**Introduction**

Healthcare costs continuously rise globally, with an increasing burden on healthcare systems, the need to manage resources efficiently while ensuring high-quality patient care as become of importance and to do this, we need to understand the factors that contribute to high healthcare costs

The objective of the report is to develop a predictive model that classifies patients into high or low healthcare cost based on attributes. This classification helps healthcare providers and health insurance providers to correctly identify which patients are most likely to incur higher costs and thereby provide a more targeted allocation of resources and treatment plans. By making use of historical data on patient demographics, health conditions, and utilization patterns, we can build a decision tree model that can assist in identifying high-cost patient.

In health care decision trees are advantageous because of their interpretability and ability to handle categorical and numeral data with the need to extensively preprocess making it perfect for large and diverse datasets, One can easily understand how factors influence each other and predictions

**Variable Description**

The dataset contains 42 columns and 4,084 entries, capturing various demographic, medical, and healthcare usage information.

The target variable” High\_low\_cost," is a binary label indicating whether a patient's total healthcare usage is classified as high (above the median) or low (below the median).

The Variables used to predict whether patients fall into high or low healthcare cost included

*Demographic data* – Age, Race, sex

*The Hierarchical Condition Category (HCC) score*- patient’s overall health risk

*Average Length of Stay (LOS)-* hospital stay duration

*Healthcare resource utilization* – visits to the ER, hospital outpatient visits, skilled nursing facility (SNF) use, and home health services, Enrolment Months, *Total Claims and Labs*

These were selected on that the bases that they would capture potential cost drivers and provide a holistic view of factors that affect healthcare costs

**Data Preprocessing**

Target Variable - The High\_Low\_Cost label was defined by calculating a Total Cost Indicator which is the sum of selected healthcare utilisation variables.

The threshold as the median with high cost being above median and low cost below median

Categorical variables such as Race and Sex were encoded using label encoding to ensure compatibility

Missing values were handled by Imputation with mean values

**Modelling**

The dataset was split into training and testing subsets - 70% training, 30% testing to evaluate model performance. It then trained a decision tree classifier with a fixed random state for reproducibility.

After training predictions on the test set, the model performance was evaluated using a detailed classification report

**Results**

**Decision tree**

A diagram of a network

Description automatically generated

***The Accuracy*** measures the percentage of correct predictions, The Decision tree achieved an accuracy of 92%

***The classification report*** provides a breakdown of the precision, recall, and F1-scores for both high and low-cost :

Classification Report:

precision recall f1-score support

0 0.90 0.93 0.92 596

1 0.93 0.90 0.92 630

accuracy 0.92 1226

macro avg 0.92 0.92 0.92 1226

weighted avg 0.92 0.92 0.92 1226

The decision tree model achieved a high accuracy and a balanced performance across both the classes,

The precision and recall were close to or above 90% for both the low-cost and high-cost predictions.

The high F1-scores indicate that the model is effective in managing both false positives and false negatives.

**The confusion matrix** provides insights on how well the mode; distinguishes between high cost and low cost

A diagram of a tree model

Description automatically generated

True Positives(TP)=556 cases( model correctly predicted the high-cost as positive)

True Negatives (TN)= 567 cases (model correctly predicted the low-cost as negative)

False Positives (FP)= 40 cases ( model incorrectly predicted the low-cost as high-cost)

False Negatives (FN= 63 (model incorrectly predicted the high-cost as low-cost)

The confusion matrix suggests there are some misclassifications however the model is overall reliable in distinguishing between high and low healthcare cost

A line graph with a point

Description automatically generated with medium confidence**ROC Curve** visualises sensitivity against the specificity

Precision, recall, and the AUC score provide a broader view of the model’s strengths and limitations

The ROC (Receiver Operating Characteristic) curve plots the true positive rate against the false positive rate.

The high AUC score of 0.92 indicates a strong predictive performance

**Conclusion**

This decision tree model provides an easy, interpretable way to classify healthcare costs.

The decision tree model performs high accuracy and AUC scores, indicating it can effectively distinguish between high and low-cost patients.

The use of metrics such as precision, recall, and the ROC curve helped to verify the model's suitability for healthcare cost prediction, making it a valuable tool for healthcare providers, given the cost implications of false negatives, further tuning or an ensemble approach might be considered to reduce these errors