

Estimation and Prediction of Hospitalization and Medical Care Costs

Naan Mudhalvan – Data Analytics Track

College: Madras Institute of Technology, Anna University (MIT, Anna University)
Chrompet

Group Members:

Team Lead: Balasubramaniam N (2020506020)

Jawahar A S (2020506035)

Sarukesh S (2020506084)

Vijay N (2020506108)

Introduction:

Background and Motivation:

The insurance industry heavily relies on accurate predictions of insurance charges to determine appropriate premiums for policyholders. Predicting insurance charges involves analyzing various factors such as age, BMI, and other relevant features. Accurate estimation of insurance charges not only benefits insurance companies but also helps individuals in making informed decisions regarding their insurance coverage.

Problem Statement:

In this project, the aim is to develop a predictive model for estimating insurance charges based on a given dataset. By leveraging data analysis techniques, machine learning algorithms, and web development tools, we aim to create a user-friendly application that provides accurate predictions of insurance charges. The project also focuses on creating an interactive dashboard using IBM Cognos to present the analysis, findings, and model performance.

By accurately predicting insurance charges, insurance companies can optimize their pricing strategies, enhance risk assessment, and improve overall profitability. Moreover, individuals seeking insurance can benefit from a transparent and reliable estimation of charges, aiding them in selecting the most suitable insurance plans for their needs.

Through this project, we aim to showcase the process of data exploration, feature analysis, model development, and deployment using modern software tools and technologies. The results obtained will contribute to the body of knowledge in insurance prediction and demonstrate the potential of data-driven approaches in the insurance industry.

The following sections of this report will provide a detailed overview of the dataset used, the software tools and technologies employed, the methodology followed for data analysis and modeling, the results obtained, and the conclusions drawn from the project. Additionally, we will discuss the development of a Flask app for prediction and an interactive dashboard using IBM Cognos to present the project's findings effectively.

Aim:

The aim of this project is to develop a predictive model to estimate insurance charges based on various features such as age, BMI, and other relevant factors. Additionally, the project involves the creation of a Flask app for deploying the model and a dashboard using IBM Cognos to present the analysis and predictions.

Software Used:

The following software tools were utilized in this project:

- Python: Used for data analysis, modeling, and Flask app development.

- Jupyter Notebook: Used for data exploration, descriptive, bivariate, univariate, and multivariate analysis.
- Scikit-learn: Employed for implementing the linear regression model.
- Flask: Utilized for creating the web application.
- IBM Cognos: Utilized for developing the dashboard, report, and storytelling.

Tech Stack:

The tech stack employed in this project includes:

- Python: Programming language for data analysis and modeling.
- Pandas: Data manipulation and analysis library.
- NumPy: Numerical computing library.
- Matplotlib and Seaborn: Data visualization libraries.
- Scikit-learn: Machine learning library for implementing linear regression.
- Flask: Python web framework for creating the web application.
- IBM Cognos: Business intelligence and analytics platform for creating the dashboard, report, and story.
- MongoDB: Database

Procedure:

a) Data Exploration and Analysis:

- Descriptive Analysis: Examined the basic statistics, distributions, and summary of the dataset, including features like age, BMI, and charges.
- Bivariate Analysis: Explored relationships between pairs of variables to identify any correlations or patterns.
- Univariate Analysis: Analyzed each feature individually to gain insights into their distributions and characteristics.
- Multivariate Analysis: Investigated the interactions between multiple variables to uncover complex relationships.

b) Linear Regression:

- Utilized the linear regression algorithm from the Scikit-learn library.
- Split the dataset into training and testing sets.
- Preprocessed the data by handling missing values, encoding categorical variables, and scaling the features if necessary.
- Trained the linear regression model using the training data.
- Evaluated the model's performance using appropriate metrics, such as mean squared error or R-squared.

c) Flask App Development:

- Created a Flask web application.
- Loaded the trained linear regression model into the Flask app.
- Implemented an interface for users to input relevant features, such as age, BMI, etc.
- Utilized the loaded model to predict insurance charges based on the user's input.

- Displayed the predicted charges on the web interface.

d) Dashboard, Report, and Story with IBM Cognos:

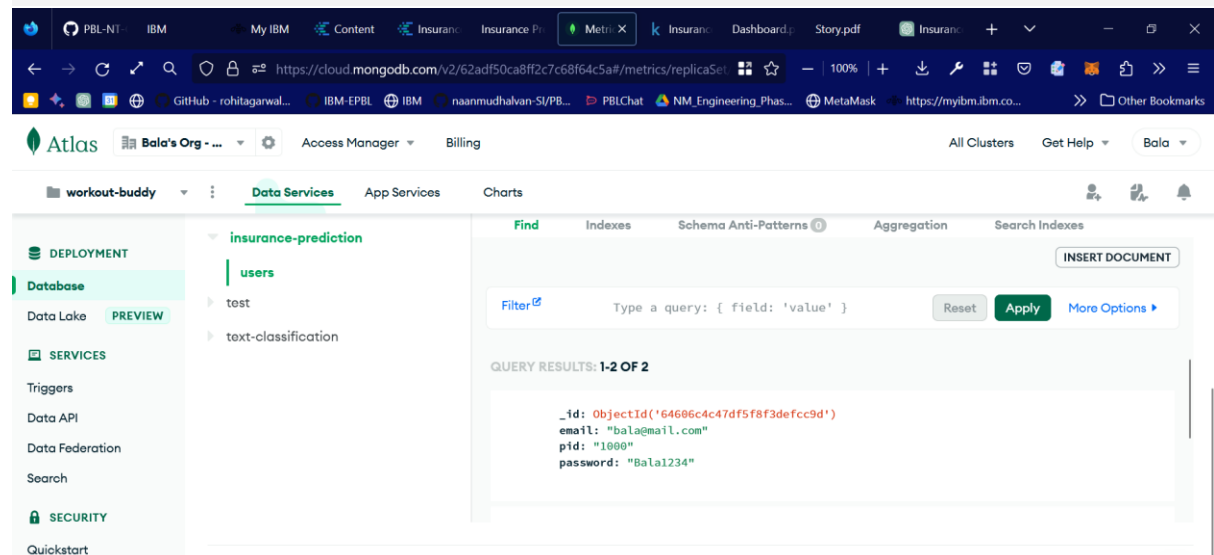
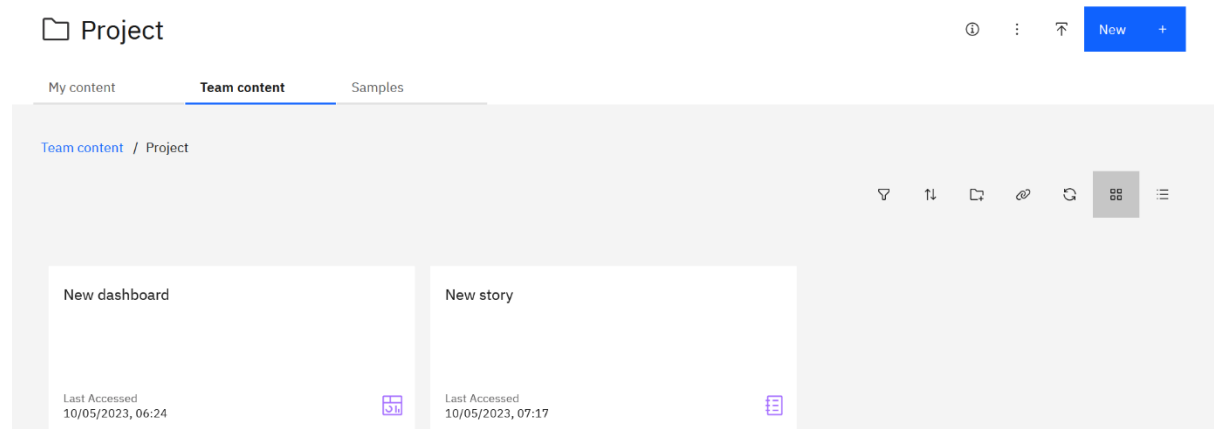
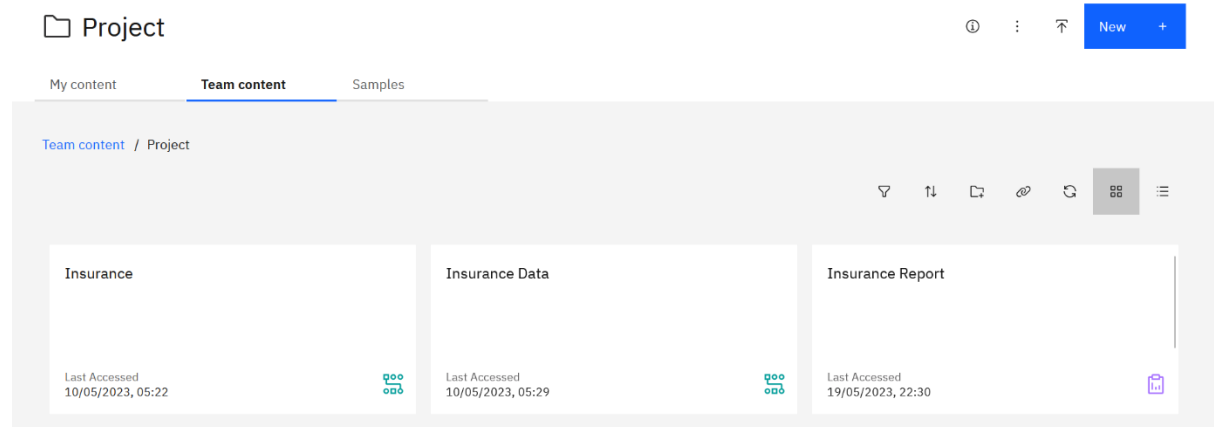
- Leveraged IBM Cognos to develop a dashboard.
- Integrated visualizations and key insights obtained from the data analysis phase.
- Created a report summarizing the project's methodology, findings, and model performance.
- Developed a story that presents the project's progression, challenges, and results using engaging visualizations and narratives.

Results:

- Descriptive Analysis: Provided a comprehensive overview of the dataset, including summary statistics, distributions, and key features.
- Bivariate Analysis: Identified correlations between variables and discovered any notable relationships.
- Univariate Analysis: Gained insights into individual feature distributions and characteristics.
- Multivariate Analysis: Uncovered complex relationships and interactions between multiple variables.
- Linear Regression: Developed a predictive model to estimate insurance charges, evaluated its performance, and identified key factors influencing charges.
- Flask App: Created a user-friendly web application that predicts insurance charges based on user input.
- IBM Cognos Dashboard, Report, and Story: Developed an interactive and visually appealing dashboard, a comprehensive report summarizing the project, and an engaging storytelling presentation.

Next we see the output and the conclusion in the next page

Output:





Search



Insurance Premium Prediction

Data Card Code (92) Discussion (0)

103

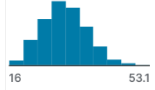
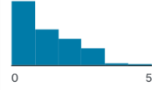
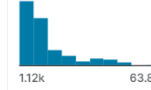
New Notebook

Download (14 kB)

About this file

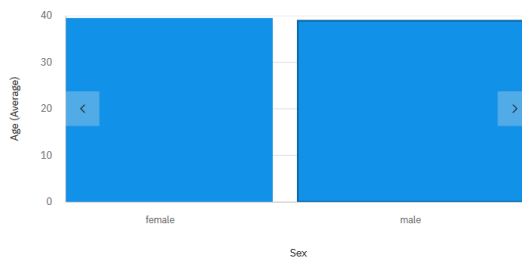
The insurance.csv dataset contains 1338 observations (rows) and 7 features (columns). The dataset contains 4 numerical features (age, bmi, children and expenses) and 3 nominal features (sex, smoker and region) that were converted into factors with numerical value designated for each level.

The purposes of this exercise to look into different features to observe their relationship, and plot a multiple linear regression based on several features of individual such as age, physical/family condition and location against their existing medical

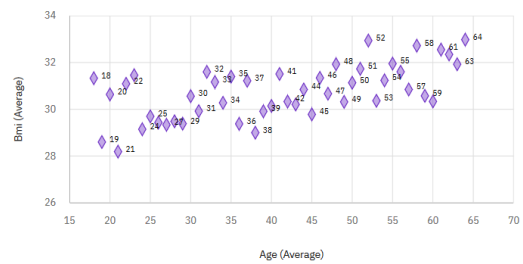
# bmi	# children	✓ smoker	Δ region	# expenses										
		<table><tr><td>true</td><td>0%</td></tr><tr><td>false</td><td>0%</td></tr></table>	true	0%	false	0%	<table><tr><td>southeast</td><td>27%</td></tr><tr><td>southwest</td><td>24%</td></tr><tr><td>Other (649)</td><td>49%</td></tr></table>	southeast	27%	southwest	24%	Other (649)	49%	
true	0%													
false	0%													
southeast	27%													
southwest	24%													
Other (649)	49%													
27.9	0	yes	southwest	16884.92										
33.8	1	no	southeast	1725.55										

Tab 1

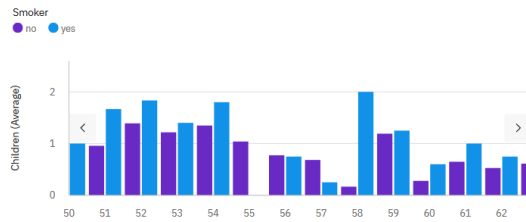
Average Age by Sex



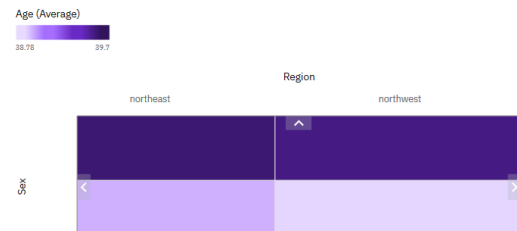
Age by Bmi with points for Age



Children by Age colored by Smoker



Age by Sex and Region

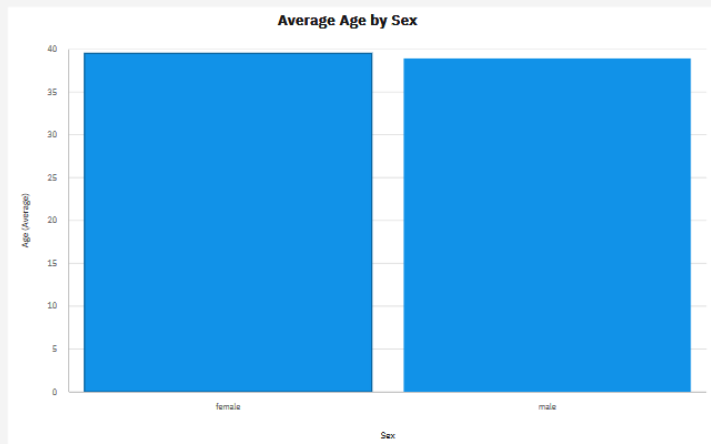


Average Age by Gender

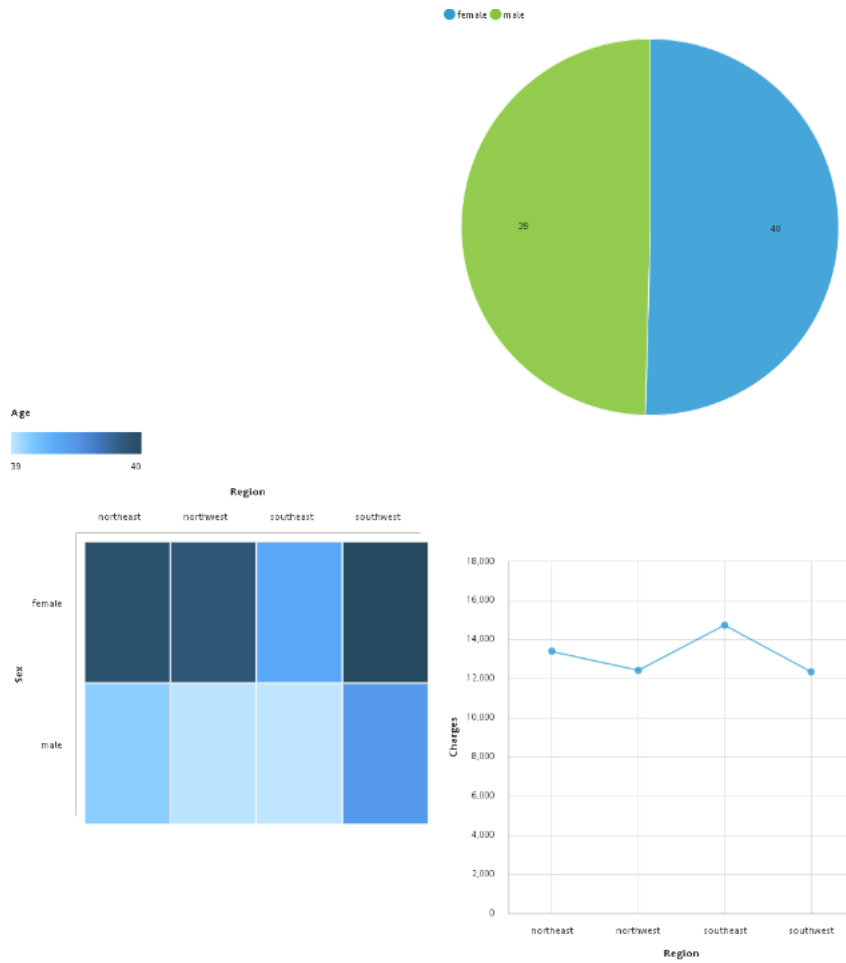
This graph gives us the average age by sex

The average age of women is higher than men.

It is also worthy to note that, the no of women in the given dataset is less.



```
bala9@XPS13 MSYS /c/Storage/College/SEM_6/NM PREED A/project/Final Deliverables/web_app (main)
$ python app.py
* Serving Flask app 'app' (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: on
* Running on http://127.0.0.1:5000 (Press CTRL+C to quit)
* Restarting with watchdog (windowsapi)
* Debugger is active!
* Debugger PIN: 216-957-582
127.0.0.1 - - [20/May/2023 08:44:58] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [20/May/2023 08:44:58] "GET /favicon.ico HTTP/1.1" 404 -
127.0.0.1 - - [20/May/2023 08:46:43] "POST /login HTTP/1.1" 302 -
127.0.0.1 - - [20/May/2023 08:46:43] "GET /home HTTP/1.1" 200 -
```



LOGIN TO ACCESS THE HEALTH INSURANCE PREDICTION

Login

UID:

Password:

LOGIN

No account? [Sign up](#)

Sign Up

Email ID:

UID:

Password:

Re-enter your password:

Sign up

Existing User? [Login](#)

Prediction Page

Insurance Premium Prediction

Age:

Sex:

BMI:

Number of children:

Smoker:

Number of children:

Smoker:

Region:

Submit

[Analysis](#)

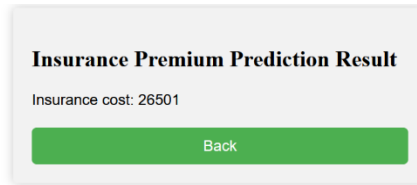
Predicted Value:

Insurance Premium Prediction Result

Insurance cost: 2280

Back

The prediction if for the same data smoker is changed to yes

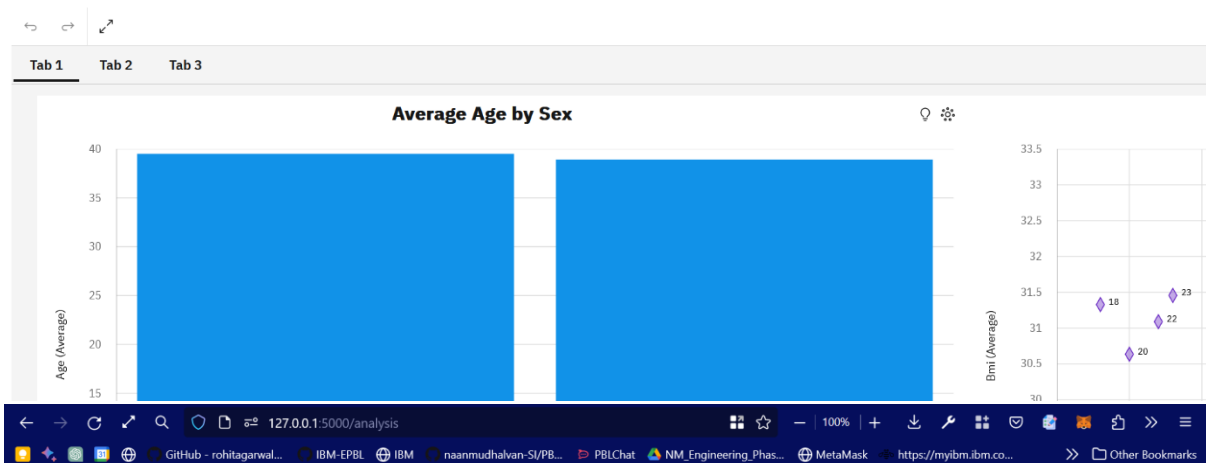


The Dashboard page of the flask Application:



IBM Cognos Analysis

Dashboard



Story



Conclusion:

In conclusion, this project aimed to predict insurance charges using a dataset containing features such as age, BMI, and other relevant factors. The analysis involved descriptive, bivariate, univariate, and multivariate techniques to gain insights into the dataset. A linear regression model was implemented to estimate charges accurately. The model was deployed using a Flask app, enabling users to obtain predicted charges based on their input.

Additionally, an IBM Cognos dashboard, report, and story were created to present the analysis, model performance, and overall project findings. This project provides a comprehensive solution for insurance charge prediction and offers valuable insights for the insurance industry.