# HEALTHCARE COST PREDICTION USING MACHINE LEARNING

MEDMINDS GROUP

*CONTENTS*

## 1. INTRODUCTION

Healthcare is a basic need nowadays that should be fair and equal for all. No matter where we live or how much money we have, everyone should have access to good healthcare.

Our project, MedMinds focuses on analyzing patient cost data in Canada across all the provinces. The data covers a six-year period from 2017 to 2022 that includes information about age, medical cases, length of stay, hospital cost and physician cost across different jurisdictions. By studying this data, we hope to discover trends in healthcare spending among different age groups and suggest how the system can be more efficient. What inspired us to choose this topic is how closely healthcare impacts all of us. Whether it's a hospital stay, a surgery, or a chronic condition for any age group, the cost of care plays a big role in the kind of treatment someone can receive. We think this project gives us a chance to make sense of real-world healthcare data and contribute ideas that can help improve the system in some way.

On the learning side, this project allows us to apply the tools we've been using, like Python, Excel, and Tableau, to work with a large and complex dataset. We'll be focusing on cleaning the data, analyzing the trends, building visualizations, and developing a predictive model. It's a great opportunity for us to grow our skills and understand how data can help solve real problems in healthcare.

Overall, this project is about more than just numbers. It's about learning how data can support better decision-making in healthcare, so resources are used wisely and everyone gets the care they need.

## 2. RELATED WORK

Healthcare cost analysis has been a key area of research, particularly due to the growing financial pressures on healthcare systems. Several studies have explored the factors driving healthcare costs, especially in countries like Canada, where the system is publicly funded.

In this section, we will compare the results of our project with those from similar studies in healthcare cost analysis to ensure that our findings are consistent with existing work.

The table below summarizes the machine learning models metrics from another groups study that focused on predicting healthcare costs:

| Machine learning models | R square | MSE |
|---|---|---|
| Random Forest | 93% | 707 |
| Decision Tree | 85% | 846 |

The Random Forest (RF), which combines various algorithms achieved the highest accuracy in this study, followed closely by Decision Tree (DT).

We use these results as a baseline to assess our model's performance. While we've employed different algorithms in our analysis, comparing our metrics with this work allows us to evaluate how well our models perform in comparison to previous attempts.

By examining our results in light of their findings, we can confirm whether our models are producing similar or better outcomes in predicting healthcare costs and if our approach contributes new insights or improvements to the field.

## 3. METHODS

This section outlines the key steps we followed in our project, including the methods used for exploratory data analysis, data preprocessing techniques, the machine learning models we applied, and the evaluation metrics chosen to assess our model's performance.

### 3.1 Importing libraries, dataset, and packages:

We began by importing all the necessary libraries and packages needed to run different parts of the code in our Python notebook (.ipynb file). After that, we loaded our dataset and saved it into a data frame to prepare it for further analysis.

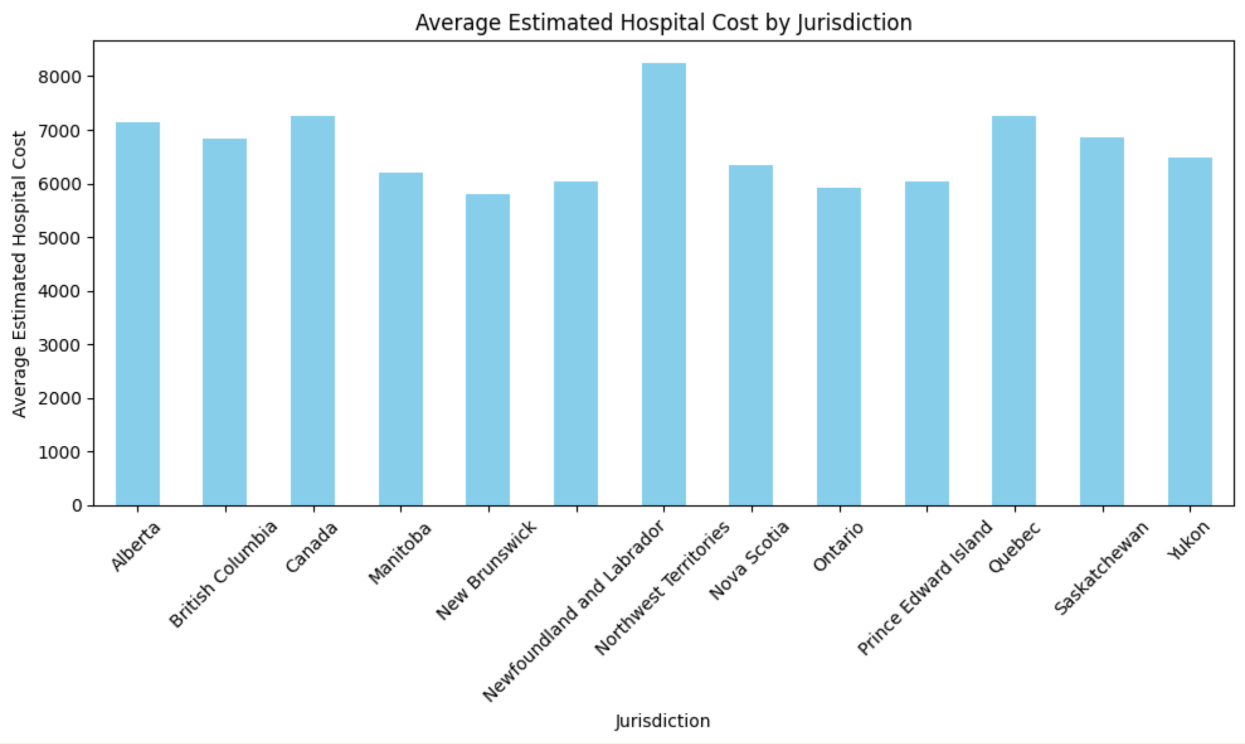
### 3.2 Data cleaning:

Data preprocessing is an important step where we clean and prepare raw data, so it becomes suitable for analysis and machine learning. Raw data often contains issues like missing or incorrect values, so preprocessing helps remove such problems and keeps only the relevant information. To better understand the structure of the dataset, we used basic functions like head, tail, describe, and info to check data types, non-null values, counts, mean, standard deviation, and quartile ranges.

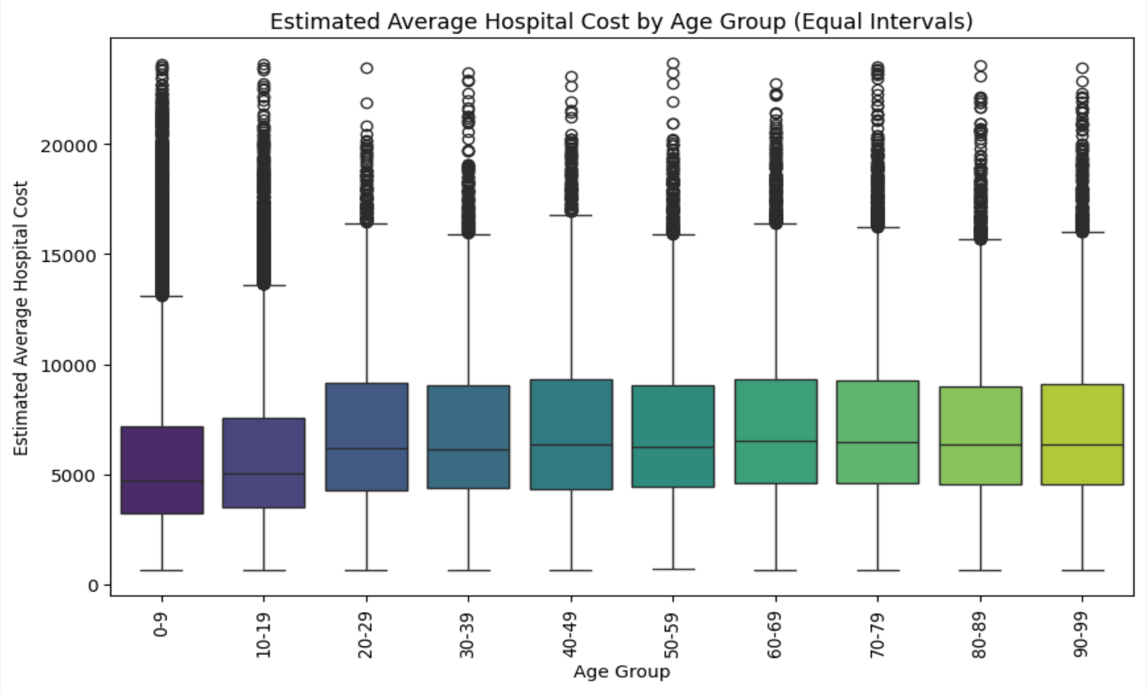
In our case, we began by checking for any missing or null values in the dataset using the isna.sum() method. Fortunately, our dataset had no null values or duplicate records. The only issue we found was the presence of outliers, which we handled separately. Outliers are values that are very different from the rest of the data. We checked for them to find any unusual or incorrect entries in our dataset. Apart from that, the dataset was clean and ready to be used for further analysis and modeling. After removing outliers, we had 74547 rows and 17 columns.

### 3.3 Exploratory analysis of the data:

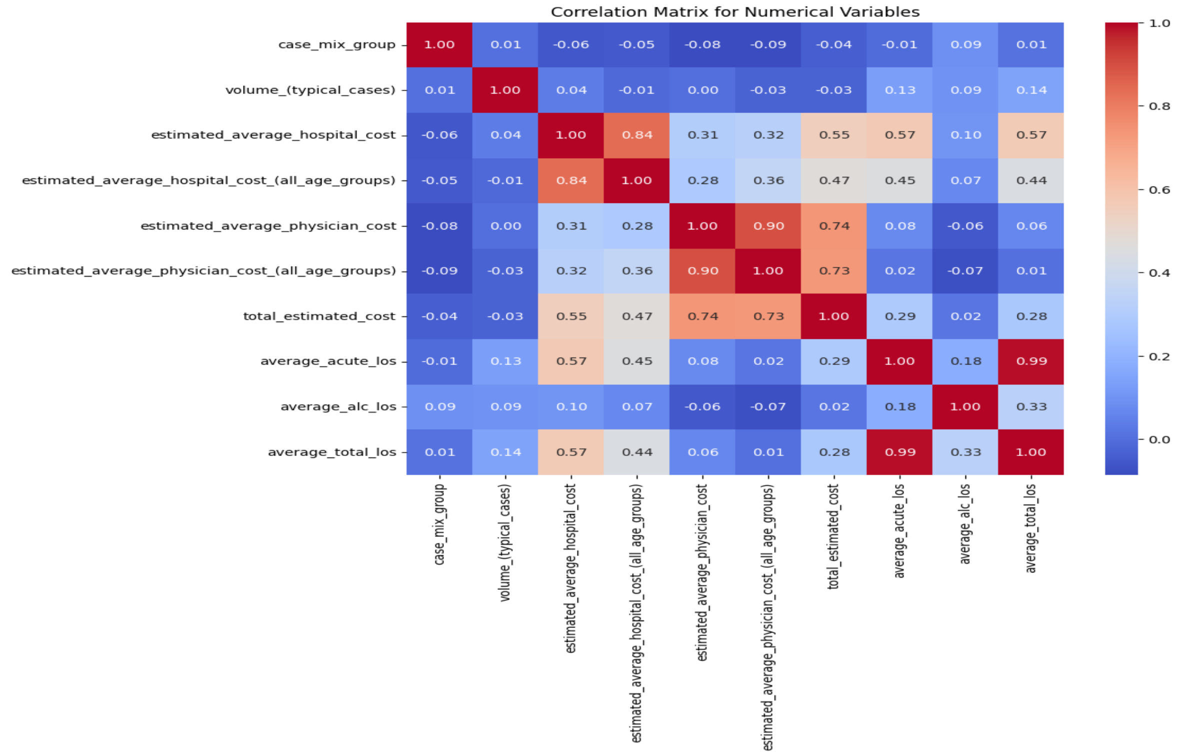
We started our data exploration by examining the contents of the data frame to get familiar with the variables present. This helped us identify the types of columns we have and determine which ones are most relevant for building our prediction model. The following chart depicts average estimated hospital cost for different jurisdictions. Newfoundland shows the highest estimated hospital cost compared to other provinces.



Further the hospital costs of different age groups were shown in boxplot to understand which age group has the highest estimated hospital cost. This plot helps us see that hospital costs tend to slightly increase with age, and that outliers (high-cost cases) are common in all groups. The data shows a wide cost range, which is important to keep in mind when building predictive models.



Next, we looked into the distribution of both numerical and categorical features using visualizations to gain more insights into the data. We also created a heatmap to show the correlation matrix, which helped us quickly identify which features had strong or weak relationships with each other.



### 3.4 Statistical analysis:

Hypothesis testing was done using chi-square, F-test(Anova).

1. Chi-square:

$H_0$: 'age group' and 'jurisdiction' are dependent
$H_{a:}$ 'age group' and 'jurisdiction' are not dependent

RESULTS: Null hypothesis is rejected -age groups and jurisdictions are dependent.

There is a statistically significant association between age group and jurisdiction, meaning that the distribution of age groups varies by jurisdiction. In other words, certain jurisdictions may have a higher concentration of certain age groups, which could influence healthcare resource planning or cost patterns.

2. F - Test:

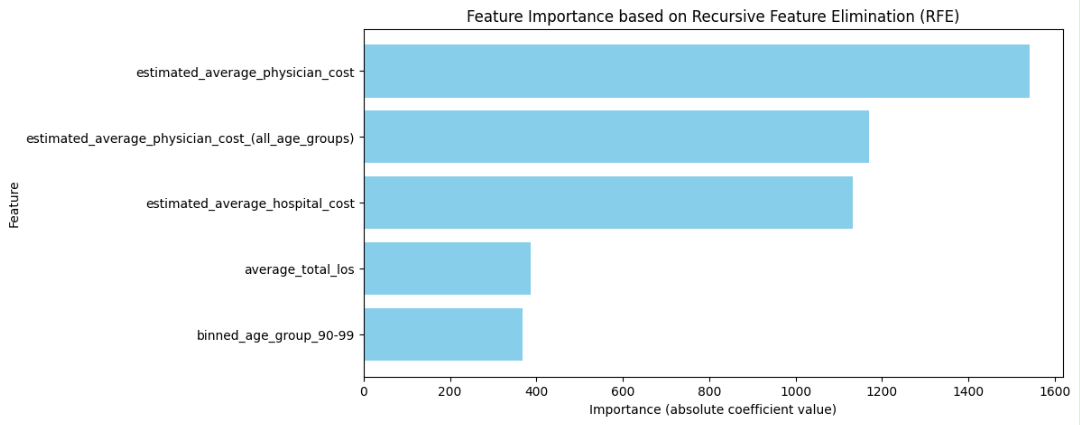
$H_0$: There is a significant difference in the mean hospital cost across age groups
$H_a$: There is no significant difference in the mean hospital cost across age groups

RESULTS: Null hypothesis is rejected – These is a significant difference in the mean hospital cost across age groups.

There is a statistically significant difference in average hospital costs between different age groups. This means age has a notable effect on hospital costs, and some age groups (likely older ones) may incur higher healthcare expenses than others.

### 3.5 Feature Selection:

Feature selection was performed using both Recursive Feature Elimination (RFE) and embedded methods to identify the most relevant predictors for our model. This helped in improving model performance and interpretability by eliminating less important variables.



Both RFE and embedded methods consistently highlighted healthcare cost variables as the most significant predictors, particularly physician and hospital costs. These insights suggest that cost-related features are the key drivers in the model, while age and length of stay play a secondary role.

### 3.6 Model Building:

To predict the total estimated cost of healthcare, a Random Forest Regression model was developed using a structured modeling pipeline.

The target variable was total estimated cost, which was separated from the feature set. Categorical variables were transformed using one-hot encoding. PyCaret's regression module was used to perform feature selection. After initializing setup, the top predictive features were automatically identified and retrieved. The data was split into 80% training and 20% testing using train test split with a fixed random state to ensure reproducibility. A Random Forest Regressor with 100 estimators was trained on the selected features from PyCaret.

### 3.7 Model Deployment:

After training and evaluating our predictive model, we proceeded with deployment to demonstrate its practical use. We created a user-friendly interface using flask, which allows healthcare professionals or analysts to input relevant patient data such as age, medical case, and length of stay and receive a predicted cost estimate in real-time.

This step showcases how machine learning can be integrated into real-world applications, making data-driven decision-making more accessible. It also helps simulate how hospitals or policymakers might use such a model to plan budgets, predict patient-specific costs, or allocate resources more efficiently.

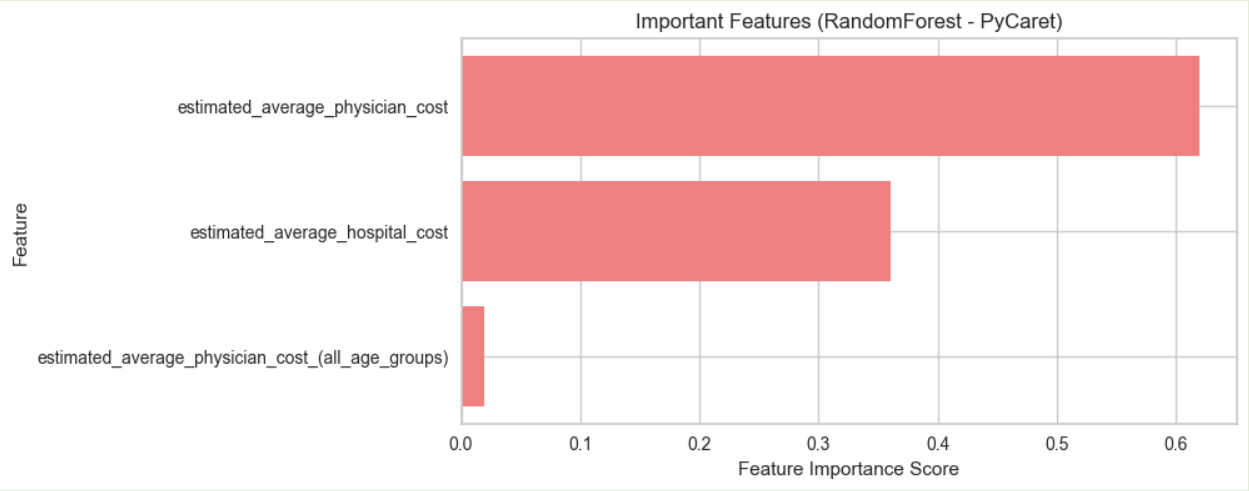
## 4. Results:

The model achieved an R-squared value of 0.99, indicating that it explained 99% of the variance in the target variable — an excellent fit. The Mean Squared Error (MSE) was 707, reflecting a low average squared difference between predicted and actual values.

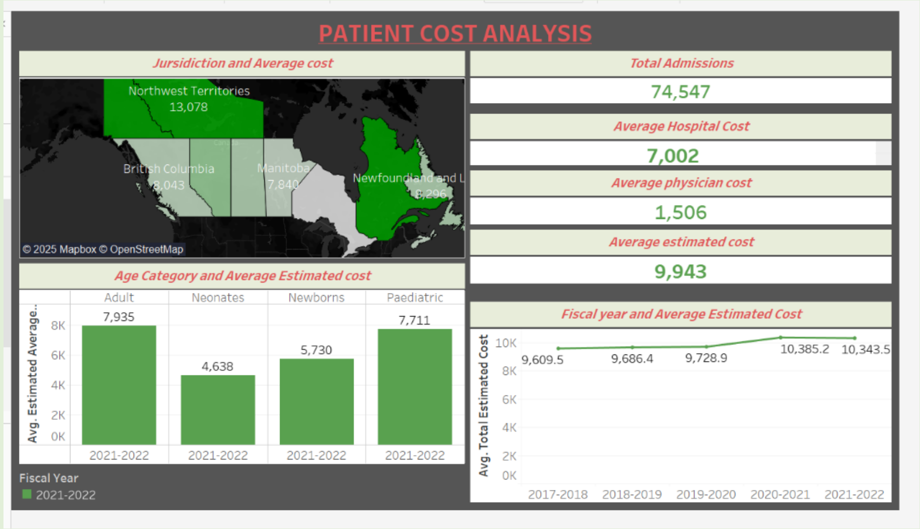
The Random Forest model performed exceptionally well in predicting healthcare costs, with high accuracy and minimal error. The robust preprocessing, combined with effective feature selection using PyCaret, contributed to building a reliable and interpretable predictive model.

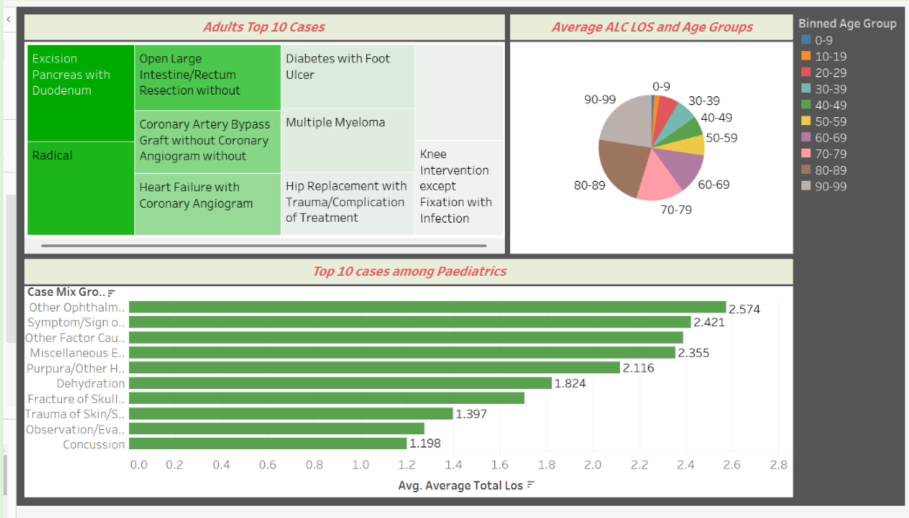
| **Metric** | **Value** |
|---|---|
| R-square | 0.99 |
| MSE | 707 |

The important features taken into consideration to estimate total cost was physician cost, and hospital cost.



From tableau we have following dashboards that gives us insight of: Healthcare costs vary significantly by region, age group, and case type, Elderly and adults incur higher average costs and longer stays, Pediatric cases generally involve shorter hospital stays but still span a range of conditions, Costs have seen a gradual rise over the years.





- The average patient healthcare cost in Canada for 2021–2022 was $9,943, with the Northwest Territories having the highest cost at $13,078, indicating significant regional disparities.
- Adults and paediatric patients had the highest average estimated hospital costs among all age groups, with adults at $7,935 and paediatrics at $7,711, while neonates had the lowest at $4,638.
- Over five fiscal years, average estimated costs have steadily increased, peaking in 2020–2021 and slightly declining in 2021–2022.
- Among adults, complex procedures like pancreas excision, coronary bypass, and hip replacements dominate in both frequency and cost.
- In paediatrics, common conditions like ophthalmologic issues and respiratory symptoms are associated with the longest hospital stays, indicating the need for targeted pediatric care planning.

## 5. DISCUSSION:

The objective of analyzing patient healthcare costs across jurisdictions, age groups, and clinical conditions was successfully achieved using interactive Tableau dashboards. The findings revealed significant regional disparities, with the Northwest Territories incurring the highest average patient cost, and variations across age groups, where adults and paediatrics bore the highest average estimated costs. Interestingly, neonates had lower average costs compared to newborns, suggesting that many neonates may not require extended care, or that newborns include a broader age range with more complex conditions.

Throughout the project, several challenges were encountered. The dataset contained inconsistent and outliers, requiring extensive preprocessing to ensure accurate analysis. Additionally, medical codes were complex, demanding careful interpretation and external validation. We also had to manage outliers and skewed cost distributions, particularly in high-cost cases, to avoid biasing the overall insights. These were addressed through data cleaning techniques and robust visual validation in Tableau.

Moreover, the project was completed under tight time constraints, which limited the depth of model testing and iteration. Despite these hurdles, the analysis produced meaningful results aligned with healthcare cost trends and laid the groundwork for further investigation. Future work could benefit from deeper patient-level data and predictive modeling to enhance resource planning and cost optimization.

## 6. CONCLUSION:

This project provided a comprehensive analysis of patient healthcare costs across Canada, leveraging both data visualization and machine learning techniques. By integrating predictive models, we enhanced the accuracy of cost estimations and uncovered key patterns related to age groups, regional variations, and clinical categories. These insights not only validated existing healthcare trends but also revealed opportunities for efficiency improvements, especially in high-cost and high-volume areas. The findings support the value of data-driven approaches in strategic healthcare planning, helping to inform better decision-making and optimize resource allocation across hospital systems. Overall, this work demonstrates the impact of combining analytics with machine learning to address real-world healthcare challenges.

## 7. CONTRIBUTION:

| Names | Contributions |
|---|---|
| Prabhleen Kaur (0857194) | • Assisted in drafting the project proposal, organizing initial research, and beginning the data cleaning process.<br>• Conducted Exploratory Data Analysis (EDA), identified missing data patterns, and visualized key trends.<br>• Conducted statistical and hypothetical analysis on key variables.<br>• Focused on building the model without using PyCaret.<br>• Lead the deployment of the model into a production environment |
| Rajwinder Kaur (0831280) | • Contributed to writing and structuring the project proposal.<br>• Detected and handled outliers in the dataset, ensuring data integrity.<br>• Performed feature transformation and scaling for model optimization.<br>• Worked on building the model using PyCaret.<br>• Focus on ensuring the model is accessible via an API for end-users. |
| Ajay Haji Korbe (0852660) | • Reviewed the methodology and statistical techniques that will be used in the analysis.<br>• Assisted with EDA and data cleaning steps in depth.<br>• Implemented initial statistical models and tested their performance.<br>• Worked on hyperparameter tuning of the model without PyCaret.<br>• Work on monitoring and evaluating the model's performance post-deployment. |
| Thrivikram Sai Teja (0856337) | • Conducted literature review, finalized sections of the proposal, and assisted with initial data cleaning.<br>• Assisted with finalizing outlier. |

| | |
|---|---|
| | • Assist with integrating the model into a user-friendly interface.<br>• Worked on hyperparameter tuning of the model using PyCaret.<br>• Validated data preprocessing and ensured dataset readiness for modeling. |

## 8.REFERENCES:

1. Canadian Institute for Health Information (CIHI). Patient Cost Estimator | CIHI
2. Patient Cost Estimator: Methodology Notes and Glossary
3. https://www.rand.org/pubs/commentary/2012/03/the-real-cost-of-healthcare.html

## 9.APPENDICES:

- Healthcare Project-MedMinds: This is jupyter notebook file that includes data preprocessing, machine learning model and deployment.
- healthcare dashboard -MedMInds: This is tableau dashboard and charts done (using business intelligence tool)
- Healthcare cost prediction using machine learning.docx: This contains final report.