*Medical Premium Insurance Cost Prediction*

PrasannaKumar Reddy, Chirra *Department of Computer Science (Graduate Student)*Richmond, Kentucky  
[prasannakumar\_chi@mymail.eku.edu](mailto:prasannakumar_chi@mymail.eku.edu)

Akhil Naidu Muthyala *Department of Computer Science (Graduate Student)*Richmond, Kentucky  
[akhil\_muthyala@mymail.eku.edu](mailto:akhil_muthyala@mymail.eku.edu)

Anudeep Reddy, Kakumani *Department of Technology Management(Graduate Student)*Richmond, Kentucky  
[nagasaianudee\_kak@mymail.eku.edu](mailto:nagasaianudee_kak@mymail.eku.edu)

Manoj Vallabhaneni. *Department of Computer Science (Graduate Student)*Richmond, Kentucky  
[manoj\_vallabhanen1@mymail.eku.edu](mailto:manoj_vallabhanen1@mymail.eku.edu)

*Abstract* — This report presents a novel approach to estimating future medical costs by developing a predictive model leveraging demographic and lifestyle data. With an emphasis on key variables such as age, gender, BMI, number of children, smoker status, and regional categorization, the study aims to equip individuals with tools for informed healthcare financial planning. Through a structured methodology encompassing data preprocessing, exploratory data analysis, and machine learning model development, the project demonstrates the efficacy of predictive modeling in navigating complex healthcare datasets. The resulting model not only offers accurate cost estimates but also serves as a practical application of machine learning in healthcare financial planning, showcasing its potential for improving decision-making processes in the healthcare domain.

Keywords — Healthcare, costs prediction, evidence regression, supervised learning.

# Introduction

In the ever-evolving landscape of healthcare, the ability to forecast and plan for future medical costs has become increasingly crucial. Our project endeavors to navigate this complex terrain by harnessing the power of predictive modeling and machine learning, with a focus on demographic and lifestyle factors. We recognize the significant financial burden that healthcare expenses impose on individuals and families, prompting a demand for robust predictive tools that can provide accurate and personalized estimates.

As healthcare systems grow in complexity and individuals assume greater responsibility for managing their medical finances, the need for reliable estimates becomes paramount. Our project seeks to address this need by developing a predictive model capable of accurately estimating future medical costs. We aim to unravel the intricate relationships between healthcare expenses and demographic features, including age, gender, BMI, number of children, smoker status, and regional categorization.

With a structured approach that encompasses data preprocessing, exploratory data analysis, and the development and evaluation of regression models, our project aims to shed light on the dynamics of future medical costs. By leveraging machine learning algorithms, we aspire not only to provide accurate predictions but also to uncover the underlying drivers influencing these predictions.

Ultimately, our project's findings have the potential to offer valuable tools for proactive healthcare financial planning, bridging the gap between uncertainty and informed decision-making in the realm of medical expenses.

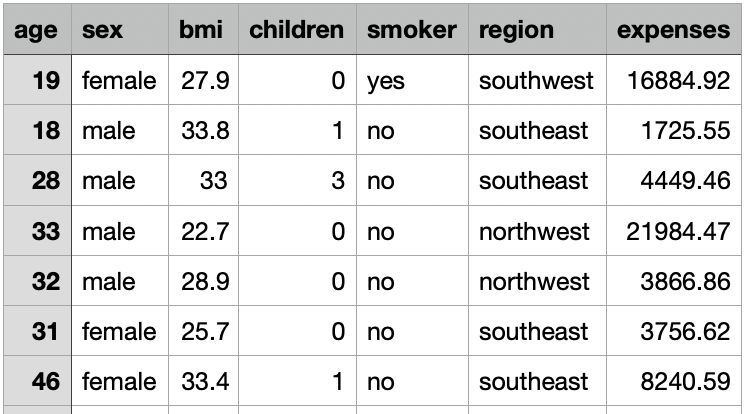
# Dataset

In this approach, we utilize a dataset sourced from Kaggle[1], characterized by its compact size of 53KB and encompassing 1330 data entries. each containing demographic and lifestyle information essential for predicting future medical costs. The dataset includes the following attributes:

* **Age:** The age of the individual.
* **Sex:** The gender of the individual (male or female).
* **BMI (Body Mass Index):** A measure of body fat based on height and weight. Calculated as weight in kilograms divided by the square of height in meters.
* **Number of Children:** The number of dependent children the individual has.
* **Smoker Status:** Indicates whether the individual is a smoker or a non-smoker.
* **Regional Categorization:** Categorizes the individual's location or region, which may impact medical costs.
* **Medical Expenses:** The target variable representing the medical expenses incurred by the individual.

This dataset serves as the foundation for our analysis and model development.

Utilizing this dataset, we embark on a structured approach, drawing inspiration from various machine learning algorithms commonly used in predictive modeling tasks.

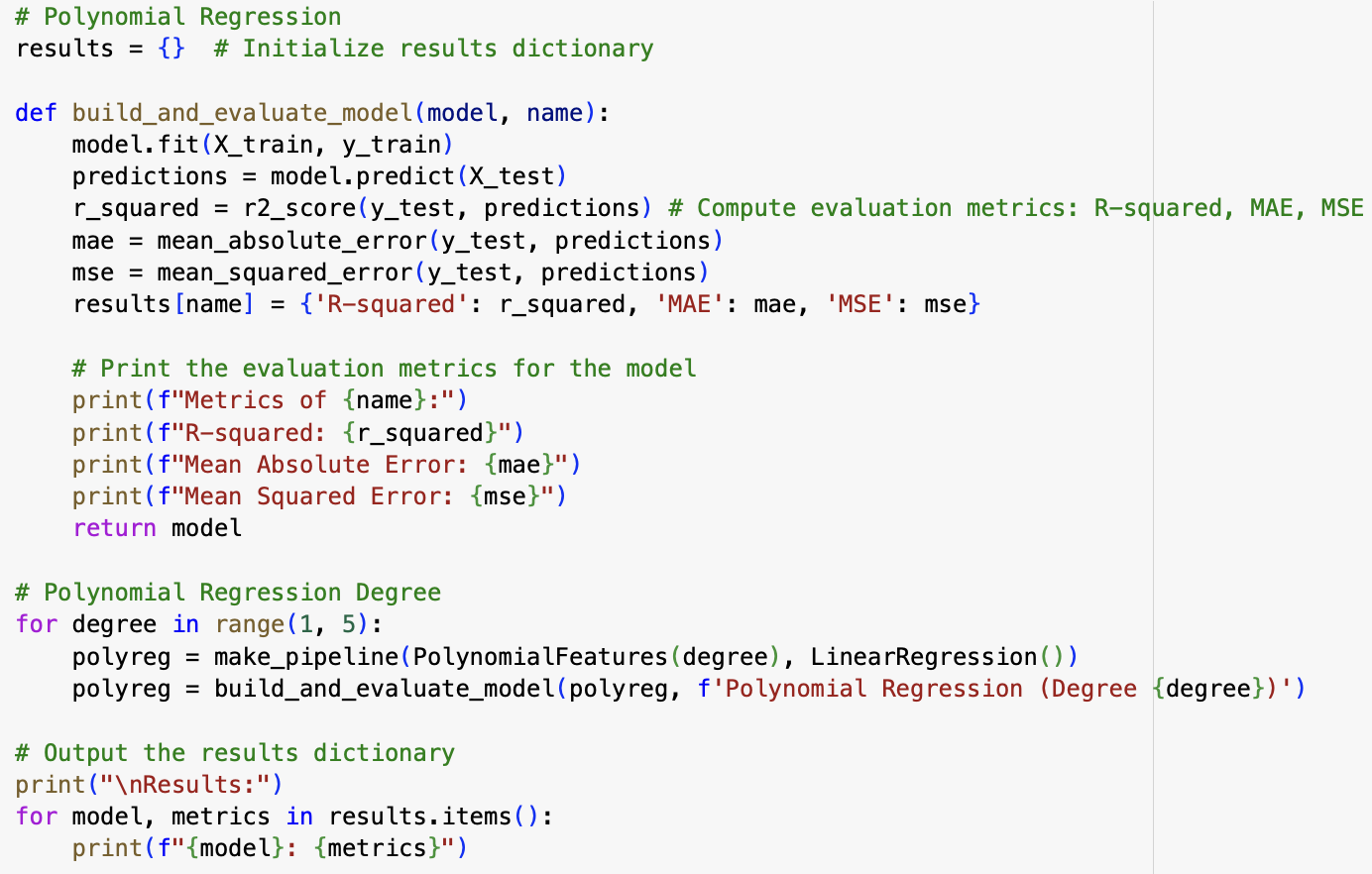


# Main Approach

Our project focuses on implementing a variety of machine learning algorithms to accurately predict future medical costs based on demographic and lifestyle factors. Each algorithm contributes unique analytical capabilities to enhance the precision and reliability of our predictions, ensuring comprehensive insights into the dataset.

A*. Polynomial Regression:*

Polynomial regression is a powerful technique used for modeling relationships between independent and dependent variables. In our project, we apply polynomial regression to capture non-linear relationships between predictors like age, BMI, and medical expenses. By transforming the features into polynomial terms of varying degrees, we create a flexible model capable of capturing complex patterns in the data. After training the polynomial regression model, we evaluate its performance using metrics such as R-squared, mean absolute error (MAE), and mean squared error (MSE), providing a comprehensive assessment of its predictive capabilities.



Polynomial Regression

*B. Decision Tree*

Decision tree regression is a non-parametric supervised learning method used for both classification and regression tasks. In our project, we employ decision tree regression to partition the dataset into subsets based on the most informative features, such as age, BMI, and smoker status. By recursively splitting the data, decision tree regression creates a hierarchy of decisions, allowing us to predict future medical costs accurately. We limit the maximum depth of the tree to prevent overfitting and evaluate the model's performance using metrics like R-squared, MAE, and MSE.

A screenshot of a computer code

Description automatically generated

Decision Tree

*C. Random Forest*

Random forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. In our project, we utilize random forest regression to predict future medical costs based on demographic and lifestyle factors. By building a forest of decision trees trained on random subsets of the data, random forest regression provides robust predictions while capturing complex interactions between variables. We specify the number of trees in the forest and evaluate the model's performance using metrics such as R-squared, MAE, and MSE.

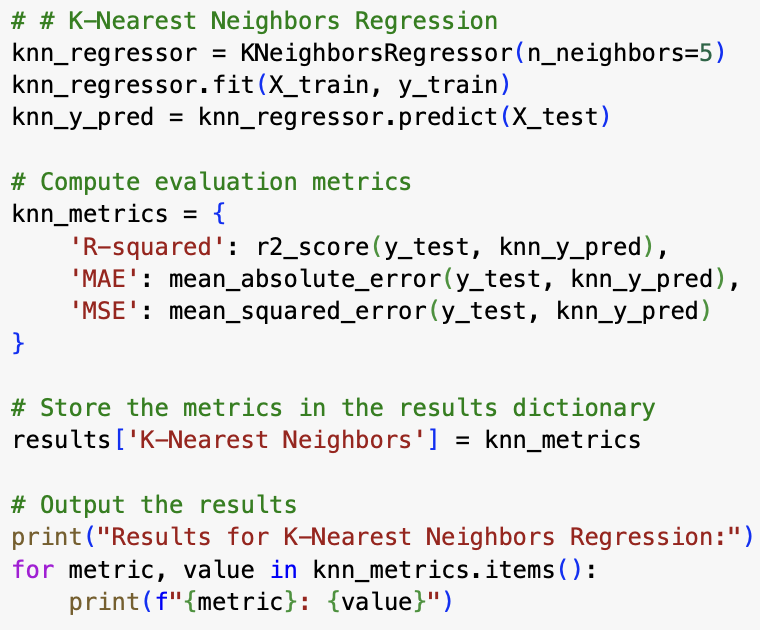
A screenshot of a computer code

Description automatically generated

Random Forest

*D. KNN*

K-Nearest Neighbors (KNN) is a non-parametric algorithm used for both classification and regression tasks. In our project, we explore the use of KNN regression to predict future medical costs based on similar individuals' demographic and lifestyle factors. KNN operates by assigning a data point to the majority class among its k nearest neighbors in the feature space, allowing us to capture local patterns and relationships in the data. We evaluate the performance of the KNN model using metrics like R-squared, MAE, and MSE, providing insights into its predictive accuracy and robustness.



KNN

# Methods

* Implemented the project using Python, utilizing common libraries such as NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn for data manipulation, visualization, and machine learning tasks.
* The Incorporated essential packages for machine learning algorithms, including Linear Regression, Polynomial Regression, Decision Tree Regression, K-Nearest Neighbors Regression, and Random Forest Regression.
* Loaded the healthcare dataset into Pandas, inspecting the first few rows to determine structure and content. Checked for missing values and evaluated overall data quality.
* Used Scikit-learn's preprocessing modules, such as StandardScaler for numerical feature scaling and One-Hot Encoder for categorical feature encoding.   
  Applied appropriate imputation strategies for missing values, resulting in a clean and complete dataset suitable for model training.
* This Based on domain knowledge and initial exploratory data analysis, we identified relevant features for predictive modeling.   
  Conducted feature selection to improve model efficiency and interpretability.
* For the initial modeling, we used Scikit-learn's Linear Regression, which considers the linear relationship between demographic and lifestyle factors and medical expenses. Polynomial regression was used to capture any potential nonlinearities in the data. The performance of each model was assessed using established metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).
* Created a comprehensive evaluation framework using R-squared, MAE, and MSE to compare the effectiveness of various regression models.   
  Model parameters are iteratively fine-tuned for optimal performance.
* Utilized Matplotlib for visualizing the most prevalent events in various states and discerning patterns across
* the United States.
* Maintained a detailed and organized documentation of the code, methodologies, and results.
* Compiled comprehensive reports showcasing the findings, insights, and visualizations derived from the predictive modeling process.

##### 

A chart of a bar graph

Description automatically generated with medium confidence Fig- Age Distribution

Fig – BMI Distribution

A green and orange rectangular bars

Description automatically generated

Fig – Region Distribution

##### Results

Upon conducting a comparative analysis of the applied algorithms, we observed that Polynomial Regression Degree 2 emerged as the top performer, achieving an impressive R-squared value of 0.866. This indicates that our model explains approximately 86.6% of the variance in the medical cost data. Decision Tree Regression with a maximum depth of 4 closely followed, achieving an R-squared value of 0.864. K-Nearest Neighbors Regression and Random Forest Regression also demonstrated strong performance, achieving R-squared values of 0.768 and 0.862, respectively.

Furthermore, Polynomial Regression Degree 2 exhibited the lowest Mean Absolute Error (MAE) of 2753.25 and the lowest Mean Squared Error (MSE) of 20,678,473.09 among all models. This suggests that our model's predictions are on average approximately $2753.25 away from the actual medical costs, with a mean squared deviation of approximately $20,678,473.09.

Our results highlight the effectiveness of Polynomial Regression in accurately predicting future medical costs based on demographic and lifestyle factors. Its superior performance in terms of R-squared, MAE, and MSE underscores its potential to provide reliable estimates of medical expenses, thereby facilitating informed decision-making in healthcare planning and resource allocation.

# Future Work

In envisioning the future trajectory of our medical insurance premium prediction system, several avenues emerge for further refinement and enhancement. Transitioning the system to a web-based platform holds the potential to significantly broaden its accessibility, offering convenience and ease of access to a wider audience. Moreover, the development of mobile applications, particularly those available in various native languages, can further extend the system's reach and usability, catering to diverse user demographics. Collaborating with government entities and healthcare organizations presents an opportunity to enrich the system's dataset with high-quality, real-time information, thereby enhancing the accuracy of predictions. Integrating advanced analytics techniques and machine learning algorithms can unlock deeper insights into healthcare trends and patterns, empowering proactive interventions, and preventive measures. As the user base expands, scalability and performance optimization become imperative, necessitating the implementation of robust cloud-based solutions and efficient data processing methodologies. Continuous updates and refinements based on user feedback and evolving healthcare landscapes are essential to ensure the system remains relevant and effective in addressing the dynamic needs of the healthcare industry. Through ongoing research and development efforts, we aim to propel the medical insurance premium prediction system toward its full potential, contributing to improved healthcare planning, decision-making, and outcomes for individuals and organizations alike.

##### Conclusion

In summary, our project aimed to develop and evaluate machine learning models for predicting medical premium insurance costs based on demographic and lifestyle factors. Through the application of various algorithms such as Polynomial Regression, Decision Tree, Random Forest, and K-Nearest Neighbors

Among the models assessed, Polynomial Regression emerged as the most promising, achieving an R-squared value of 0.867, indicating a strong correlation between the predictor variables and the target variable. Decision Tree and Random Forest also demonstrated competitive performance with R-squared values of 0.864 and 0.862 respectively, showcasing their efficacy in capturing complex relationships within the data. However, K-Nearest Neighbors lagged with an R-squared value of 0.768, suggesting limited predictive power for our dataset and problem domain.

Furthermore, Polynomial Regression exhibited superior performance in terms of Mean Absolute Error (MAE) and Mean Squared Error (MSE), indicating its ability to provide more accurate predictions compared to other models. Its interpretability and simplicity make it a valuable tool for estimating medical premium insurance costs based on demographic and lifestyle factors.

Looking ahead, our project lays the groundwork for future enhancements in medical premium insurance cost prediction. Potential avenues for improvement include expanding the dataset, exploring advanced feature engineering techniques, and incorporating additional variables such as medical history and pre-existing conditions. By refining our models and leveraging emerging technologies, we can enhance the accuracy and accessibility of medical premium insurance cost predictions, ultimately empowering individuals, and insurance providers to make informed decisions regarding healthcare coverage.

##### References

1. <https://www.kaggle.com/datasets/>
2. S. Patidar, S. Dudi and Rohit, "Estimating Medical Insurance Cost using Linear Regression with HyperParameterization, Decision Tree and Random Forest Models," *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2023, pp. 504-508, doi: 10.1109/Confluence56041.2023.10048836*.
3. G. S. Eswari, I. Gayathri Devi, C. K. Rupesh Kumar, U. V. Ramesh and A. Lakshmanarao, "An Efficient Medical Cost Prediction System with Machine Learning and Deep Learning Model, *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI), Tiruchengode, India, 2023, pp. 1-5, doi: 10.1109/ICAEECI58247.2023.10370936.*
4. K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar and R. K. Bhatia, "Health Insurance Cost Prediction using Machine Learning," *2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-5, doi: 10.1109/INCET54531.2022.9824201.*