic Regression model and the KNN model- demonstrated a comparable accuracy of ~94%, indicating consistent predictive capabilities. This suggests that the model selection had *little* bearing on the predicted result within the parameters of the original dataset. But when the models were applied to the training dataset that had been PCA-transformed, few notable differences became apparent. The KNN model showed an improvement in accuracy in this instance, going from 93% to 97%. This shift suggests that the KNN model, when applied to the PCA-transformed data, was more effective in capturing underlying patterns and relationships, leading to an improvement in predictive accuracy compared to the initially implemented model. The KNN model, in this instance, demonstrated superior performance on the PCA-transformed data, *highlighting the importance of considering both the model type and the nature of the dataset in predictive modeling*.

On the other hand, the use of the K Means Clustering method to identify patient subgroups based on similar characteristics exposed limits in terms of offering conclusive information, particularly with regard to the prediction of high-risk patient subgroups. Among the possible reasons are that the dataset's inherent complexity makes it difficult for the clustering algorithm to identify significant patterns.

Going forward, the results emphasize the necessity of a *more tailored* and *nuanced* approach in *clustering analysis* as well as *predictive modeling*. Alternative clustering techniques that better fit the complicated features of high-risk patient characteristics and the intricate nature of the dataset should be investigated in future areas. To boost the predictive models' robustness and yield more informative findings for healthcare applications, further research is required to comprehend the interactions between various model types and dataset variables.

BIBLIOGRAPHY

[1] “Key Findings.” *Www.lung.org*, www.lung.org/research/state-of-lung-cancer/key-findings.
[2] Li C, Lei S, Ding L, Xu Y, Wu X, Wang H, Zhang Z, Gao T, Zhang Y, Li L. Global burden and trends of lung cancer incidence and mortality. Chin Med J 2023;136:1583–1590. doi: 10.1097/CM9.0000000000002529
Global burden and trends of lung cancer incidence and mortality - PMC.
[3] (2017). ’Supervised learning’, Mathworks, [Online]. Available: https://se.mathworks.com/discovery/supervised-learning.html (accessed January 31, 2018).
Supervised Learning - MATLAB & Simulink
[4] (2017). ’Unsupervised learning’, Mathworks, [Online]. Available: https://se.mathworks.com/discovery/unsupervised-learning.html (accessed
January 31, 2018)
Unsupervised Learning - MATLAB & Simulink
[5] IBM. “What Is Logistic Regression?” Www.ibm.com, 2022, https://www.ibm.com/topics/logistic-regression
What is Logistic regression? | IBM

[6] ---. “What Is the K-Nearest Neighbors Algorithm? | IBM.” *Www.ibm.com*, 2023, www.ibm.com/topics/knn.

What is the k-nearest neighbors algorithm? | IBM

[7] “Introduction to Feature Selection - MATLAB & Simulink.” *Www.mathworks.com*, www.mathworks.com/help/stats/feature-selection.html.

Introduction to Feature Selection - MATLAB & Simulink

[8] Brownlee, Jason. “Introduction to Dimensionality Reduction for Machine Learning.” Machine Learning Mastery, 5 May 2020, machinelearningmastery.com/dimensionality-reduction-for-machine-learning/.

Introduction to Dimensionality Reduction for Machine Learning - MachineLearningMastery.com

Resources:

Dataset: Lung Cancer Prediction

Code file: Notebook.ipynb

Book: Machine Learning with PyTorch and Scikit-Learn by sebastian Raschka, Yuxi (Hayden) Liu and Vahid Mirjalili

Website: Scikit-learn

Article: What is a confusion matrix?. Everything you Should Know about… | by Anuganti Suresh | Analytics Vidhya | Medium

Article: Logistic Regression. Simplified.. After the basics of Regression, it’s… | by Apoorva Agrawal | Data Science Group, IITR | Medium

Article: Logistic Regression Simply Explained in 5 minutes | by Serafeim Loukas, PhD | MLearning.ai | Medium

Article: Demystifying the Confusion Matrix | by Mattison Hineline | Medium

Article: Evaluating machine learning models: How to tackle metrics