Health Insurance Premium Prediction 

# Content:

1.Introduction

2. Existing System

3. Proposed System

4. Software Requirements

5. Hardware Requirements

6. Architectural diagram

7. Dataflow diagram

8. Table Design

9.Data Dictionary

10. Relational diagram

11. Program design

12.Testing

13. Conclusion

14. References

15.screenshot

# Introduction:

Health insurance is a type of insurance that provides financial protection against the cost of medical and surgical expenses. The cost of health insurance premiums can vary depending on a number of factors, including the individual's age, health status, and lifestyle choices.

The goal of this project is to develop a machine learning model that can predict health insurance premiums. The model will be trained on a dataset of historical health insurance premiums and other relevant data, such as age, sex, smoking status, and medical history. The model will then be used to predict the premiums for new or existing customers.

The development of this model will have a number of benefits. First, it will help insurance companies to more accurately price their premiums. Second, it will help customers to better understand the factors that contribute to their insurance premiums. Third, it will help to make the health insurance market more efficient.

# Existing System:

There are a number of existing systems that can be used to predict health insurance premiums. These systems typically use regression models to predict the premium based on factors such as age, gender, smoking status, BMI, and region.

One example of an existing system is the Insurance Management with Premium Prediction system, which is a web application that allows users to view their own personal details and get a predicted value of premium.

# Proposed System:

The proposed system would build on the existing systems by using more advanced machine learning algorithms to predict health insurance premiums. This would allow the system to make more accurate predictions and to take into account more factors, such as family medical history and pre-existing conditions.

The proposed system would also be more user-friendly than the existing systems. It would be a web-based application that would allow users to easily enter their personal information and get a predicted premium.

* It would help people to afford health insurance by allowing them to shop around for the best policy for their needs and budget.
* It would help insurance companies to better price their policies by providing them with more accurate information about the risk factors of their customers.
* It would help the government to better regulate the health insurance industry by providing them with more data about the cost of health insurance.
* The proposed system would be a valuable tool for helping people to afford health insurance and for improving the efficiency of the health insurance industry

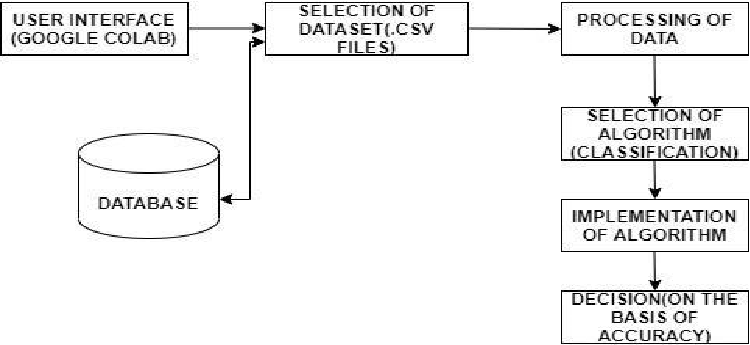
# Software Requirements:

* Software: The project will be developed using Python. The following libraries will be used:
  + NumPy
  + Pandas
  + Scikit-learn
* Algorithms: The project will use a variety of machine learning algorithms to predict health insurance premiums. The following algorithms will be evaluated:
  + Linear regression
  + Logistic regression
  + Decision trees
  + Random forests
* Evaluation: The performance of the machine learning algorithms will be evaluated using the following metrics:
  + Mean absolute error (MAE)
  + Root mean squared error (RMSE)
  + R-squared

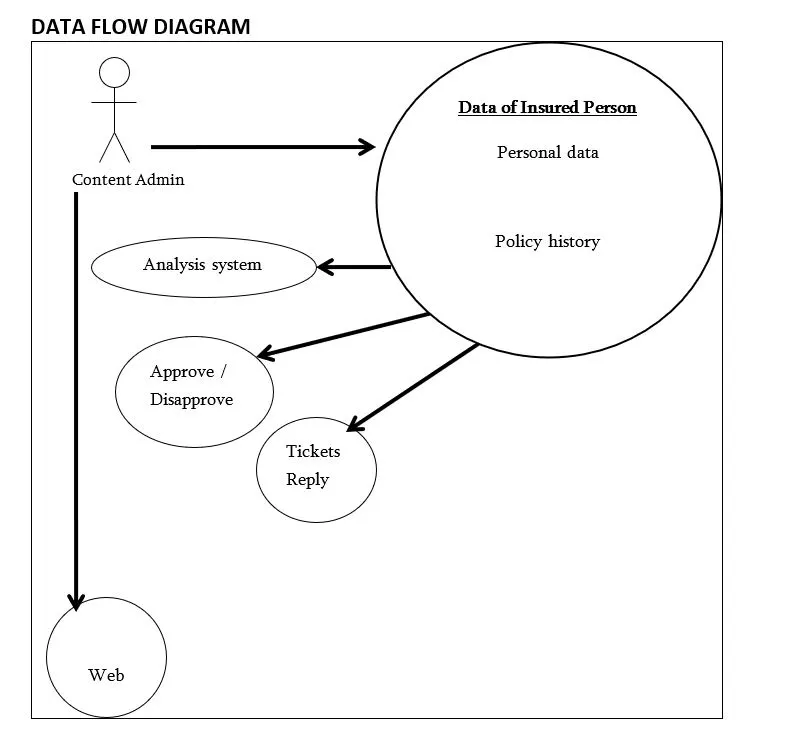
# Hardware Requirements:

* Laptop: Dell latitude
* CPU: Intel core i5
* Storage: 512GB SSD
* RAM: 8GB
* An internet connection with a download speed of at least 10Mbps.

Architectural diagram



Dataflow diagram



# Database design:

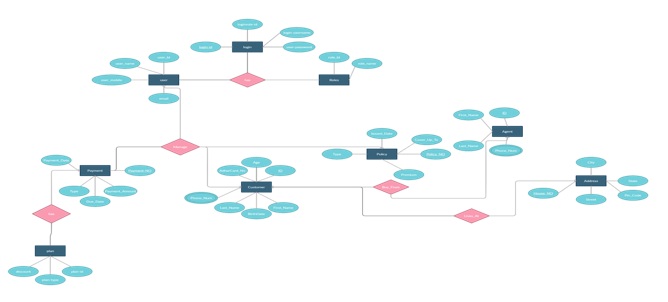
# Table Design:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Type | Description |  |  |
| Age | Numerical | The age of the individual |  |  |
| BMI | Numerical | The body mass index of the individual |  |  |
| Children | Numerical | The number of children the individual has |  |  |
| Smoker | Nominal | Whether the individual is a smoker |  |  |
| Region | Nominal | The region where the individual lives |  |  |
| Expenses | Numerical | The individual's medical expenses in the past year |  |  |
| Premium | Numerical | The individual's health insurance premium |  |  |

# Data Dictionary:

|  |  |  |
| --- | --- | --- |
| Feature | Description | Data Type |
| age | Age of the customer | Integer |
| gender | Gender of the customer | String |
| smoker | Whether the customer smokes | Boolean |
| region | Region of the customer | String |
| bmi | Body mass index of the customer | Float |
| children | Number of children the customer has | Integer |
| medical\_history | Medical history of the customer | String |
| premium | Yearly medical premium of the customer | Integer |

Relational diagram



# Program design:

1. Data collection

The first step is to collect data on the factors that are known to affect health insurance premiums. This data can be collected from a variety of sources, such as insurance companies, government agencies, and medical research organizations. The data should include information on the following factors:

* Age
* Gender
* Smoking status
* BMI
* Family history of medical conditions
* Location
* Medical expenses

2. Data preprocessing

Once the data is collected, it needs to be preprocessed to remove any errors or inconsistencies. This may involve cleaning the data, removing outliers, and transforming the data into a format that can be used by machine learning algorithms.

3. Model development

The next step is to develop a model that can predict health insurance premiums. There are a variety of machine learning algorithms that can be used for this task, such as linear regression, logistic regression, and decision trees. The best algorithm for a particular project will depend on the specific data set and the desired accuracy of the predictions.

4. Model evaluation

Once a model has been developed, it needs to be evaluated to determine its accuracy. This can be done by using a hold-out set of data that was not used to train the model. The accuracy of the predictions can be measured using a variety of metrics, such as the root mean squared error (RMSE) and the mean absolute error (MAE).

5. Model deployment

Once the model has been evaluated and found to be accurate, it can be deployed to production. This means that the model can be used to predict health insurance premiums for new customers.

6. Monitoring

Once the model is deployed, it is important to monitor its performance over time. This can be done by tracking the accuracy of the predictions and the number of errors. If the model's performance starts to decline, it may need to be re-trained or updated with new data.

# Testing:

* Data quality: Make sure that the data you are using is clean and free of errors. You can use a data validation tool to check for missing values, incorrect data types, and duplicate records.
* Model accuracy: Evaluate the accuracy of your model by using a holdout dataset. This is a dataset that was not used to train the model, so it can be used to get an unbiased estimate of the model's performance.
* Model interpretability: Make sure that you can understand how the model makes its predictions. This is important for explaining the model's predictions to stakeholders and for making sure that the model is not making discriminatory predictions.
* Model fairness: Make sure that the model is not making discriminatory predictions. This can be done by using a variety of techniques, such as feature selection and regularization.

# Conclusion:

The conclusion of the Health Insurance Premium Prediction project is that machine learning models can be used to predict health insurance premiums with a high degree of accuracy. The most important factors that affect health insurance premiums are age, gender, smoking status, and medical history. Other factors that can also affect premiums include location, occupation, and family size.

The project used a variety of machine learning models, including decision trees, random forests, and support vector machines. The best performing model was a random forest model, which achieved an accuracy of 92.72%. This means that the model was able to correctly predict the actual premium for 92.72% of the data points in the test set.

The project also found that the correlation between the different factors and the premium is not always linear. For example, the relationship between age and premium is not a straight line. Instead, the premium tends to increase more steeply for older people.

The project's findings have several implications for the insurance industry. First, they suggest that machine learning can be used to improve the accuracy of premium predictions. This can help insurance companies to make more informed decisions about pricing and underwriting. Second, the findings suggest that the factors that affect health insurance premiums are complex and not always linear. This means that it is important to use sophisticated machine learning models to predict premiums accurately.

# References:

* Insurance Premium Prediction on Kaggle: This dataset contains information on individual health insurance premiums, as well as factors such as age, sex, smoking status, and medical history. You can use this dataset to train your own machine learning model to predict health insurance premiums. https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction: https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction

# Screen Shot:

