Predictive Modeling of Disease Outcomes Using Patient Data: A Comparative Analysis of Linear Regression and Decision Tree Approaches

1st Yogadisha Sendhil Kumar
Computer and Information Technology
Purdue University
Lafayette, United States
ysendhil@purdue.edu

***Abstract*—Predictive modeling plays a crucial role in healthcare, enabling early diagnosis and informed decision-making. In this study, the investigation focuses on the accuracy of predicting disease outcomes (Positive/Negative) using patient data. The comparative analysis centers around two popular approaches: linear regression and decision trees. By leveraging symptoms, demographic information, and health indicators, the study aims to uncover patterns that contribute to accurate predictions.**

***Keywords*—predictive modeling, disease outcomes, linear regression, decision trees**

I. Introduction

The landscape of healthcare has undergone a profound transformation propelled by remarkable strides in medical technology and the unprecedented availability of healthcare data. In this era of innovation, predictive modeling stands out as a powerful instrument, offering tantalizing prospects for foreseeing disease outcomes based on comprehensive analyses of patient data. Recognizing the pivotal role of predictive modeling in shaping healthcare strategies, it becomes imperative to unravel the intricate web of factors that underpin these outcomes, thereby facilitating the formulation of tailored treatment regimens and more astute patient management protocols.

This research embarks on a journey delving deep into the predictive potential inherent within patient data, meticulously scrutinizing a spectrum of variables encompassing symptoms - such as fever, cough, fatigue, and difficulty breathing - alongside demographic markers like age and gender, and critical health indicators including blood pressure and cholesterol levels. At its core lies a comparative examination of two venerable methodologies - linear regression and decision trees - each vying for supremacy in the realm of predictive analytics. Through a rigorous evaluation of their efficacy, interpretative nuances, and resilience, the aim is to cast a revealing light upon the feasibility of accurately forecasting disease outcomes, thereby illuminating pathways towards more effective healthcare strategies.

Central to this exploration is the outcome variable, dichotomously delineating the positive and negative manifestations of diagnoses or assessments for specific ailments. In this meticulous dissection of patient data, the overarching goal is to not only contribute substantively to the burgeoning field of personalized medicine but also to elevate the precision and efficacy of clinical decision-making processes. By optimizing the allocation of resources and fostering a more nuanced understanding of disease dynamics, the findings of this study hold the promise of catalyzing substantial improvements in patient care paradigms and the overarching optimization of healthcare systems, propelling them towards unprecedented levels of efficacy and responsiveness.

II. Methodology

A. Data Collection

The research utilized a dataset containing information on disease symptoms, demographic details, and health indicators. The dataset was obtained from a reliable source and encompassed various attributes such as fever, cough, fatigue, difficulty breathing, age, gender, blood pressure, cholesterol level, and outcome variable (positive/negative diagnosis).

B. Study Design

This study harnessed the power of supervised learning, a fundamental paradigm in machine learning, to forecast the outcomes of disease diagnoses using patient data. Supervised learning involves training algorithms on labeled data, where the inputs (features) are associated with corresponding outputs (labels), allowing the model to learn patterns and relationships between the input features and the output labels.

In this context, two distinct machine learning algorithms were deployed for predictive modeling: Multiple Linear Regression (MLR) and Decision Tree Classification.

Multiple Linear Regression is a statistical method used for modeling the relationship between a dependent variable (target) and two or more independent variables (predictors) by fitting a linear equation to observed data. It assumes a linear relationship between the input features and the target variable, making it well-suited for analyzing the effects of multiple predictors on the outcome.

On the other hand, Decision Tree Classification is a non-parametric supervised learning method used for both classification and regression tasks. Decision trees recursively partition

the feature space into distinct regions, with each partition representing a decision based on the input features. They are particularly adept at capturing non-linear relationships and interactions between variables, making them suitable for complex classification tasks like predicting disease outcomes.

C. Participants

The participants in this study were represented by the entries in the dataset, each corresponding to a unique patient profile. The dataset encapsulated a diverse range of individuals across different age groups, genders, and health conditions.

D. Materials

In this study, we employed a robust set of tools to facilitate our analysis. The primary toolset revolved around the Python programming language, renowned for its versatility and extensive ecosystem of libraries tailored for data science tasks. Specifically, we made extensive use of libraries such as pandas, NumPy, scikit-learn, seaborn, and matplotlib. These libraries collectively provided a comprehensive suite of functionalities for data manipulation, visualization, and the implementation of machine learning models. Leveraging these tools, we were equipped to handle various aspects of our research, from data preprocessing to model evaluation.

E. Exploratory Data Analysis

Before embarking on the construction of predictive models, we conducted an essential phase known as exploratory data analysis (EDA). This preliminary step was instrumental in gaining a deep understanding of the dataset's characteristics and uncovering valuable insights that informed our subsequent modeling decisions. During EDA, we employed a variety of visualization techniques, including box plots, pie charts, and count plots. These visualizations allowed us to explore the distribution of variables within the dataset and elucidate potential relationships between variables and the outcome of interest. By conducting thorough EDA, we laid a solid groundwork for our predictive modeling efforts, ensuring that our subsequent analyses were built upon a comprehensive understanding of the data.

F. Statistical Methods

The Multiple Linear Regression (MLR) analysis begins with preprocessing steps to ensure the dataset's compatibility with the model. Categorical variables like 'Fever', 'Cough', 'Fatigue', 'Difficulty Breathing', 'Gender', 'Blood Pressure', and 'Cholesterol Level' are encoded into dummy variables, facilitating numerical computation. Subsequently, the dataset is divided into features and the target variable, with 'Outcome Variable' designated as the target and all other columns except 'Disease' and 'Outcome Variable' considered as features. This separation enables the model to learn the relationship between the independent variables and the target variable. The dataset is then split into training and testing sets using the train-test-split function from sklearn.model-selection, allocating 80% of the data for training and 20% for testing.

Following the data split, label encoding is applied to the target variable using LabelEncoder from sklearn.preprocessing. This transformation converts categorical labels into numerical format, a prerequisite for model training. With the target variable encoded, the MLR model is initialized and trained using the training set. The model seeks to establish a linear relationship between the features and the encoded target variable. Once trained, the MLR model predicts outcomes for the testing set using the predict method, generating predicted values for evaluation.

Predicted outcomes are then converted into binary predictions based on a threshold of 0.5. If the predicted probability exceeds 0.5, it is classified as 1; otherwise, it is classified as 0. The accuracy of the MLR model is evaluated using the accuracy-score function from sklearn.metrics, comparing the predicted values with the actual values. Additionally, a classification report is generated to provide detailed evaluation metrics such as precision, recall, and F1-score. This comprehensive approach ensures that the MLR model is effectively trained, evaluated, and capable of making accurate predictions on unseen data, offering valuable insights into the dataset's underlying relationships.

The Decision Tree Classification process follows similar preprocessing steps as the Multiple Linear Regression (MLR) analysis. Initially, the dataset is prepared by defining the features and the target variable. In this case, the features consist of all columns except 'Disease' and 'Outcome Variable', while 'Outcome Variable' is designated as the target variable, representing disease outcomes.

Next, categorical variables within the feature set are converted into dummy variables using one-hot encoding via the get-dummies function from pandas. This transformation ensures that categorical data are appropriately represented for machine learning algorithms. The drop-first parameter is set to True to avoid multicollinearity issues.

Subsequently, the dataset is split into training and testing subsets using the train-test-split function from sklearn.model-selection. The split allocates 80% of the data for training the model and 20% for testing its performance. This separation enables the model to learn from a portion of the data and assess its generalization ability on unseen data.

Following data splitting, a Decision Tree Classifier model is initialized and trained on the training set. The Decision Tree Classifier algorithm builds a predictive model in the form of a tree structure by recursively partitioning the feature space based on the values of the features.

Once trained, the Decision Tree model is utilized to predict disease outcomes for the testing set using the predict method. Predicted outcomes are compared with the actual outcomes to evaluate the model's accuracy using the accuracy-score function from sklearn.metrics. Additionally, a classification report is generated to provide detailed evaluation metrics such as precision, recall, and F1-score, offering insights into the model's performance across different classes. This comprehensive approach ensures that the Decision Tree Classifier is

effectively trained, evaluated, and capable of making accurate predictions regarding disease outcomes based on the provided features.

III. Results

The analysis commences with a thorough investigation of the dataset, examining various attributes including disease type, symptoms (fever, cough, fatigue, difficulty breathing), demographic details (age, gender), and health indicators (blood pressure, cholesterol level). Figure 1 illustrates the distribution of these attributes within the dataset, providing a visual overview of the data landscape. Notably, there are no missing values in the dataset, ensuring its integrity and completeness for subsequent analysis (Figure 2).

	Disease	Fever	Cough	Fatigue	Difficulty Breathing	Age	Gender	Blood Pressure	Cholesterol Level	Outcome Variable
332	Osteoporosis	Yes	No	No	No	70	Male	Normal	Normal	Negative
289	Hepatitis B	No	Yes	Yes	No	60	Male	Normal	Low	Positive
299	Parkinson's Disease	Yes	Yes	No	No	60	Male	High	Normal	Positive
197	Pneumonia	Yes	Yes	Yes	Yes	45	Male	High	High	Positive
83	Kidney Cancer	No	No	Yes	No	35	Male	High	High	Positive

Fig. 1. Data Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 349 entries, 0 to 348
Data columns (total 10 columns):
Disease                 349 non-null object
Fever                   349 non-null object
Cough                   349 non-null object
Fatigue                 349 non-null object
Difficulty Breathing    349 non-null object
Age                     349 non-null int64
Gender                  349 non-null object
Blood Pressure          349 non-null object
Cholesterol Level       349 non-null object
Outcome Variable        349 non-null object
dtypes: int64(1), object(9)
memory usage: 27.3+ KB
```

Fig. 2. Null Value Count

Exploring the age distribution (Figure 3) reveals a wide range of ages, spanning from 19 to 90 years, with an average age of approximately 46 years. Interestingly, individuals in older age groups appear to have a higher likelihood of testing positive for diseases, as observed in the box plot shown in Figure 4, indicating a possible outlier in the dataset.

	Age
count	349.000000
mean	46.323782
std	13.085090
min	19.000000
25%	35.000000
50%	45.000000
75%	55.000000
max	90.000000

Fig. 3. Age Description

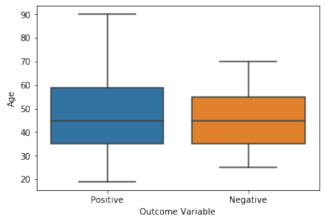

Fig. 4. Age vs Outcome Variable

Gender distribution is fairly balanced, as depicted in the pie chart in Figure 5, with nearly equal representation of male and female participants. Further examination of symptoms reveals that cough is a common occurrence across both positive and negative disease outcomes, as shown in Figure 6. As observed in Figure 7, fever is a significant indicator of a positive diagnosis. Conversely, fatigue emerges as a prevalent symptom among a substantial number of individuals, regardless of their disease status, as depicted in Figure 8.

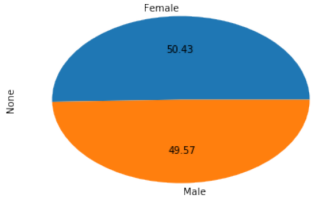

Fig. 5. Gender vs Outcome Variable

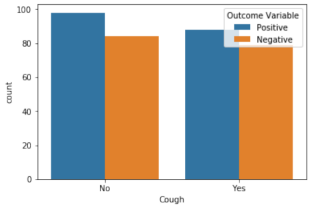

Fig. 6. Cough vs Outcome Variable

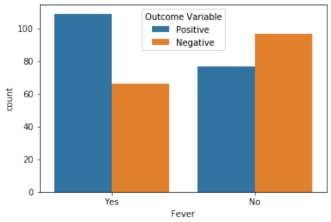

Fig. 7. Fever vs Outcome Variable

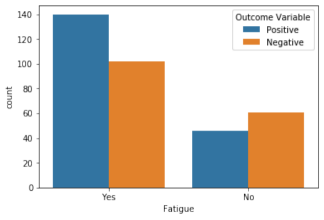

Fig. 8. Fatigue vs Output Variable

Analysis of health indicators indicates that the majority of subjects exhibit normal or high blood pressure levels (Figure 9), while high cholesterol levels are more common than low cholesterol levels (Figure 10). Additionally, a significant proportion of individuals, irrespective of disease outcome, do not experience difficulty in breathing, as illustrated in the count plot shown in Figure 11.

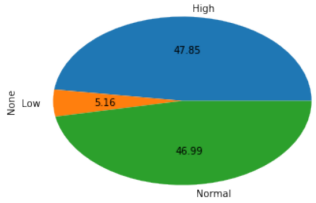

Fig. 9. Blood Pressure vs Output Variable

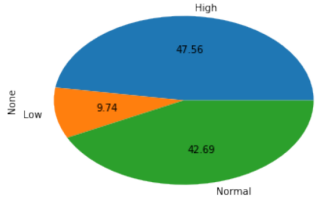

Fig. 10. Cholesterol Level vs Output Variable

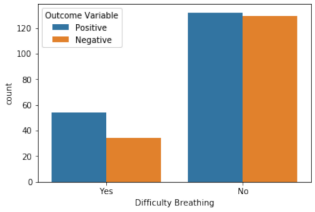

Fig. 11. Difficulty Breathing vs Output Variable

Moving on to model evaluation, both Multiple Linear Regression (MLR) and Decision Tree models are employed to predict disease outcomes based on the provided dataset. The MLR model achieves an accuracy of 57.14% (Figure 12). In contrast, the Decision Tree model outperforms with an accuracy of 72.86% (Figure 13). Precision, recall, and F1-score metrics are provided for both positive and negative disease outcomes, offering a comprehensive assessment of model performance. These findings suggest that while both models demonstrate predictive capabilities, the Decision Tree model exhibits higher accuracy in predicting disease outcomes. Further refinement and analysis may enhance the predictive power of the models and contribute to a deeper understanding of disease prognosis.

Accuracy of MLR model with label encoding: 0.5714285714285714
Classification Report:

	precision	recall	f1-score	support
0	0.50	0.53	0.52	30
1	0.63	0.60	0.62	40
avg / total	0.58	0.57	0.57	70

Fig. 12. MLR Matrix

Accuracy of Decision Tree model: 0.7285714285714285
Classification Report:

	precision	recall	f1-score	support
Negative	0.64	0.83	0.72	30
Positive	0.84	0.65	0.73	40
avg / total	0.75	0.73	0.73	70

Fig. 13. Decision Tree Matrix

IV. Discussion

In our comparative analysis, the decision tree model emerged as the more accurate predictor, achieving an accuracy rate of 72.86% compared to 57.14% for linear regression. This suggests that decision trees may be more suitable for this type of classification task, as they can capture non-linear relationships and interactions between variables more effectively. However, it's essential to consider the trade-offs between model complexity and interpretability. While linear regression offers simplicity and interpretability through its coefficient analysis, decision trees provide a more intuitive representation of the decision-making process, making it easier to understand and interpret the model's predictions. However, decision trees can become complex and prone to over-fitting, especially with large datasets. It's crucial to strike a balance between model complexity and interpretability to ensure accurate and reliable predictions.

Linear regression models offer insights into the strength and direction of relationships between independent and dependent variables through their coefficients. On the other hand, decision trees offer a more intuitive representation of the decision-making process, making it easier to understand and interpret the model's predictions. However, decision trees can become complex and prone to over-fitting, especially with large datasets. It's crucial to strike a balance between model complexity and interpretability to ensure accurate and reliable predictions.

The analysis of feature importance revealed that certain symptoms, such as fever and cough, were strong predictors of positive disease outcomes, while others, such as difficulty

breathing, had less influence on the prediction. This underscores the importance of selecting relevant features and understanding their impact on the predictive model. Additionally, exploring variable relationships uncovered interesting insights, such as the association between age and disease outcomes, with older individuals showing a higher likelihood of testing positive for diseases.

While this study provides valuable insights into disease outcome prediction, it is not without limitations. The dataset used in this study may not capture the full complexity of disease processes, and additional variables or data sources could further improve predictive accuracy. Future research could explore more advanced machine learning techniques, such as ensemble methods or deep learning, to enhance predictive modeling performance. Additionally, external validation of the predictive models on independent datasets would strengthen the generalizability of the findings.

The findings of this study have important implications for personalized medicine and healthcare decision-making. Accurate prediction of disease outcomes based on patient data can enable early intervention, optimize resource allocation, and improve patient outcomes. By leveraging predictive modeling techniques, healthcare providers can tailor treatment plans to individual patient profiles, leading to more effective and efficient healthcare delivery. Overall, this study highlights the potential of predictive modeling in healthcare and underscores the importance of further research in this area to unlock its full potential in improving patient care and clinical decision-making.

V. Conclusion

In summary, our study underscores the pivotal role that predictive modeling techniques play in healthcare, particularly in the domain of predicting disease outcomes. By harnessing patient data encompassing symptoms, demographic profiles, and health indicators, we meticulously evaluated the efficacy of both linear regression and decision tree methodologies. Our findings underscore the critical importance of selecting appropriate modeling techniques tailored to the unique characteristics of the dataset and the intricacies of the predictive task.

Through our comparative analysis, we gleaned valuable insights into the respective accuracies and interpretability of linear regression and decision tree models. While decision trees exhibited superior predictive capabilities, their inherent complexity underscores the imperative of carefully considering model interpretability. Additionally, our exploration into feature importance shed light on the relevance of specific symptoms and demographic factors in forecasting disease outcomes, offering actionable insights for personalized medicine and informed healthcare decision-making.

Looking ahead, there is a compelling need for further research to delve into advanced machine learning methodologies and validate predictive models across diverse datasets. Such endeavors are indispensable for enhancing the broader applicability and generalizability of predictive modeling in real-world healthcare settings. Ultimately, the promise held by predictive modeling in revolutionizing patient care, by enabling timely interventions, optimizing resource allocation, and enhancing clinical outcomes, is immense. By embracing these innovative approaches, healthcare providers stand poised to usher in a new era of tailored and efficacious healthcare delivery.

VI. Future Work

In considering future research directions, several avenues emerge for advancing the field of predictive modeling in healthcare. First, exploring more sophisticated machine learning techniques, including ensemble methods, deep learning architectures, or hybrid models, could enhance predictive accuracy and robustness. Additionally, external validation of predictive models on independent datasets from diverse populations or healthcare settings is essential to ensure their generalizability and real-world applicability. Incorporating additional variables, such as genetic markers or socio-economic factors, may enrich predictive models and improve their performance. Moreover, efforts to enhance the interpretability of predictive models, particularly decision trees, could facilitate their adoption in clinical practice. Finally, investigating the feasibility of real-time implementation of predictive models in clinical settings holds promise for proactive patient care and timely interventions. By pursuing these avenues, future research can contribute to the development of more accurate, interpretable, and clinically relevant predictive models in healthcare.

References

[1] Disease Symptoms and Patient Profile Dataset. (2024, January 21). Kaggle. Retrieved from https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset?resource=download

[2] Sethi, P. (n.d.). Classification Model Comparison (Diseases). Kaggle. Retrieved from https://www.kaggle.com/code/priyanshusethi/classification-model-comparison-diseases

[3] Analytics Vidhya. (2021, May). Multiple Linear Regression using Python and scikit-learn. Retrieved from https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/

[4] Springboard. (n.d.). Decision Tree Implementation in Python. Retrieved from https://www.springboard.com/blog/data-science/decision-tree-implementation-in-python/

[5] Raychev, V., Bielik, P., & Vechev, M. (2016). Probabilistic model for code with decision trees. *ACM SIGPLAN Notices*, *51*(10), 731–747. DOI: 10.1145/3022671.2984041.

[6] Vangara, V. K. M., Vuddanti, S., & Kakani, B. (2021). An Accurate and Fast Computational Python Based Module for Linear Regression Analysis in Data Science Applications. In *2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT)* (pp. 167-170). Visakhapatnam, India. DOI: 10.1109/ICISSGT52025.2021.00043.

[7] Podgorelec, V., Kokol, P., Stiglic, B., et al. (2002). Decision Trees: An Overview and Their Use in Medicine. *Journal of Medical Systems*, *26*, 445–463. DOI: 10.1023/A:1016409317640.

[8] Kotsiantis, S.B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*, 261–283. DOI: 10.1007/s10462-011-9272-4.

[9] Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(5), 1467-1474. ISSN 1871-4021. DOI: 10.1016/j.dsx.2020.07.045. Available at: https://www.sciencedirect.com/science/article/pii/S1871402120302939.

[10] Breiman, L., & Friedman, J. H. (1997). Predicting Multivariate Responses in Multiple Linear Regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *59*(1), 3–54. DOI: 10.1111/1467-9868.00054.