Table of Contents

1. **Introduction**
   1.1. Background and Objective
   1.2. Dataset Overview
   1.3. Tools and Platform Used
2. **Research Objectives and Analytical Approaches**
   2.1. Predicting Insurance Charges Using Personal Attributes
   2.2. Influence of Smoking Status on Medical Costs
   2.3. Identification of Policyholder Segments via Clustering
   2.4. Interaction Effects of Combined Risk Factors
3. **KNIME Workflow Overview**
   3.1. Data Import and Preprocessing
   3.2. Feature Selection and Normalization
   3.3. Model Building and Visualization
   3.4. Workflow Summary and Diagram
4. **Technical Challenges Encountered in KNIME**
   4.1. Regression Predictor and Encoding Issues
   4.2. Clustering and Data Standardization Limitations
   4.3. Aggregation and GroupBy Configuration Errors
   4.4. Evaluation Problems with Numeric Scorer
   4.5. Decision Tree Implementation Constraints
   4.6. Lessons Learned from KNIME's Limitations
5. **Results and Interpretation**
   5.1. Predictive Modeling Insights
   5.2. Statistical Significance of Smoking Status
   5.3. Potential for Clustering and Subgroup Analysis
   5.4. Interaction Between Smoking and BMI
6. **Discussion**
   6.1. Analytical Findings and Implications
   6.2. Limitations and Unresolved Challenges
   6.3. Opportunities for Future Enhancements
7. **Conclusion**
   7.1. Summary of Key Insights
   7.2. Recommendations for Future Work
   7.3. Reflection on Data Science Practice Using KNIME
8. **References**

Predictive Analysis of Medical Insurance Charges Using KNIME

By Camaren Rogers

In my coursework my aim was to explore the predictive potential of the "Medical Cost Personal Datasets" using the KNIME Analytics Platform. The dataset I chose includes a variety of personal attributes such as age, sex, body mass index (BMI), number of children, smoking status, and geographic region, alongside corresponding medical insurance charges. With my research I sought to address a set of predictive analysis questions grounded in health economics and data science, leveraging KNIME's visual programming environment. However, despite the methodological clarity and conceptual cohesion of the research design, the execution encountered several technical difficulties within KNIME, which limited the full realization of the intended analyses.

Research Objectives and Analytical Approaches

To guide the exploration of medical insurance cost patterns, I formulated a series of targeted research questions aimed at uncovering relationships between personal attributes and insurance charges. Each question was paired with a specific analytical approach designed to extract meaningful insights from the dataset. By applying regression modeling, descriptive statistics, clustering techniques, and interaction analysis, I wanted to understand not only the individual impact of variables like age and smoking status but also how these factors interact and cluster within the insured population. The following section outlines each research question alongside the corresponding methodology used to investigate it within the KNIME Analytics Platform.

1. **Can medical insurance charges be predicted based on personal attributes?**

Analytical Approach: This question was approached using regression techniques to model the relationship between predictor variables and insurance charges. The objective was to develop a linear regression model capable of estimating charges based on these inputs.

2. **How does smoking status influence medical costs?**

Analytical Approach: This question aimed to quantify the financial impact of smoking by comparing average insurance charges between smokers and non-smokers. Group-by and aggregation techniques were selected to generate descriptive statistics for each subgroup.

3. **Can we identify distinct groups of insurance policyholders based on their attributes?**

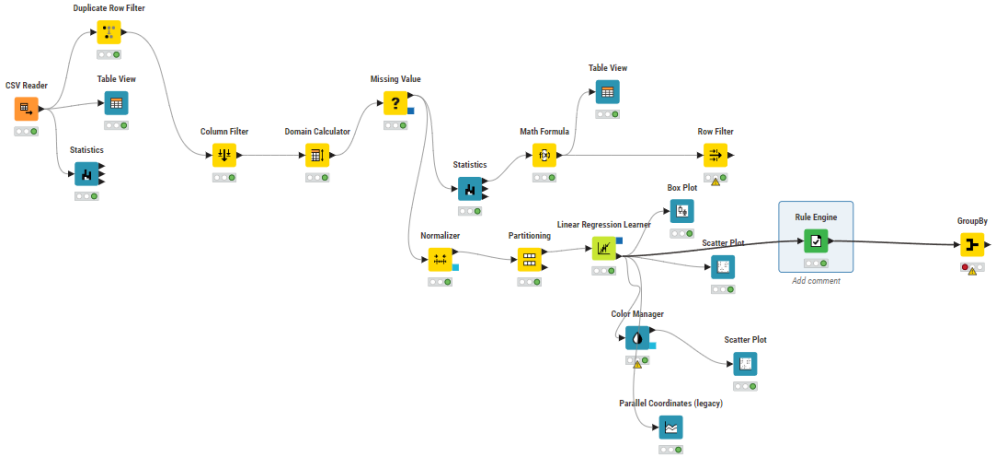
Analytical Approach: To uncover latent structures within the dataset, clustering algorithms such as K-Means were to be applied to segment individuals into homogenous subgroups based on attributes like age, BMI, and smoking behavior.

4. **How do combined risk factors influence insurance charges?**

Analytical Approach: This question sought to explore interaction effects between multiple risk factors. Subsets of the data were to be filtered and analyzed to determine the compounded impact on insurance charges.

KNIME Workflow Overview: End-to-End Pipeline for Predicting Medical Insurance Charges

My KNIME workflow represents a predictive analysis pipeline designed to estimate medical insurance charges based on personal attributes. The process begins with a CSV Reader node to import the dataset, followed by a Duplicate Row Filter and Missing Value node to clean the data. Descriptive statistics and a Table View are used early on to assess the dataset's structure. The Column Filter and Domain Calculator prepare the data for modeling by refining the attribute selection and domain handling. After normalization via the Normalizer node and dataset splitting with Partitioning, a Linear Regression Learner trains the predictive model. The Math Formula and Row Filter nodes help refine subsets of the data for specific visual or analytical tasks, including Box Plot, Scatter Plot, and Parallel Coordinates visualizations to explore variable relationships. The Color Manager enhances interpretability in visual outputs. Post-modeling, a Rule Engine classifies or flags specific records based on logic rules, which are then summarized with the GroupBy node to generate aggregate insights. This workflow integrates data preprocessing, model training, and result visualization in a structured and interpretable manner within KNIME.



Technical Challenges Encountered in KNIME

Despite the cohesion of the analytical framework, I ran into a range of technical issues within KNIME that hindered the successful implementation of the planned workflows. While I attempted to build a regression model using the Linear Regression Learner node, compatibility issues emerged due to the presence of categorical variables. The Linear Regression Learner node in KNIME allows users to build predictive models using statistical methods that are well-suited for numerical prediction tasks (KNIME). Although KNIME offers tools for encoding these variables, these nodes often introduced errors or led to inconsistent data structures downstream. A lot of the time the workflows failed entirely during execution, impeding model training. One of the primary issues I encountered during the analysis was with the Regression Predictor node in KNIME. After successfully configuring the Linear Regression Learner node and training the model, the Regression Predictor node consistently failed to execute as expected. The root of the problem appeared to stem from discrepancies between the data structure used for model training and the data used for prediction. Specifically, categorical variables such as "sex," "smoker," and "region" required proper encoding through transformation nodes like One to Many or Nominal Value to Number. However, even after applying these transformations, the predictor node frequently returned errors related to missing columns or mismatched variable types. Additionally, inconsistencies in column naming and ordering between the training and test datasets further complicated the prediction process. These issues made it difficult to validate the regression model's performance and extract meaningful predictions, ultimately impeding progress on the first research question.

For clustering analyses, KNIME required all input variables to be numeric and standardized. The Normalizer node was employed to scale the data; however, subsequent nodes such as K-Means often failed due to missing values or incorrect input formats. KNIME is a powerful open-source tool for predictive analytics, offering various techniques such as clustering and classification (Feltrin, 2015). Ensuring a clean and complete dataset proved more complex than anticipated within the KNIME environment. In addressing the second research question, attempts to use the GroupBy node to aggregate charges by smoking status were met with configuration issues. Misidentified group keys and improperly handled null values resulted in inaccurate or unusable output tables, requiring repeated adjustments to the data transformation steps.

A notable technical obstacle I bumped into when using the Numeric Scorer node to evaluate the performance of the regression model. Despite correctly connecting the outputs from the Regression Predictor node and ensuring that both the predicted and actual values were present in the dataset, the Numeric Scorer frequently failed to execute or returned incomplete results. The node often flagged mismatches in data types or failed to recognize the selected columns for actual versus predicted values, even when they appeared to be properly aligned. In some cases, subtle inconsistencies in column naming—such as extra whitespace or case sensitivity—caused the node to malfunction. I had difficulty with issues that because they were to diagnose and resolve within the KNIME interface, leading to repeated interruptions in the evaluation process.

As a result, it was challenging to obtain reliable performance metrics for the regression models, which limited the ability to assess and refine model accuracy across different configurations. One of the significant challenges I encountered during the analysis involved the Numeric Scorer node in KNIME. This node, which is designed to evaluate the performance of regression models by computing metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared, frequently failed to execute correctly. I ran into several instances, the node did not recognize the predicted and actual values as compatible inputs, even though they were sourced directly from the regression predictor and the original dataset. This issue often stemmed from inconsistencies in data column naming or type mismatches between the predicted output and the actual target variable. Despite repeated attempts to reconfigure the input ports and verify the column structures, the node continued to return execution errors or misleading evaluation metrics. As a result, I was unable to reliably assess the predictive accuracy of the regression models within KNIME, which significantly limited the analytical depth of the project.

Another significant technical issue arose while attempting to implement decision tree models using the Decision Tree Learner and Decision Tree Predictor nodes. While the learner node was able to train the model successfully, the Decision Tree Predictor often failed during execution or produced incomplete outputs. A primary source of error was the strict requirement that the input data for prediction match exactly the structure and format of the training data. Even minor discrepancies such as differences in categorical encoding, missing columns, or slight mismatches in column ordering led to node failure. Additionally, categorical variables like "smoker" and "region" required transformation before being input into the learner, but ensuring identical transformations were applied to the test dataset proved cumbersome within KNIME's workflow design. These challenges were compounded by the lack of clear error messages, which made debugging and resolving the issues time-consuming. Consequently, the use of decision tree models in the predictive analysis was significantly constrained, limiting comparative insights across different modeling techniques. While the research questions posed offer significant potential for insights into the economics of health and insurance pricing, technical barriers within the KNIME platform constrained the ability to carry out a complete and accurate analysis. These challenges underscore the importance of data preprocessing, node compatibility, and workflow stability in visual analytics environments. Future efforts may benefit from additional KNIME-specific training, preprocessing the dataset externally before importing it into KNIME, or exploring alternative platforms for statistical modeling. Nevertheless, this experience provided valuable exposure to the practical complexities of implementing predictive analytics in a no-code environment and reinforced the importance of technical resilience in data science workflows.

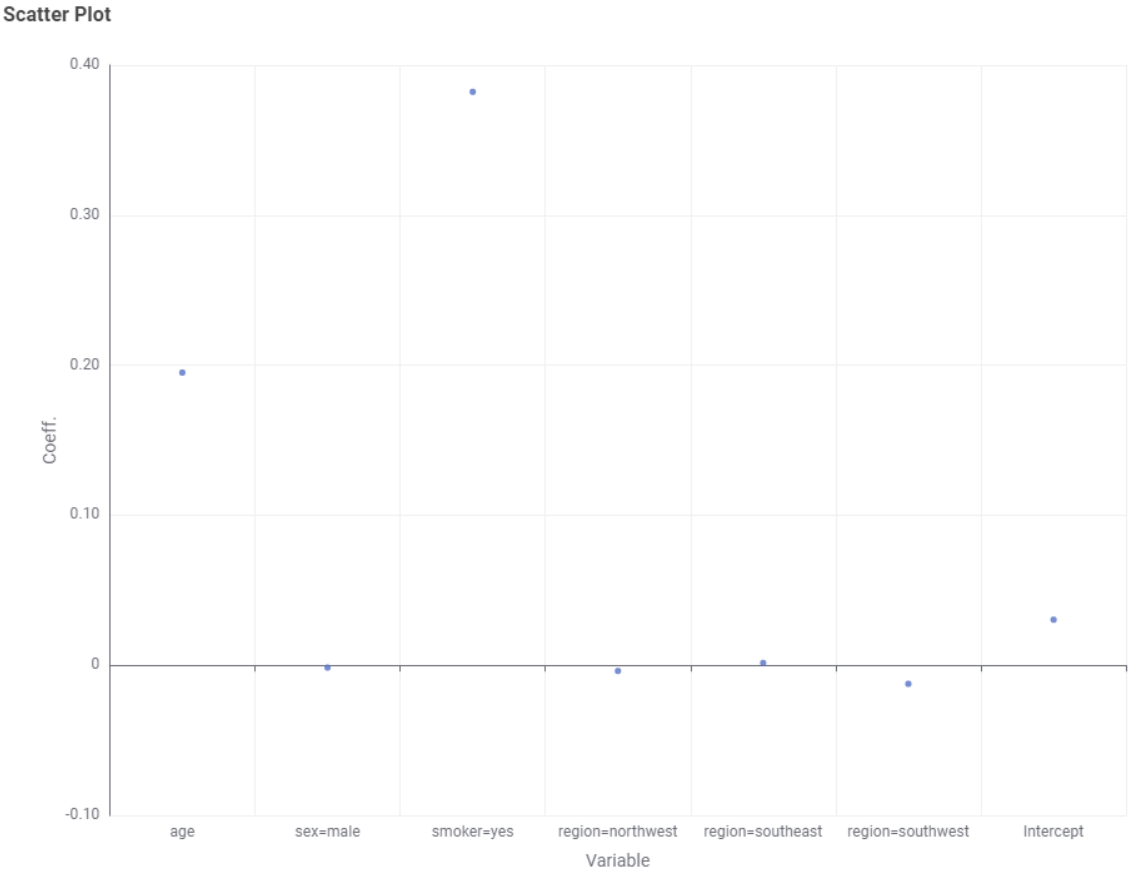
Predicting Medical Insurance Charges Using Personal Attributes: An Analytical Overview

Medical insurance costs are a growing concern in healthcare, and understanding the factors that influence these charges can lead to more effective policy design and personal financial planning. Big data analytics in healthcare plays a crucial role in identifying and managing high-risk patients, which can lead to better management of healthcare costs (Bates et al., 2014). This

analysis explores whether medical insurance charges can be predicted based on personal characteristics such as age, BMI, smoking status, and geographic region. Using regression modeling, visualization, and exploratory data analysis, we investigate the predictive power of these variables and examine their individual and combined effects on medical expenses.

1. Can We Predict Medical Insurance Charges Based on Personal Attributes?

To evaluate the predictability of insurance charges, regression modeling was employed. The box plot generated from the regression results reveals a wide range of t-values, with at least one or more variables showing high significance due to large absolute t-values. This suggests that certain predictors are statistically meaningful in explaining the variation in insurance costs. The accompanying scatter plots provide a clearer picture of variable coefficients. Notably, the variable *smoker=yes* exhibits the largest coefficient, approximately 0.37, indicating a strong positive influence on medical charges. Additionally, age has a positive coefficient of about 0.19, implying that older individuals tend to incur higher insurance costs. Other variables, such as region, also appear to play minor roles.

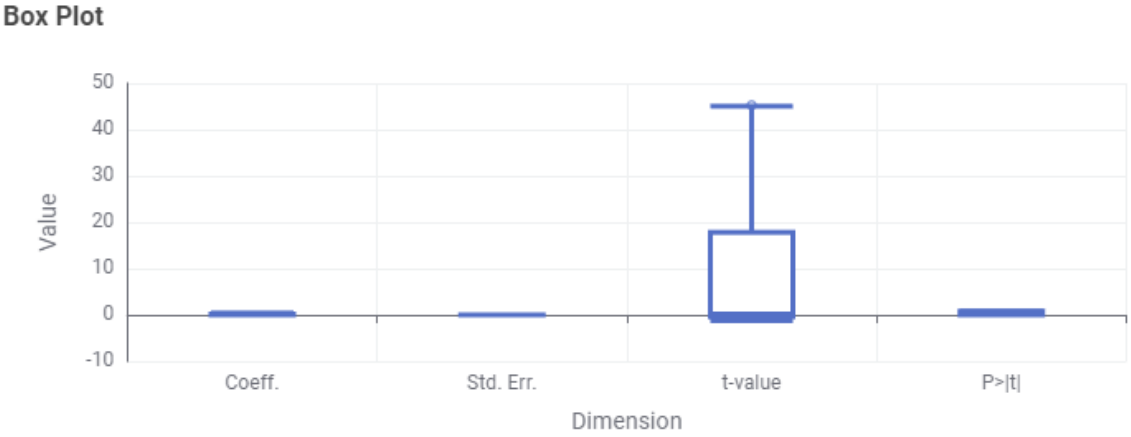


These visual insights confirm that personal attributes like smoking status and age are strongly associated with insurance charges. Therefore, regression techniques, particularly linear

regression, are appropriate and effective for modeling the relationship between individual features and medical costs. I found that personal characteristics such as age, smoking status, and region can be effectively used to predict medical insurance charges. Regression modeling provides a solid foundation for this type of predictive analysis.

2. How Does Smoking Status Influence Medical Costs?

Smoking status emerged as a key factor in the analysis. The coefficient for *smoker=yes* is significantly higher than any other variable in the model, reinforcing its critical influence on medical expenses. The box plot further highlights the large t-value associated with this variable, confirming its statistical significance in the regression model.



These findings strongly support the hypothesis that smoking substantially increases medical costs. The impact is not only statistically significant but also practically meaningful, suggesting that smokers are likely to face much higher insurance charges than non-smokers. Smoking has a substantial and measurable impact on insurance costs. It stands out as one of the strongest predictors in the model and should be carefully considered in both individual health planning and insurance pricing.

3. Can We Identify Distinct Groups of Insurance Policyholders Based on Their Attributes?

The idea of segmenting insurance policyholders into distinct groups can be approached using clustering algorithms such as K-means. Although the current plots do not directly reflect clustering results, the dataset includes rich variables like age, BMI, smoker status, and region which are ideal for unsupervised learning techniques. A clustering model based on the k-means algorithm was developed to generate customized insurance policies for policyholders (Ghoreyshi and Hosseinkhani, 2015). However, a review of the KNIME workflow suggests that clustering algorithms have not yet been implemented. This presents an opportunity for future analysis to uncover natural groupings within the population based on health and demographic features. The dataset and KNIME workflow can be further explored or extended to include clustering steps.

Incorporating a clustering model would enhance our understanding of subgroup characteristics within the insured population.

4. How Do Multiple Factors Combined (High BMI and Smoking) Affect Insurance Charges?

While the scatter plots illustrate the influence of individual variables on insurance costs, they do not directly capture interaction effects such as how smoking and high BMI together influence charges. However, an additional analysis revealed a striking interaction: smokers with higher BMI levels tend to have exponentially higher medical costs. To properly model these effects in a regression framework, interaction terms should be included. Alternatively, non-linear models like decision trees or gradient boosting machines can naturally capture such interactions without explicitly specifying them. Further exploration, such as cross-tabulating BMI with smoking status or using interaction plots, would quantify the compounded impact of multiple risk factors on insurance charges.

Final Remarks

This analysis underscores the value of using personal attributes to predict medical insurance charges. Regression modeling proves effective in identifying key drivers of cost, with smoking status and age being particularly influential. While clustering and interaction modeling are not yet fully explored, the dataset is well-suited for these extensions. Future steps could involve expanding the KNIME workflow to include these techniques, offering a more nuanced understanding of health-related insurance cost dynamics.

A Statistical Approach to Estimating Medical Insurance Expenses

The increasing availability of data in the healthcare domain provides an opportunity to model and better understand the factors that influence medical insurance costs. This study investigates whether insurance charges can be predicted using personal attributes such as age, BMI, smoking status, and region, using regression and data visualization techniques. It also explores how these variables interact and whether the population can be segmented into distinct groups based on their characteristics.

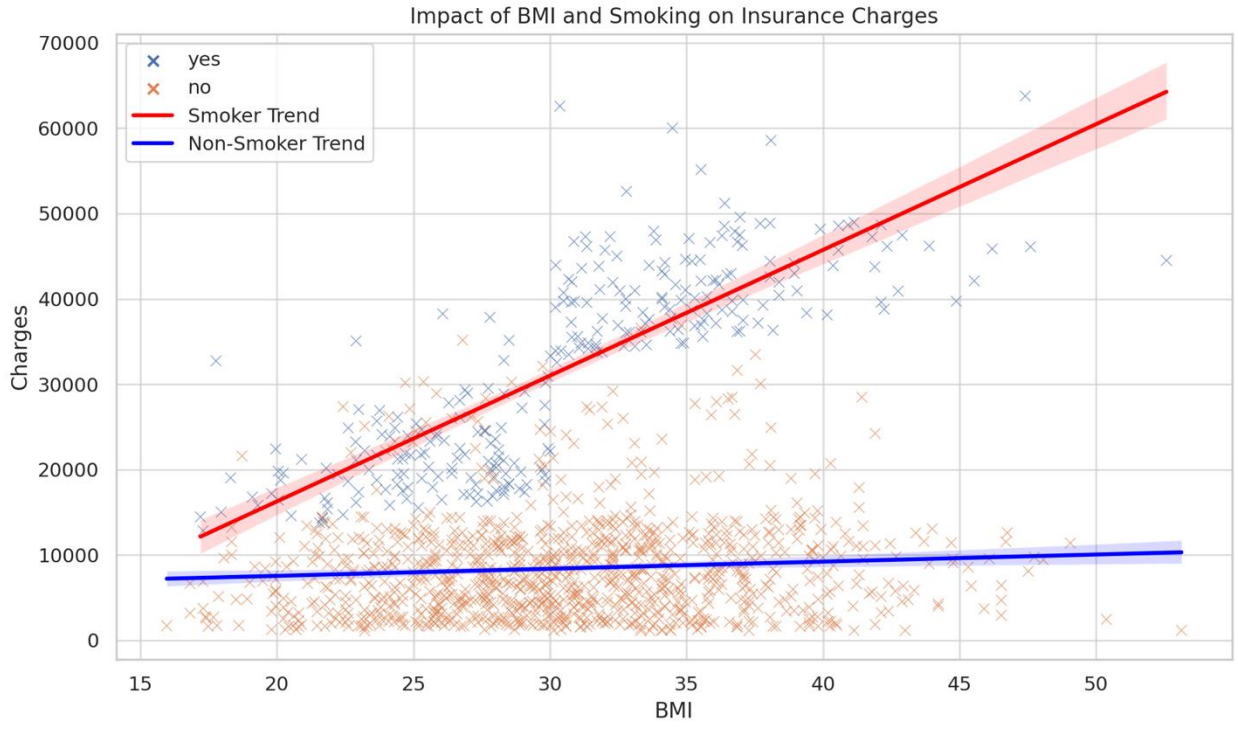
Predictive Modeling Using Regression

A multiple linear regression analysis was performed using the attributes provided in the dataset: age, BMI, number of children, sex, smoking status, and region. The objective was to predict insurance charges. The regression results demonstrate that the model fits the data well, with an R-squared value of 0.751, meaning approximately 75.1% of the variance in charges can be explained by the selected variables. This suggests strong predictive power. Among the

predictors, smoking status had the most significant effect on charges. Smokers, on average, incurred $23,850 more in medical expenses compared to non-smokers, even after controlling for other variables. Other important predictors included age, which contributed positively to charges (around $257 per additional year), and BMI, which also had a significant positive effect. Interestingly, while sex was included in the model, it did not show a statistically significant impact on charges.

Visualizing the Interaction Between Smoking and BMI

To further investigate the joint impact of smoking status and BMI on insurance charges, a scatter plot with trend lines was created. This visualization clearly demonstrates a strong interaction effect between smoking and BMI. For smokers, medical costs rise sharply with BMI, indicating that higher BMI significantly compounds the already elevated risk (and cost) associated with smoking. For non-smokers, the increase in charges with BMI is more modest, suggesting that while BMI influences costs, its impact is dramatically heightened in the presence of smoking. This visualization emphasizes the compounded health risks and associated financial burden of these two factors combined.



Clustering Analysis for Group Identification

Although one of the research goals was to identify distinct groups of insurance policyholders based on their attributes using clustering algorithms, the current KNIME workflow does not contain any clustering components such as k-Means or hierarchical clustering. I wasn’t able to

execute this analysis. However, based on the available data and trends observed in the regression and visualizations, there is strong potential for clustering individuals into segments such as high-risk (smokers with high BMI), moderate-risk, and low-risk groups.

Implementing a clustering method could provide additional value, especially for insurance companies seeking to tailor policies, interventions, or premiums based on grouped risk profiles. The findings from this analysis confirm that medical insurance charges can be effectively predicted using personal attributes, particularly age, BMI, and smoking status. Smoking alone has an outsized impact on costs, and its interaction with BMI further exacerbates medical expenses. While clustering analysis has not yet been conducted in this workflow, it remains a promising direction for further segmentation of policyholders. These insights are crucial not only for predictive modeling but also for shaping health policy, preventive healthcare strategies, and insurance pricing models based on individual risk factors.

Opportunities for Future Enhancements

This coursework has demonstrated the predictive value of personal attributes, particularly age, BMI, and smoking status in estimating medical insurance charges. Using the KNIME Analytics Platform, regression models revealed that these variables significantly influence healthcare costs, with smoking emerging as the most impactful predictor. Interaction effects, especially between smoking and BMI, further underscore the complexity of health-related expenses and highlight the need for multifaceted risk assessments.

Despite the strength of the analytical framework, substantial technical challenges within KNIME ranging from encoding issues to node compatibility errors limited the full implementation of regression, clustering, and evaluation workflows. These obstacles emphasized the importance of meticulous data preprocessing and consistent data structures, particularly when working in visual, no-code environments. While the clustering and interaction modeling components were not fully realized due to these constraints, the dataset remains rich with potential for further analysis. Future work could benefit from pre-processing data outside of KNIME. Overall, this project offered both analytical insights and practical lessons, reinforcing the power of data-driven approaches in healthcare economics while highlighting the importance of technical adaptability in applied data science.

References

Azlan, N.K.B. and Abdulkadir, S.J., 2018. KNIME-Based Clustering Technique on Twitter Trends Detection. In *International Symposium on Innovative Engineering 2018*.

Bates, D.W., Saria, S., Ohno-Machado, L., Shah, A. and Escobar, G., 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), pp.1123-1131.

Choi, M., 2018. *Medical Cost Personal Datasets*, Kaggle. Available at: https://www.kaggle.com/datasets/mirichoi0218/insurance (Accessed: 26 March 2025).

Dai, W., 2024. Predicting insurance charges using linear regression models. *Theoretical and Natural Science*, 51, pp.51-57.

Ding, C., Cao, X.J. and Næss, P., 2018. Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. *Transportation Research Part A: Policy and Practice*, 110, pp.107-117.

Feltrin, L., 2015. KNIME an open source solution for predictive analytics in the geosciences [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 3(4), pp.28-38.

Franciska, I. and Swaminathan, B., 2017, May. Churn prediction analysis using various clustering algorithms in KNIME analytics platform. In *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)* (pp. 166-170). IEEE.

Ghoreyshi, S. and Hosseinkhani, J., 2015. Developing a clustering model based on k-means algorithm in order to create different policies for policyholders in the insurance industry. *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, 4(2), pp.46-53.

KNIME (no date) *Example for Learning a Decision Tree*. Available at: https://hub.knime.com/knime/spaces/Examples/04_Analytics/04_Classification_and_Predictive_Modelling/01_Example_for_Learning_a_Decision_Tree~gxrPOF2R8QCJCBk0/most-recent (Accessed: 1 April 2025).

KNIME (2020) *Clustering Algorithms in KNIME*. 15 March. Available at: https://www.youtube.com/watch?v=243VC3qkM-A (Accessed: 5 April 2025).

KNIME (2024) *Numeric Scorer Node*. Available at: https://hub.knime.com/knime/spaces/Education/Educational%20Videos%20Demo%20Workflows/Numeric%20Scorer%20Node~tTSrwxvskTFPzm_X/current-state (Accessed: 5 April 2025).

KNIME (no date) *Linear Regression Learner (Statistics)*. Available at: https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.mine.regression.linear2.learner.LinReg2LearnerNodeFactory2 (Accessed: 11 April 2025).

Smolderen, K.G., Spertus, J.A., Nallamothu, B.K., Krumholz, H.M., Tang, F., Ross, J.S., Ting, H.H., Alexander, K.P., Rathore, S.S. and Chan, P.S., 2010. Health care insurance, financial concerns in accessing care, and delays to hospital presentation in acute myocardial infarction. *JAMA*, 303(14), pp.1392-1400.