Architecture

Insurance Premium Prediction

Prepared By: Mangesh Devkate

Data Science Intern

iNeuron.ai

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1 | 1-12-2022 | Mangesh Devkate |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[Document Version Control 2](#_Toc120366836)

[Abstract 4](#_Toc120366837)

[1. Introduction 5](#_Toc120366838)

[1.1 What is Architecture? 5](#_Toc120366839)

[1.2 Scope 5](#_Toc120366840)

[1.3 Constraints 5](#_Toc120366841)

[2. Technical Specifications 5](#_Toc120366842)

[2.1 Dataset 5](#_Toc120366843)

[2.2 Logging 6](#_Toc120366844)

[2.3 Deployment 6](#_Toc120366845)

[3. Technology Stack 7](#_Toc120366846)

[4. Proposed Solution 7](#_Toc120366847)

[5. Architecture 7](#_Toc120366848)

[5.1 Data Gathering 8](#_Toc120366849)

[5.2 Raw Data Validation 8](#_Toc120366850)

[5.3 Exploratory Data Analysis 8](#_Toc120366851)

[5.4 Feature Engineering 8](#_Toc120366852)

[5.5 Feature Selection 8](#_Toc120366853)

[5.6 Train-Test-Split 8](#_Toc120366854)

[5.7 Model Building 8](#_Toc120366855)

[5.8 Model Evaluation 8](#_Toc120366856)

[5.8 Model Saving 8](#_Toc120366857)

[5.9 Flask Setup for Web Application 9](#_Toc120366858)

[5.10 GitHub 9](#_Toc120366859)

[5.11 Deployment 9](#_Toc120366860)

[6. User Input/Output Workflow 9](#_Toc120366861)

# Abstract

Machine learning is a part of Artificial Intelligence and it is a process in which machine is having ability to learn without being explicitly programmed. Machine learning algorithms build a model based on sample data, known as [training data](https://en.wikipedia.org/wiki/Training_data), in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, [email filtering](https://en.wikipedia.org/wiki/Email_filtering), [speech recognition](https://en.wikipedia.org/wiki/Speech_recognition), [agriculture](https://en.wikipedia.org/wiki/Agriculture), health, insurance and [computer vision](https://en.wikipedia.org/wiki/Computer_vision), where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. In this project, we will estimate the amount of insurance premium on the basis of personal health information. Taking various aspects of a dataset collected from people, and the methodology followed for building a predictive model.

# 1. Introduction

## 1.1 What is Architecture?

The machine learning architecture defines the various layers involved in the machine learning cycle and involves the major steps being carried out in the transformation of raw data into training data sets capable for enabling the decision making of a system.

## 1.2 Scope

Architecture is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software, architecture, source code, and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work. And the complete workflow.

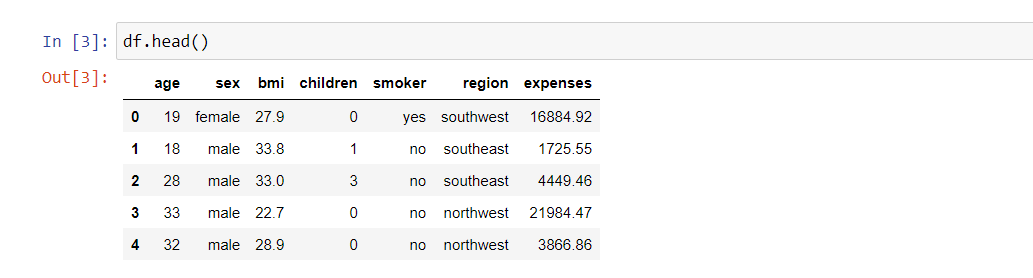
## 1.3 Constraints

We only predict the expected estimating cost of insurance premium of customers based on their personal health information.

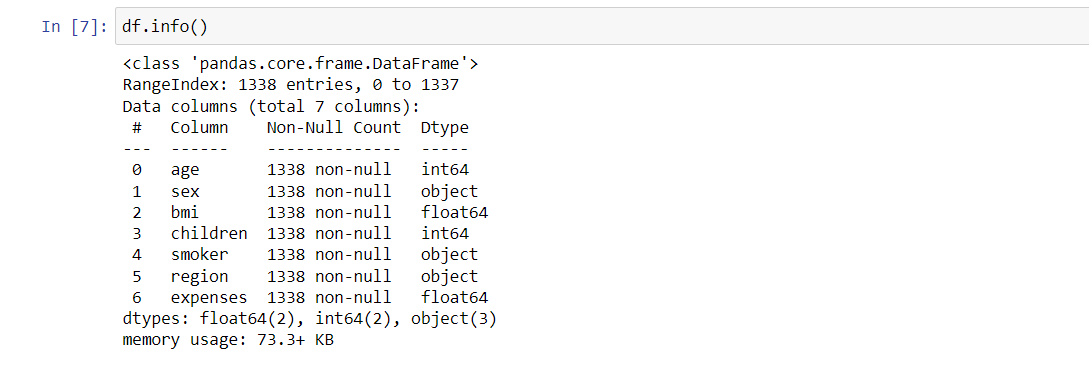
# 2. Technical Specifications

## 2.1 Dataset

The dataset containing historical data of 1338 customers personnel information and their actual medical expenses. It consists 1338 rows and 7 columns. The objective is to estimate the customers expenses based on their personnel information like age, sex, bmi, children, smoker, region etc. The dataset looks like as follow



Data contains numerical as well as categorical columns. The columns in the dataset consists of various data types like int, float, object as shown in fig.



The categorical variables are sex, smoker, region and the numerical variables are age, bmi, children and expenses. Below is the summary of the numerical variables which includes statistical information of the variables like, mean, std, min, max, percentile value of the numerical variables. as shown in below fig.



Exploratory Data Analysis is doing analysis of dataset. Doing univariate analysis and bivariate analysis is the best way to analyse and understand the dataset. Univariate analysis can be done by using countplot, distplot, histogram, boxplot and bivariate analysis can be done by using scatterplot, pairplot. Feature Engineering includes removing duplicate rows in the dataset, filling NULL/missing values in the variables by imputing the variables, checking data distribution of variables by which we get to know that whether data distribution is normal or schewed, outlier detection and handling outliers, transformation of categorical variables into numerical variables by encoding techniques.

## 2.2 Logging

We should be able to log every activity done by the user

* The system identifies at which step logging require.
* The system should be able to log each and every system flow.
* The system should be not be hung even after using so much logging. Logging just because we can easily debug issuing so logging is mandatory to do.

## 2.3 Deployment

We have used Heroku for project deployment.



## 3. Technology Stack

|  |  |
| --- | --- |
| Front End | HTML/CSS |
| Back End | Python/Flask |
| Deployment | Heroku |

## 4. Proposed Solution

We have used EDA to find the important relation between different attributes and machine-learning algorithms to estimate the cost of expenses. The user will provide the required feature values as input and will get results through the web application. The system will get features values and it will be passed into the backend where the features will be validated and pre-processed and then it will be passed to a hyperparameter tuned machine learning model to predict the final outcome.

## 5. Architecture

Data Gathering

Feature Engineering

Start

Importing Necessary Libraries

Prediction

Pushing App On The Cloud

Cloud Setup

Stop

Push All Files Into GitHub

Flask Setup

Feature Selection

Data From User (Input)

Model Evaluation

Application Start

Save Model

Train-Test-Split

Model Building

## 5.1 Data Gathering

Data Source : <https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction>

## 5.2 Raw Data Validation

After data is loaded, various types of validation are required before we proceed further with any operation. Validations like checking for zero standard deviation for all the columns, checking for complete missing values in any columns, etc. These are required because the attributes which contain these are of no use. It will not play role in contributing to the estimating cost of the premium.

## 5.3 Exploratory Data Analysis

In EDA we have done analysis of dependent and independent variables. We have done univariate analysis and bivariate analysis to get insights of the data.

## 5.4 Feature Engineering

Feature Engineering consists various steps like removing duplicate rows in the dataset, filling NULL/Missing values, outlier detection and handling outliers, transformation of categorical columns into numerical columns using encoding techniques like label encoding, one hot encoding. Adding to these feature engineering consists other steps like feature scaling, feature binning.

## 5.5 Feature Selection

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. Feature selection techniques are wrapper method, filter method, embedded method.

## 5.6 Train-Test-Split

The train\_test\_split() method is used to split our data into train and test sets. First, we need to divide our data into features (X) and labels (y). The dataframe gets divided into X\_train, X\_test , y\_train and y\_test. X\_train and y\_train sets are used for training and fitting the model.

## 5.7 Model Building

Building an ML Model requires splitting of data into two sets, such as ‘training set’ and ‘testing set’ in the ratio of 80:20 or 70:30. A set of supervised (for labelled data) and unsupervised (for unlabeled data) algorithms are available to choose from depending on the nature of input data and business outcome to predict.

## 5.8 Model Evaluation

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance.

## 5.8 Model Saving

Model is saved using pickle library in pickle` format.

## 5.9 Flask Setup for Web Application

After saving the model, the API building process started using Flask. Web application creation was created in Flask for testing purpose. Data entered by user extracted by the model to estimate the premium of insurance.

## 5.10 GitHub

The whole project directory pushed into the GitHub repository.

## 5.11 Deployment

The project was deployed from GitHub into the Heroku platform.



# 6. User Input/Output Workflow

Pre-Processing

Predicted Result

Start Application

Showing Error Page

Submit Detail

User Input