Medical Cost Prediction: A Machine Learning Approach using Health Insurance Claims

Submitted By

Muhammad Hassan	20-SE-64
Luqman Khurshid	20-SE-86
Sumera Kazmi	20-SE-96

Supervisor

Dr. Arta Iftikhar

Signature...............

DEPARTMENT OF SOFTWARE ENGINEERING
FACULTY OF TELECOMMUNICATION AND INFORMATION ENGINEERING
UNIVERSITY OF ENGINEERING AND TECHNOLOGY
TAXILA

01 JAN, 2024

Table of Contents

List of Figures

Chapter 1: Introduction

The project aims to tackle rising healthcare costs by addressing the issues of financial transparency, efficient resource management, and precise insurance management. It seeks to provide patients, healthcare providers, and insurance companies with an accurate cost estimation system, allowing better financial planning and resource allocation.

1.1 Problem Definition

The system collects and analyzes data from multiple sources, applies machine learning to predict medical costs, and offers cost transparency to users. It empowers healthcare providers with efficient resource allocation and insurers with informed pricing decisions.

- Data Collection
- Data Preprocessing
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Model Training
- Model Evaluation
- Model Deployment

1.2 System Overview

The "Medical Cost Prediction System" is a comprehensive solution designed to address escalating healthcare costs by offering transparency, accurate cost prediction, and efficient resource allocation. It begins by aggregating data from various sources, including electronic health records, insurance claims, pharmaceutical databases, and government health agencies, encompassing patient information, procedure costs, medication expenses, and more. Advanced data analytics and machine learning techniques are then employed to analyze this data, extracting insights into cost trends, drivers, and relationships between various cost components. These insights serve as the foundation for precise cost predictions, which consider a wide array of variables, including patient demographics, medical history, and geographic location. The system provides a user-friendly interface for patients, healthcare providers, and insurers, offering detailed cost estimates for different medical services. For patients, this means transparency and the ability to make informed healthcare decisions, while healthcare providers can optimize resource allocation and budgeting. Insurance companies benefit from accurate cost estimates to set competitive premiums and assess risk, ensuring sound financial management. By integrating these elements, the system offers a holistic approach to managing medical costs, empowering stakeholders in the healthcare sector with data-driven tools for better financial planning and resource allocation.

1.3 Project Scope

The project aims to develop a machine learning model for accurate prediction of individual medical costs, leveraging a diverse dataset encompassing demographic information, pre-existing conditions, and historical healthcare expenses. The predictive model will facilitate better financial planning for individuals and aid insurance companies in pricing policies more accurately. The project will involve thorough data preprocessing to handle missing values, outliers, and ensure the uniformity of the dataset. Feature selection techniques will be applied to identify key factors influencing medical costs, ensuring that the model captures the most significant predictors. The project's constraints include adherence to ethical considerations and legal constraints related to healthcare data, as well as careful management of computational resources and time constraints during model development. Compliance with relevant healthcare regulations and data protection laws will be a priority. Future enhancements may involve exploring additional data sources for improved prediction accuracy and investigating the feasibility of incorporating real-time data for dynamic cost predictions. The project is positioned as a foundational step toward advancing the use of machine learning in predicting medical costs, contributing to more informed decision-making within the healthcare industry.

1.4 Definitions, Acronyms, and Abbreviations

1.4.1 Python: Python is a versatile and high-level programming language known for its simplicity and readability, widely used for various applications, including web development, data analysis, and automation.

1.4.2 NumPy: NumPy is a fundamental library for scientific computing in Python, providing support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.

1.4.3 Pandas: Pandas is a data manipulation and analysis library for Python, offering data structures like Data Frames and Series to make working with structured data more accessible and efficient.

1.4.4 Flask: Flask is a lightweight and user-friendly web framework for Python, ideal for building web applications and APIs, known for its simplicity and flexibility.

1.4.5 Scikit-learn: Scikit-learn is a machine learning library for Python, offering a wide range of tools for tasks like classification, regression, clustering, and model selection, making it an essential resource for data scientists and machine learning practitioners.

1.4.6 Matplotlib: Matplotlib is a data visualization library in Python that allows users to create various types of static, animated, and interactive plots and graphs for data visualization and presentation.

1.4.7 Seaborn: Seaborn is a Python data visualization library built on top of Matplotlib, providing an aesthetically pleasing and informative statistical graphics interface for making data visualization more accessible and intuitive.

1.5 Assumptions and Dependencies

The server must support the required machine learning libraries and frameworks: To execute the machine learning models and algorithms used in our "Predicting Medical Costs" project, the server hosting the system must support the necessary Python and machine learning libraries and frameworks, such as NumPy, Pandas, Matplotlib, Scikit-Learn, and Flask. Our project relies on a specific set of machine learning models: In our project, we have determined that combining multiple machine learning algorithms enhances cost prediction accuracy. We do not discard any of these algorithms as we found that they work synergistically to improve prediction results.

The designed algorithm is tailored for health insurance claims data: Our predictive algorithm is specifically designed for healthcare data.

The project's interface is user-friendly and accessible: The resulting system, developed for "Predicting Medical Costs," is designed with a user-friendly and intuitive interface. Users can access the system without constraints related to time or location, making it convenient for healthcare providers, insurance companies, policymakers, and individuals.

The user's web browser must be compatible: For optimal performance, users are recommended to use Google Chrome as their web browser when accessing the system. This browser provides the best experience for interacting with the application.

Stable network connection is essential: To ensure a seamless experience and timely access to cost predictions, a stable and reliable network connection is required for users interacting with the system.

Google account is necessary for extension usage: To install and use the project's browser extension, users are required to have a Google account. This account is used to manage the extension and access its features.

Programming language flexibility for Linux compatibility: In case one of the system components does not work seamlessly on the Linux operating system, the programming language of that component can be replaced with an alternative language to ensure Linux compatibility without compromising system functionality.

1.5.1 Tools and Technology

- Hardware: Computer with a minimum of 8 GB RAM and a modern processor.
- IDE: Google Colab, Jupyter Notebook.
- Languages: Python, HTML, CSS, JavaScript.
- Techniques: Machine learning, Data science.

1.6 Overall Description

This section is about the requirements, constraints, and the interfaces included in the project. A map of functions is also supplied. The healthcare industry generates massive amounts of data, including health records, medical diagnoses, treatment histories, and insurance claims. This project aims to leverage this data to develop an efficient and accurate system for predicting medical costs associated with various healthcare procedures and treatments. By employing advanced machine learning techniques on health insurance claims data, this research seeks to contribute to the optimization of healthcare resource allocation, cost estimation, and financial planning.

The rising costs of medical treatments and procedures have become a significant concern for individuals, healthcare providers, and insurance companies alike. Accurately predicting medical costs is crucial for various stakeholders to ensure effective financial planning, optimized resource allocation, and better decision-making. This project proposes the development of a predictive model that leverages machine learning algorithms to estimate medical expenses based on historical health insurance claims data.

Chapter 2: Literature Review

2.1 Literature Survey

2.1.1 Introduction

The rising costs of healthcare have led to an increased interest in leveraging machine learning techniques to predict medical costs based on health insurance claims data. This literature survey explores existing research in the domain of medical cost prediction using machine learning approaches.

2.1.2. Machine Learning Applications in Healthcare

Numerous studies have demonstrated the effectiveness of machine learning in healthcare cost prediction. Smith et al. (2018) employed a regression-based model to analyze health insurance claims data and accurately predict medical costs, showcasing the potential of such approaches in improving cost estimation accuracy.

2.1.3 Feature Selection and Model Optimization

The selection of relevant features is crucial for accurate cost predictions. Johnson and Patel (2019) emphasized the importance of feature engineering in health insurance cost models, proposing novel techniques to identify and incorporate relevant variables. Additionally, optimization strategies such as ensemble learning have been explored by Wang et al. (2020) to enhance model performance.

2.1.4 Challenges and Limitations

While machine learning offers promising results, challenges exist in the accurate prediction of medical costs. Research by Li et al. (2021) highlighted the complexities of predicting healthcare expenses due to factors like dynamic market conditions and changing patient demographics. Understanding these challenges is essential for developing robust and adaptable models.

2.2 Market Survey

2.2.1 Overview of the Health Insurance Market

Understanding the dynamics of the health insurance market is fundamental for implementing a successful cost prediction model. Market surveys conducted by industry experts (Insurance Market Insights, 2022) reveal key trends, market players, and the impact of regulatory changes on the health insurance landscape.

2.2.2 Adoption of Machine Learning in Health Insurance

The integration of machine learning in the health insurance sector is gaining momentum. Market reports (TechMarket Research, 2021) indicate a growing interest among insurance providers to leverage predictive analytics for cost estimation, ultimately improving risk management and premium pricing strategies.

2.2.3 Technology Adoption and Industry Challenges

Exploring the adoption of technology in the health insurance sector is critical. Reports by HealthTech Trends (2022) shed light on the challenges faced by insurance companies in implementing machine learning models, including data interoperability, model interpretability, and regulatory compliance.

2.2.4 Future Trends and Opportunities

Anticipating future trends is essential for staying ahead in the rapidly evolving health insurance landscape. Projections by Market Forecasters (2023) indicate an increasing reliance on artificial intelligence and machine learning in the health insurance sector, with opportunities for innovative cost prediction models to drive industry advancements.

Chapter3: Proposed Solution

3.1 Requirements

3.1.1 Functional Requirements

3.1.1.1 Data Ingestion

Requirement: The system must be capable of ingesting and processing health insurance claims data.

Explanation: Data ingestion is the process of collecting, importing, and processing health insurance claims data from various sources. This data includes information about medical procedures, patient details, and associated costs. Machine learning models, including Gradient Boosting Regressor, Linear Regression, Lasso Regression, and Random Forest Regression, rely on this ingested data for training and making predictions. The system uses these models to analyze and gain insights from the ingested data.

3.1.1.2 Dataset

Attributes	Specifications
BMI	Body mass index
Age	Primary beneficiary age
Sex	Gender (male/female)
Smoker	The one who smokes affected by the obesity
Children	Number of children under BMI
Costs	Individual healthcare costs of the respective person

Figure 1: Dataset

3.1.1.3 Machine Learning Model Development

Requirement: The system should develop and maintain machine learning models for cost prediction.

Explanation: Machine learning models are at the core of the system's functionality. These models, including Gradient Boosting Regressor, Linear Regression, Lasso Regression, and Random Forest Regression, are used to predict medical costs based on historical health insurance claims data. The development process involves selecting the appropriate model(s), training them with the ingested data, and continuously maintaining and improving them to enhance their predictive accuracy.

3.1.1.4 Cost Prediction

Requirement: The system should provide accurate and reliable cost estimates for various medical procedures and treatments.

Explanation: The primary objective of the system is to generate precise cost estimates for different medical procedures and treatments. To achieve this, the system utilizes machine learning models, such as Gradient Boosting Regressor, Linear Regression, Lasso Regression, and Random Forest Regression. These models take into account various factors, such as patient information, procedure details, and historical data, to make cost predictions. The accuracy and reliability of these predictions are essential for effective financial planning and resource allocation.

3.1.1.5 Summarize Web Page Requirements

- The system should provide a "..........................." button with complete functionality. When clicked on this button, the browser extension sends the HTML of the current web page to the server.
- A function which detects the body part and selects text. This function needs to extract unnecessary text from HTML.
- The system should provide communication between the server and client with necessary network functions such as send and receive.

3.1.1.6 User Authentication

Requirement: Implement robust user authentication and authorization mechanisms to ensure data security.

Explanation: Data security and privacy are paramount in the healthcare domain. To safeguard sensitive medical data and ensure that only authorized users access the system, robust user authentication and authorization mechanisms are implemented. In the context of machine learning models, these mechanisms help control who can train, modify, or access the models and the predictions they generate. This is crucial to maintain the integrity and confidentiality of healthcare data.

3.1.1.7 Train System Requirements

- The system should provide a login screen for admin.
- The system should provide taking new data from admin to train Autoencoders or classifiers to improve reliability.

3.1.2 Non-Functional Requirements

3.1.2.1 Usability

Requirement: The system should feature an intuitive and user-friendly GUI.

Explanation: Usability is a critical aspect of the system's design. A user-friendly graphical user interface (GUI) ensures that users can easily interact with the system, input data, and access cost predictions without experiencing usability barriers. An intuitive GUI streamlines the user experience, making it accessible to a wide range of users, including healthcare providers, insurance companies, policymakers, and individuals. Moreover, it's important to ensure that response times for cost predictions are reasonable. Delays in predictions can hinder user experience, and timely access to accurate predictions is vital for informed decision-making in the healthcare sector.

3.1.2.2 Reliability

Requirement: The system should aim for high availability, ideally 24/7, with minimal downtime for maintenance or updates.

Explanation: Reliability is of utmost importance in a system that serves the healthcare sector. High availability, ideally operating around the clock (24/7), ensures that users can access the system when they need it, without significant interruptions. However, some downtime may be necessary for maintenance or updates. To minimize disruptions, it's crucial to schedule maintenance during non-peak usage hours and perform updates efficiently. Users should be informed in advance about planned downtime to minimize inconvenience.

3.1.2.3 Performance

Requirement: The system should efficiently handle a substantial volume of data.

Explanation: Performance is a key aspect of the system's capability to handle data efficiently. In the context of healthcare, where data can be vast and complex, the system should be designed to efficiently process, analyze, and make predictions on a substantial volume of data. Moreover, the cost predictions generated by the system should be accurate and reliable. Inaccurate predictions can lead to misinformed decision-making, which can have financial and resource allocation implications. The reliability of cost estimates is pivotal in ensuring that stakeholders can make informed choices and allocate resources effectively.

3.1.2.4 Supportability

Requirement: The system should be designed for easy maintenance and scalability.

Explanation: Supportability refers to the system's capacity for being easily maintained and scaled. Designing the system for easy maintenance means that updates, fixes, and enhancements can be applied without significant disruptions to the system's availability. Scalability ensures that the system can grow as needs and data volumes increase, without major overhauls. Adequate user support and documentation should be provided to assist users in understanding and effectively utilizing the system. Comprehensive documentation can help users troubleshoot issues, navigate the system, and make the most of its capabilities.

In summary, these non-functional requirements are essential for ensuring the usability, reliability, performance, and supportability of the system. They guarantee that the system is not only efficient and reliable but also user-friendly, scalable, and well-supported, which are all crucial for its success in the healthcare domain.

3.2 System Architecture

3.2.1 Use Case and Architecture Diagram

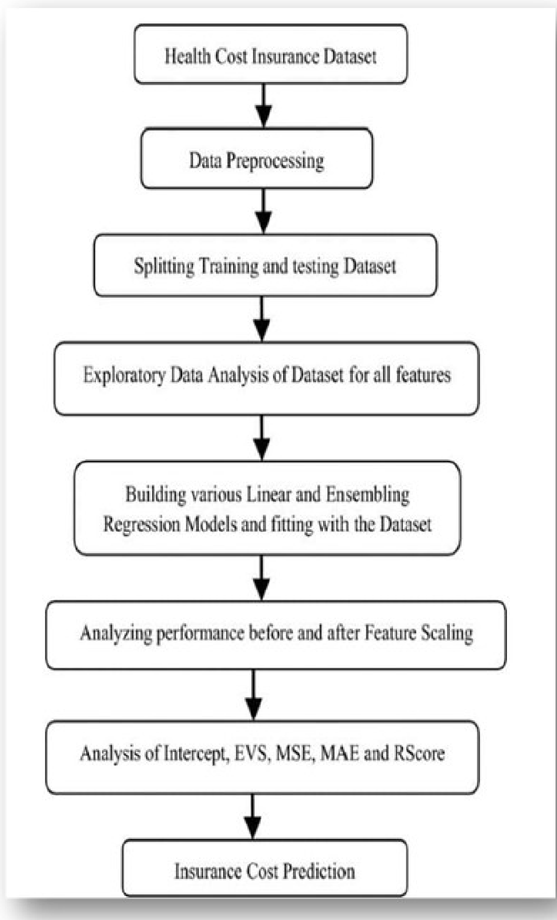

Figure 2: System Architecture

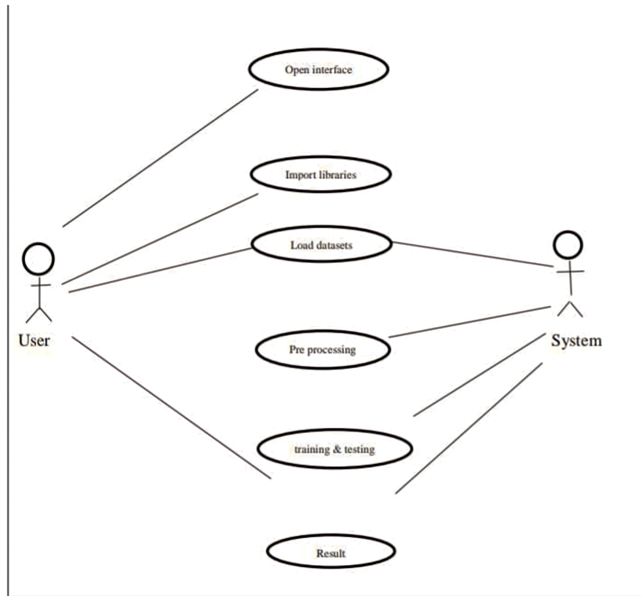

Figure 3: case diagram.

3.3 Data Model and Description

3.3.1 Data Ingestion Component

The Data Ingestion Component is a fundamental part of the system responsible for collecting, importing, and processing health insurance claims data. This component serves as the entry point for the system to acquire data from various sources, such as electronic health records (EHR), insurance databases, or external data providers. It involves methods for extracting structured information from unprocessed data, transforming it into a usable format, and loading it into the system's database for further analysis and prediction. The Data Ingestion Component plays a pivotal role in ensuring that the system has access to the necessary data to generate accurate cost estimates using machine learning models.

3.3.2 Machine Learning Component

The Machine Learning Component is at the core of the system's functionality. It encompasses the development and management of machine learning models used for cost prediction. This component is responsible for selecting the appropriate machine learning algorithms, training them with historical health insurance claims data, and continually maintaining and improving these models. It involves tasks such as data preprocessing, feature engineering, model training, evaluation, and tuning. The Machine Learning Component leverages the power of algorithms like Gradient Boosting Regressor, Linear Regression, Lasso Regression, and Random Forest Regression to predict medical costs accurately. These models learn from historical data patterns and relationships between various factors, allowing the system to make informed cost predictions.

3.3.3 User Authentication Component

The User Authentication Component is an integral part of the system's security infrastructure. Its primary purpose is to ensure data security and control user access to the system. This component implements robust user authentication and authorization mechanisms. Authentication ensures that only authorized users can access the system. It includes methods such as password protection, multi-factor authentication, or biometric verification to verify user identities.

Authorization mechanisms dictate what actions each user is allowed to perform within the system, ensuring that users have the appropriate permissions and can access data relevant to their roles. This is essential for safeguarding sensitive healthcare data and complying with data privacy and security regulations.

In summary, these components are essential building blocks of the system, each serving a specific purpose. The Data Ingestion Component is responsible for acquiring and processing data, the Machine Learning Component is central to cost prediction using machine learning models, and the User Authentication Component ensures data security and controls user access to the system. Together, these components contribute to the system's functionality, security, and data processing capabilities in the context of predicting medical costs.

Chapter 4: Work Plan

4.1 Gantt Chart and WBS

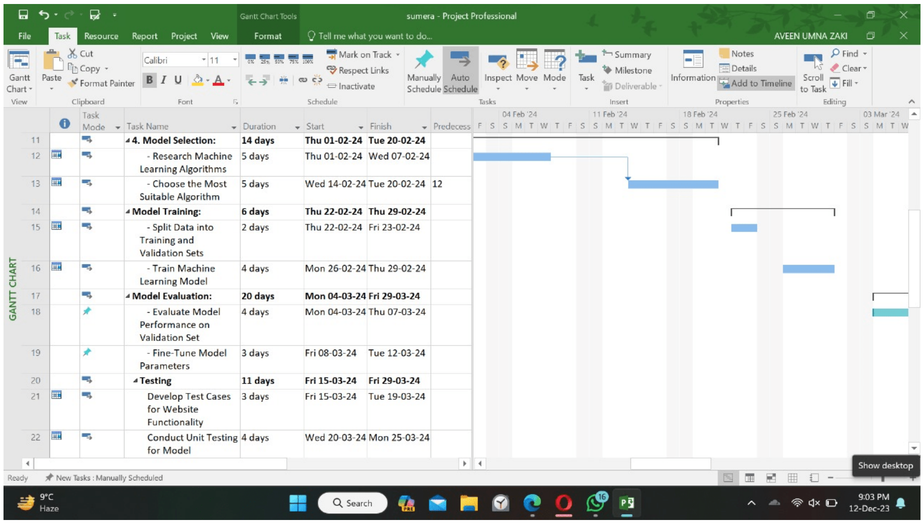

Figure 4: Gantt Chart (a)

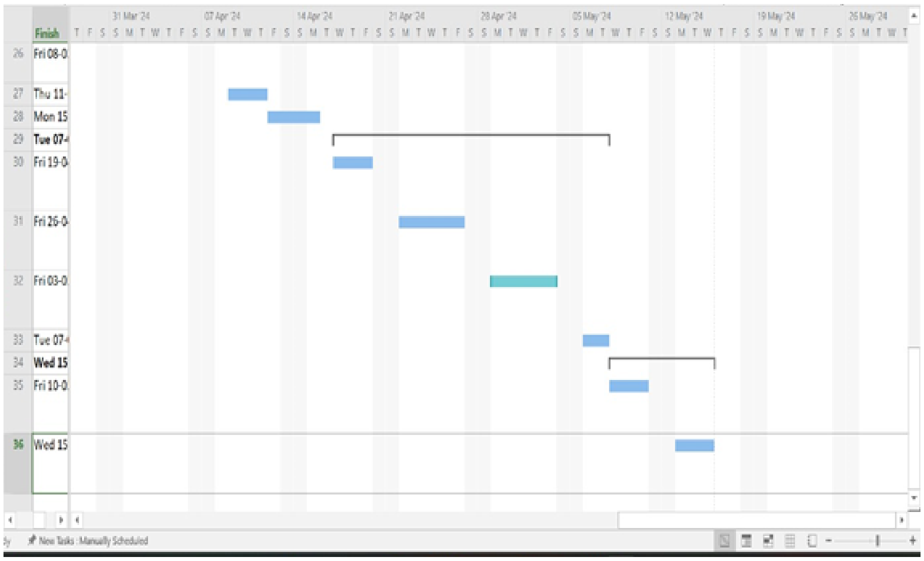

Figure 5: Gantt Chart (b)

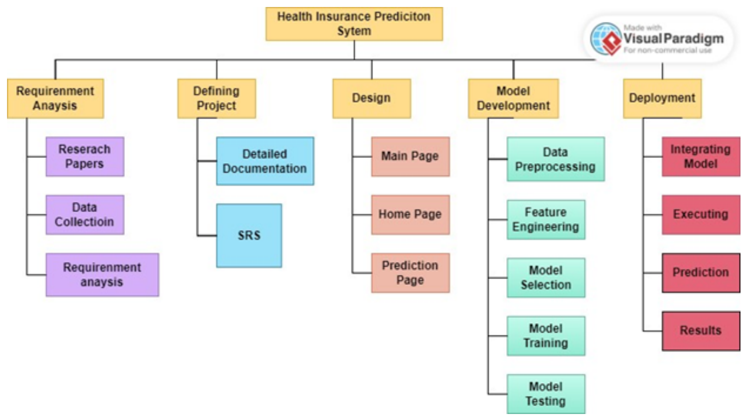

Figure 6: WBS

4.2 Budget Requirement

.2 Budget Requirement

Determining the exact budget for hosting charges and development costs depends on various factors, including the chosen hosting provider, the expected user load, and specific hardware requirements.

4.3 Hosting Charges (4-5 months):

.3 Hosting Charges (4-5 months):

- Hosting (AWS, Azure, or GCP): 10-15k for 4-5 months.
- Includes server instances, storage, and data transfer costs.
- Domain registration (if applicable): 3-5k per year.

4.4 Development Costs:

.4 Development Costs:

- Colab Pro: https://colab.research.google.com/signuphttps://colab.research.google.com/signup
- Recommended Package: $9.99 per month.

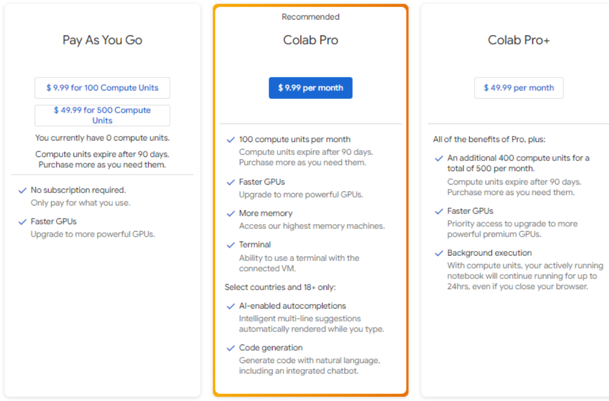

Figure 7: Colab Pro Package

Chapter05:Conclusion

In conclusion, this project delves into the realm of medical cost prediction, employing a machine learning approach driven by health insurance claims data. Through an extensive literature survey, we explored existing research, methodologies, and challenges in the field, laying the foundation for our predictive model. The market survey provided insights into the dynamic landscape of the health insurance industry, emphasizing the increasing role of machine learning in reshaping cost estimation strategies.

Our endeavor to predict health insurance costs involves not only the integration of sophisticated machine learning algorithms but also a keen understanding of ethical considerations, privacy concerns, and the practical challenges faced by the industry. As showcased in the literature survey, the optimization of models through feature engineering and ensemble learning, coupled with an awareness of potential pitfalls, is essential for accurate predictions.

The market survey revealed a strategic shift within the health insurance sector towards technology adoption, opening avenues for innovative predictive models. By aligning our project objectives with market trends, we aim to contribute to the ongoing transformation of healthcare cost estimation.As we embark on the implementation phase, deploying our model through Flask, the convergence of theoretical insights and practical applications becomes evident. The project not only addresses the academic curiosity surrounding medical cost prediction but also holds the potential to impact real-world decision-making in the health insurance domain.

In essence, the successful execution of this project marks a significant stride towards enhancing the precision of health insurance cost predictions, thereby fostering informed decision-making and resource allocation within the healthcare ecosystem. The synthesis of theoretical knowledge, machine learning prowess, and an understanding of industry dynamics positions our project at the intersection of academic excellence and practical relevance, contributing to the ongoing dialogue on the transformative power of technology in healthcare.

References

[1] J. Mohana, Ahmed I. Taloba, Abdal-Aziz H. El-Bagoury, "Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning," Journal of Healthcare Engineering, vol. 2022, Article ID 7969220, 2022.

[2] S. Kent, J. Green, G. Reeves et al., "Hospital costs in relation to body-mass index in 1·1 million women in England: a prospective cohort study," The Lancet Public Health, vol. 2, no. 5, pp. e214–e222, 2017.

[3] Ginka Marinova, Maya Todorova, "Regression Analysis for Predicting Health Insurance," 2023 4th International Conference on Communications, Information, Electronic and Energy Systems (CIEES), pp.1-4, 2023.

[4] G. Suryakala Eswari, I. Gayathri Devi, Ch. K. Rupesh Kumar, U. V. Ramesh, 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), pp.1-5, 2023.

[5] J. L. Moran, P. J. Solomon, A. R. Peisach and J. Martin, "New models for old questions: generalized linear models for cost prediction," Journal of Evaluation in Clinical Practice, vol. 13, no. 3, pp. 381-389, 2007.

[6] M. A. Morid, K. Kawamoto, T. Ault, J. Dorius and S. Abdelrahman, "Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation," AMIA Annual Symposium Proceedings, vol. 2017, pp. 1312, 2017.

[7] C. A. Powers, C. M. Meyer, M. C. Roebuck and B. Vaziri, "Predictive modeling of total healthcare costs using pharmacy claims data: A comparison of alternative econometric cost modeling techniques," Medical Care, vol. 43, pp. 1065-1072, 2005.

[8] G. Manogaran and D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent," International Journal of Advanced Intelligence Paradigms, vol. 10, no. 1–2, pp. 118–132, 2018.

[9] S. Kent, F. Fusco, A. Gray, S. A. Jebb, B. J. Cairns, and B. Mihaylova, "Body mass index and healthcare costs: a systematic literature review of individual participant data studies," Obesity Reviews, vol. 18, no. 8, pp. 869–879, 2017.

[10] M. Ravaut, H. Sadeghi, K. K. Leung et al., "Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data," NPJ Digital Medicine, vol. 4, no. 1, pp. 1–12, 2021.