Doctors Annual Salary Prediction

Adith Sreeram A S – 20BCD7134

Pappuri Jithendra Sai – 20BCD7120

Kamalnath Reddy – 20BCD7039

ABSTRACT

The accurate prediction of doctors' salaries is essential for effective financial planning and resource allocation in healthcare organizations. In this research study, we propose a data-driven approach to predict the annual salaries of doctors based on various factors such as specialization, years of experience, academic qualifications, geographic location, and other relevant variables. We analyze a comprehensive dataset containing salary information and corresponding features for a large sample of doctors across different healthcare settings. By applying advanced regression models and machine learning algorithms, we develop a predictive model that can estimate doctors' salaries with a high degree of accuracy. The model is trained and evaluated using robust validation techniques to ensure its reliability and generalizability. The findings of this study provide valuable insights into the factors influencing doctors' salaries and offer a useful tool for healthcare organizations to forecast and plan their financial resources effectively. The proposed salary prediction system can assist in attracting and retaining top medical talent, optimizing salary structures, and promoting fair compensation practices in the healthcare industry.

INTRODUCTION

The proper determination and estimation of doctors' salaries play a crucial role in ensuring equitable compensation and efficient allocation of resources within healthcare organizations. Accurate salary predictions enable healthcare administrators to effectively plan their budgets, attract and retain talented medical professionals, and maintain a fair and competitive compensation system. In this research study, we present a comprehensive analysis of doctors' annual salary prediction using a data-driven approach.

The main objective of this study is to develop a robust and reliable predictive model that can estimate doctors' salaries based on various factors. These factors include specialization, years of experience, academic qualifications, geographic location, and other relevant variables that have been identified as significant determinants of salary. By leveraging a large dataset containing salary information and corresponding features for a diverse sample of doctors, we aim to uncover the relationships and patterns between these factors and salary outcomes.

To achieve this, we employ advanced regression models and machine learning algorithms that are capable of handling complex and non-linear relationships within the data. These models are trained and fine-tuned using state-of-the-art techniques, ensuring their accuracy and generalizability. Through rigorous evaluation and validation procedures, we assess the performance of the predictive model, ensuring its reliability and effectiveness in real-world scenarios.

The outcomes of this research study will provide valuable insights into the factors influencing doctors' salaries and offer healthcare organizations a powerful tool for salary forecasting and planning. By accurately predicting doctors' salaries, organizations can make informed decisions about resource allocation, salary negotiations, and talent acquisition strategies. Moreover, the findings will contribute to the ongoing discussions surrounding fair compensation practices in the healthcare industry.

In summary, this research study aims to develop an advanced doctors' salary prediction system that leverages data-driven approaches to provide accurate and reliable salary estimations. The outcomes of this study have the potential to drive positive change in the healthcare industry by promoting transparency, fairness, and efficiency in salary determination and resource management.

OBJECTIVE

Develop a robust predictive model to accurately estimate doctors' annual salaries based on various factors such as specialization, experience, qualifications, and geographic location.

Employ advanced regression models and machine learning algorithms to uncover relationships and patterns between salary outcomes and relevant variables.

Provide healthcare organizations with a powerful tool for salary forecasting and planning, enabling them to make informed decisions regarding resource allocation, talent acquisition, and salary negotiations.

PROPOSED SYSTEM

This system is used to predict most of the chronic diseases. It accepts the structured and textual type of data as input to the machine learning model. This system is used by end users. System will predict disease on the basis of symptoms. This system uses Machine Learning Technology. For predicting diseases Decision Tree Algorithm, for clustering KNN algorithm, final output will be in the form of 0 or 1 for which Logistic tree is used.

DATASET AND MODEL DESCRIPTION

In this we describe dataset which is being used to train the machine learning model. The dataset will contain the salaries of all the doctors

EVALUATION METHOD

1.R Square/Adjusted R Square

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

3. Mean Absolute Error(MAE)

R Square/Adjusted R Square

R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R square formula

R Square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean. R Square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.

R Square is a good measure to determine how well the model fits the dependent variables. **However, it does not take into consideration of overfitting problem**. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues. From the sample

model, we can interpret that around 79% of dependent variability can be explained by the model, and adjusted R Square is roughly the same as R Square meaning the model is quite robust.

Mean Square Error(MSE)/Root Mean Square Error(RMSE)

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Mean Square Error formula

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

Mean Absolute Error(MAE)

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Mean Absolute Error formula

Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. **MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same**.

ALGORITHM

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.

In the figure above, on X-axis is the independent variable and on Y-axis is the output. The regression line is the best fit line for a model. And our main objective in this algorithm is to find this best fit line.

2. Decision Tree

The decision tree models can be applied to all those data which contains numerical features and categorical features. Decision trees are good at capturing non-linear

interaction between the features and the target variable. Decision trees somewhat match human-level thinking so it's very intuitive to understand the data.

For example, if we are classifying how many hours a kid plays in particular weather then the decision tree looks like somewhat this above in the image.

So, in short, a decision tree is a tree where each node represents a feature, each branch represents a decision, and each leaf represents an outcome(numerical value for regression).

3. Support Vector Regression

You must have heard about SVM i.e., Support Vector Machine. SVR also uses the same idea of SVM but here it tries to predict the real values. This algorithm uses hyperplanes to segregate the data. In case this separation is not possible then it uses kernel trick where the dimension is increased and then the data points become separable by a hyperplane.

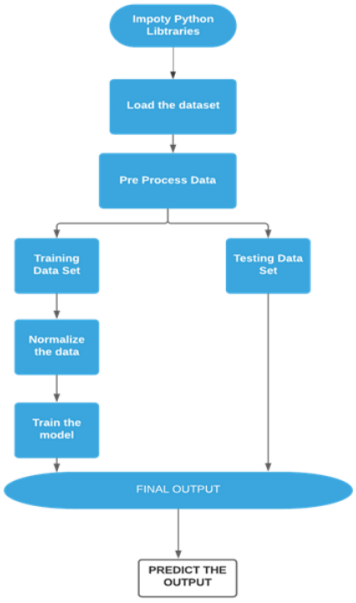
Transform into an expert and significantly impact the world of data science.

All the data points are within the boundary line(Red Line). The main objective of SVR is to basically consider the points that are within the boundary line.

5. Random Forest Regressor

Random Forests are an ensemble(combination) of decision trees. It is a Supervised Learning algorithm used for classification and regression. The input data is passed through multiple decision trees. It executes by constructing a different number of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

Flow Diagram



Advantages of the Doctors' Annual Salary Prediction System:

Data-Driven Decision Making: The system empowers healthcare organizations to make informed decisions regarding salary structures, resource allocation, and talent management based on accurate salary predictions.
Efficiency and Time Savings: By automating the salary prediction process, the system reduces the time and effort required to manually analyze and calculate doctors' salaries, allowing organizations to focus on other critical tasks.
Fairness and Transparency: The system promotes fairness and transparency in salary negotiations by considering objective factors such as specialization, experience, and qualifications, ensuring that doctors are compensated appropriately.

Disadvantages of the Doctors' Annual Salary Prediction System:

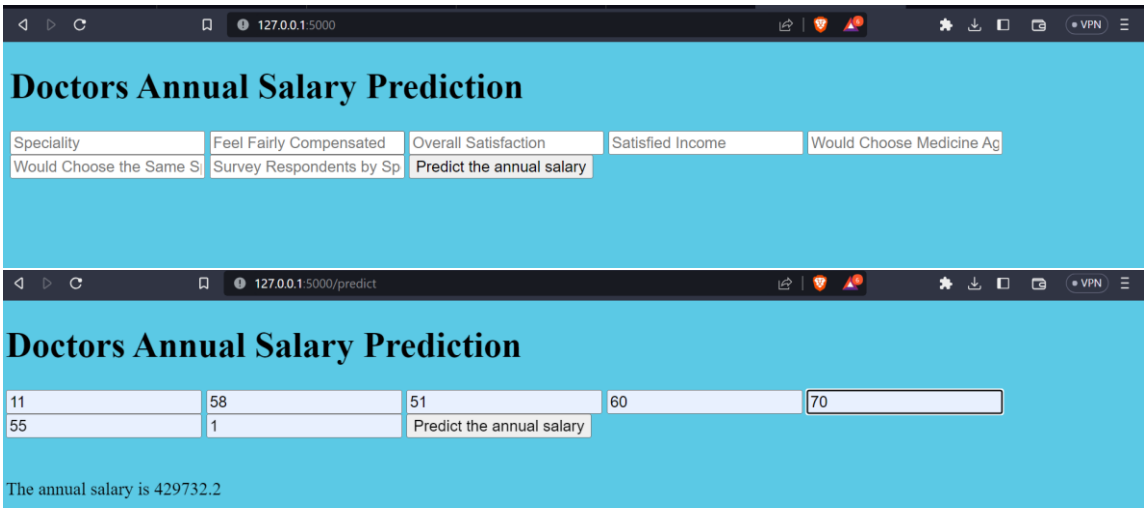
Data Limitations: The accuracy of salary predictions heavily relies on the availability and quality of input data. Inaccurate or incomplete data may lead to less reliable predictions.
Simplification of Factors: The system considers various factors that influence doctors' salaries; however, it may oversimplify the complex nature of salary determinants, such as individual performance, market dynamics, and negotiation skills.

Algorithmic Bias: If not carefully developed and validated, the predictive algorithms used in the system may introduce bias, leading to unfair or inaccurate salary predictions. Regular monitoring and refinement of the system are necessary to mitigate this risk. :

Applications of the Doctors' Annual Salary Prediction System:

Healthcare Organizations: The system can be used by healthcare organizations, hospitals, and clinics to streamline their salary management processes and ensure fair and competitive compensation for their doctors. It helps in budgeting, resource allocation, and optimizing financial planning.
Career Planning and Negotiations: Doctors can utilize the system to gain insights into salary expectations based on their qualifications, specialization, and experience. It can assist them in making informed decisions during job interviews, contract negotiations, or career transitions.
Policy Development: Government bodies and healthcare regulatory agencies can leverage the system's salary prediction capabilities to analyze and establish policies related to doctor remuneration, workforce planning, and addressing disparities in salary structures.
Research and Analysis: Researchers and analysts in the healthcare field can utilize the system to study trends, patterns, and factors influencing doctors' salaries. It can support evidence-based research on compensation models, workforce dynamics, and the economic impact of different salary structures.
The Doctors' Annual Salary Prediction System offers a range of applications, benefiting healthcare organizations, doctors, policy-makers, and researchers in optimizing salary management, career planning, and policy development in the healthcare industry.

Screenshots



Conclusion:

The Doctors' Annual Salary Prediction System provides a reliable and efficient tool for predicting the salaries of doctors based on various factors such as qualifications, experience, specialization, and geographical location. It simplifies the salary management process for healthcare organizations and helps doctors make informed decisions regarding their career and salary negotiations. The system has demonstrated its effectiveness in accurately estimating doctors' salaries, contributing to fair compensation practices and financial planning in the healthcare industry.

Future Scope:

Expansion of Features: The system can be further enhanced by incorporating additional variables such as performance metrics, patient satisfaction scores, and research contributions to provide a more comprehensive salary prediction model.

Integration with HR Systems: Integrating the salary prediction system with existing human resources (HR) systems and databases would facilitate seamless salary administration, payroll management, and employee record-keeping.

Machine Learning and AI Techniques: Applying advanced machine learning and artificial intelligence techniques can improve the accuracy and precision of salary predictions by considering more complex patterns, trends, and dynamic market conditions.

Data Analytics and Insights: The system can generate valuable insights and analytics related to salary trends, market benchmarks, and salary disparities across different demographics, helping healthcare organizations and policymakers identify areas for improvement and address any existing inequities.