INSURANCE PREMIUM PREDICTION

Architecture Design

I. Document Version Control

Date Issued	Version	Description	Author
19.02.2023	V1.0	Architecture Design	Shikha

A. Abstract

Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. 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

❖ What is Architecture Design?

Architecture Design (AD) aims to give the internal design of the actual program code for the `Insurance Premium Prediction`. AD describes the class diagrams with the methods and relation between classes and program specifications. It describes the modules so that the programmer can directly code the program from the document.

❖ Scope

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

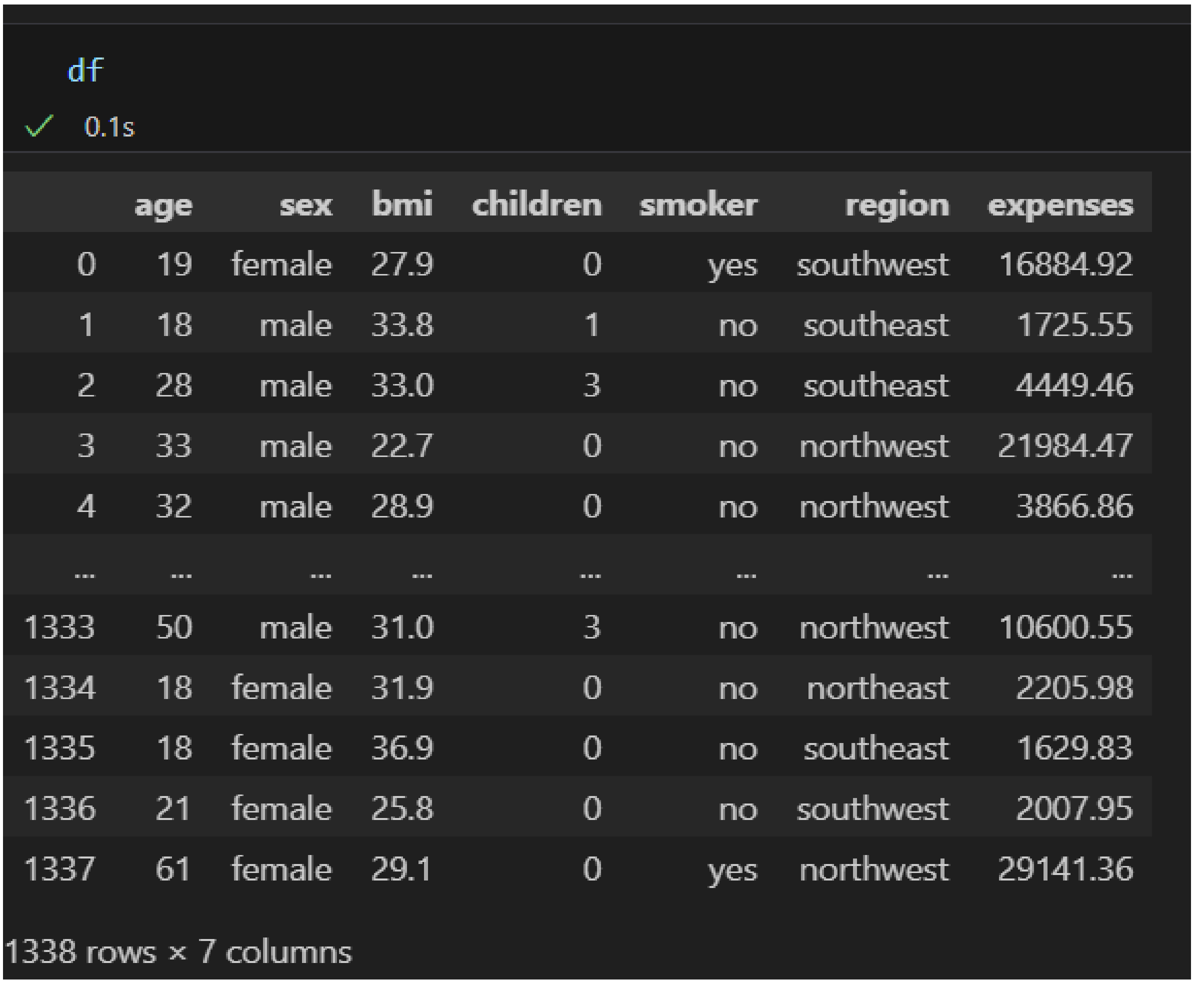
❖ Constraints

We only predict the expected estimated cost of expenses for customers based on some personal health information.

2. Technical Specification

❖ Dataset

The dataset contains validated historical data, encompassing the aforementioned details and the real medical expenses paid by more than 1338 customers. The goal is to devise a method for estimating the values in the "expenses" column based on the information provided in other columns such as age, gender, BMI, number of children, smoking habits, and region. By examining all the observations, it is possible to infer the impact of certain user characteristics on their expenses. The dataset is structured in the following format:

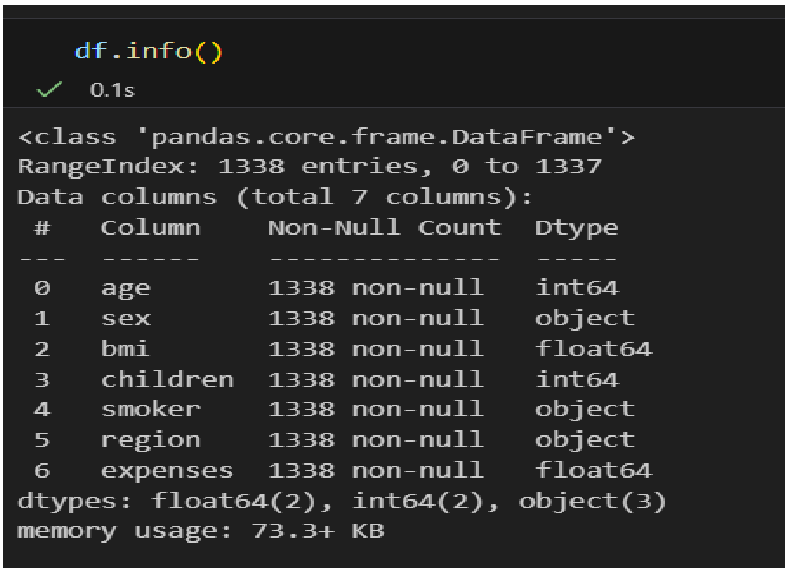


```
df
✓ 0.1s
```

	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86
...	...	...	...	...	...	...	...
1333	50	male	31.0	3	no	northwest	10600.55
1334	18	female	31.9	0	no	northeast	2205.98
1335	18	female	36.9	0	no	southeast	1629.83
1336	21	female	25.8	0	no	southwest	2007.95
1337	61	female	29.1	0	yes	northwest	29141.36

1338 rows × 7 columns

The data set consists of various data types from integer to floating to object as shown in Fig.



```
df.info()
✓ 0.1s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   expenses  1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

The dataset comprises different types of underlying patterns that offer a comprehensive understanding of the subject matter and provide valuable insights into the problem at hand. It appears that 'age,' 'children,' 'BMI' (body mass index), and 'expenses' are represented as numerical values, while 'sex,' 'smoker,' and 'region' are represented as strings, likely representing categories.

For the numerical attributes, various statistical measures such as mean, standard deviation, median, count of values, and maximum value are presented below to highlight their importance:

Mean: It represents the average value of the numerical attribute across the dataset.

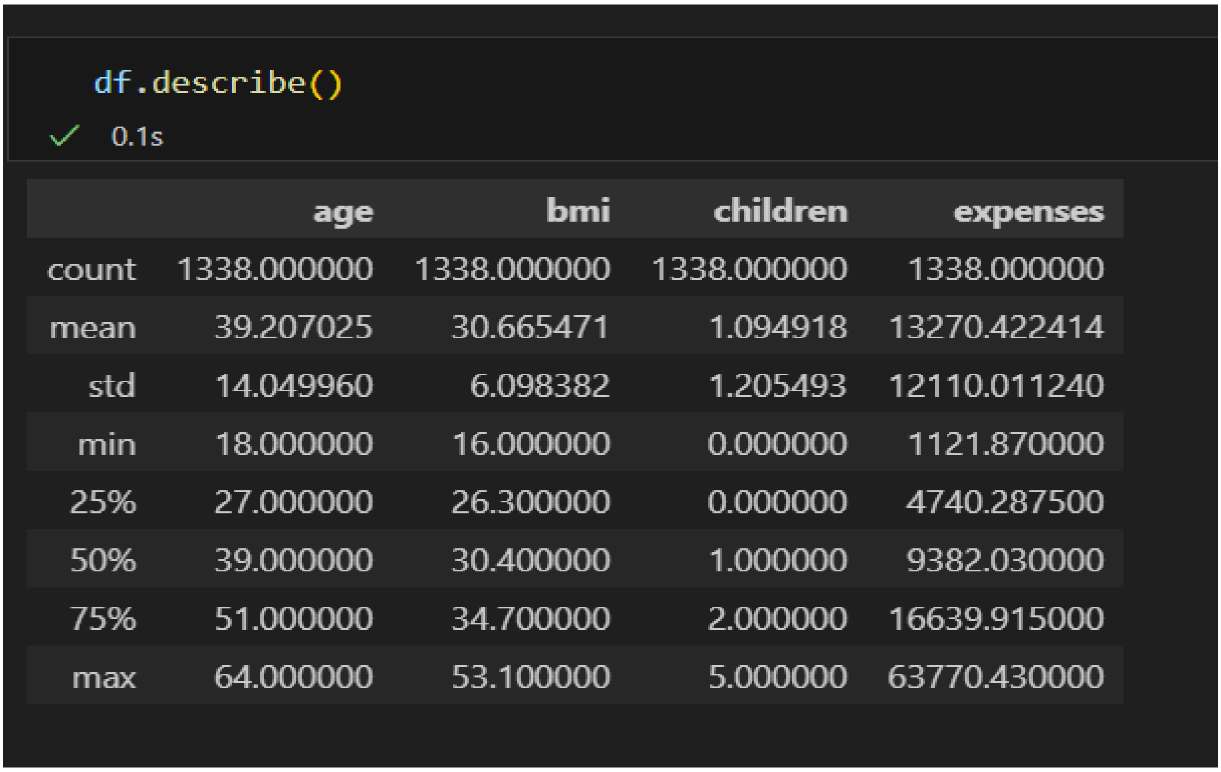
Standard deviation: It indicates the dispersion or variability of the numerical attribute values around the mean.

Median: It represents the middle value in a sorted list of the numerical attribute, dividing the data into two equal halves.

Count of values: It denotes the number of instances or observations present for the numerical attribute.

Maximum value: It represents the highest numerical value observed for the attribute in the dataset.

These statistical measures provide valuable information about the distribution and characteristics of the numerical attributes in the dataset



```
df.describe()
```

✓ 0.1s

```
               age          bmi     children      expenses
count  1338.000000  1338.000000  1338.000000   1338.000000
 mean    39.207025    30.665471     1.094918  13270.422414
  std    14.049960     6.098382     1.205493  12110.011240
  min    18.000000    16.000000     0.000000   1121.870000
  25%    27.000000    26.300000     0.000000   4740.287500
  50%    39.000000    30.400000     1.000000   9382.030000
  75%    51.000000    34.700000     2.000000  16639.915000
  max    64.000000    53.100000     5.000000  63770.430000
```

The pre-processing of this dataset involves conducting an analysis on the independent variables. This analysis includes checking for any null values present in each column and subsequently replacing or filling them with appropriate data types. This ensures that the accuracy of the analysis and model fitting process is not impeded.

The representations displayed above are generated using tools provided by Pandas. These representations provide information such as the count of variables for numerical columns and the distinct values for categorical columns.

The maximum and minimum values observed in the numerical columns, as well as their percentile values such as the median, play a crucial role in determining which values should be prioritized for further exploration and analysis tasks.

The data types assigned to different columns are subsequently utilized in label processing and a one-hot encoding scheme during the construction of the model. This enables the effective handling of categorical variables in the model-building process.

- **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 not be hung even after using so much logging. Logging is just because we can easily debug issuing so logging is mandatory to do.

- **Deployment**

  For hosting we will use Heroku

3. Technology Stack

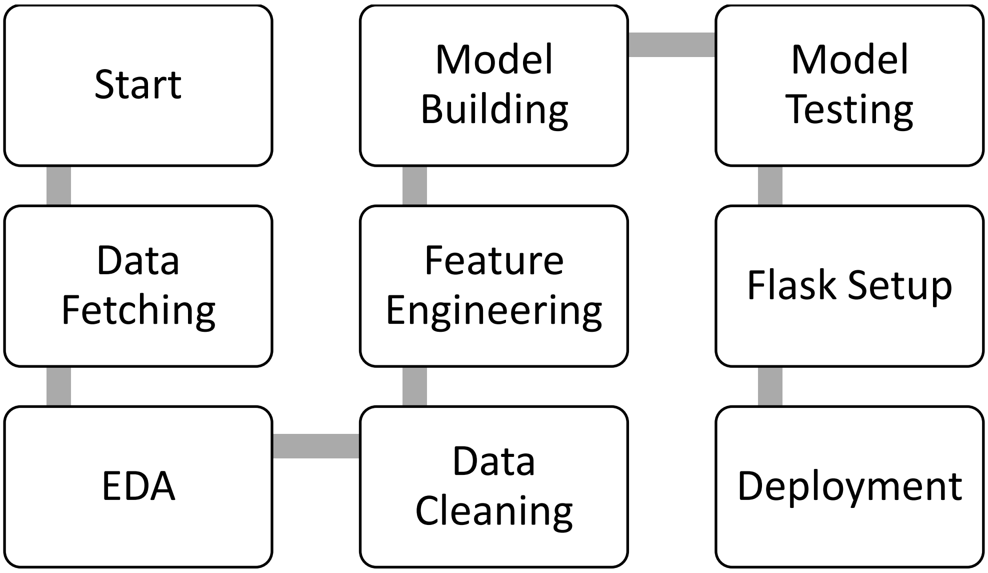
Front End	HTML/CSS
Backend	Python/Flask
Deployment	Heroku

4. Proposed Solution

To gain valuable insights and establish meaningful relationships among various attributes, we will conduct Exploratory Data Analysis (EDA). Subsequently, a machine-learning algorithm will be employed to estimate the cost of expenses. The client will provide the necessary input features through a web application and receive the results accordingly.

The system will receive the input features and pass them to the backend. Here, the features will undergo validation and pre-processing steps to ensure their quality and compatibility with the model. Once the pre-processing is complete, the features will be passed to a machine-learning model that has been fine-tuned with hyperparameters. This model will predict the final outcome, which represents the estimated cost of expenses based on the input provided by the client.

5. Architecture



❖ Data Gathering

Data source: https://www.kaggle.com/noordeen/insurance-premium-prediction

The dataset is stored in .csv format.

❖ Raw Data Validation

B. Once the data is loaded, it is crucial to perform several validations before proceeding with any further operations. These validations include checking for zero standard deviation across all columns and identifying any columns with complete missing values. Such validations are necessary because attributes that exhibit zero standard deviation or contain completely missing values do not contribute to the estimation of premium cost and are therefore irrelevant.

C. By conducting these validations, we can identify and exclude attributes that lack variability or meaningful information, ensuring that they do not impact the accuracy of the premium cost estimation.

❖ Exploratory Data Analysis

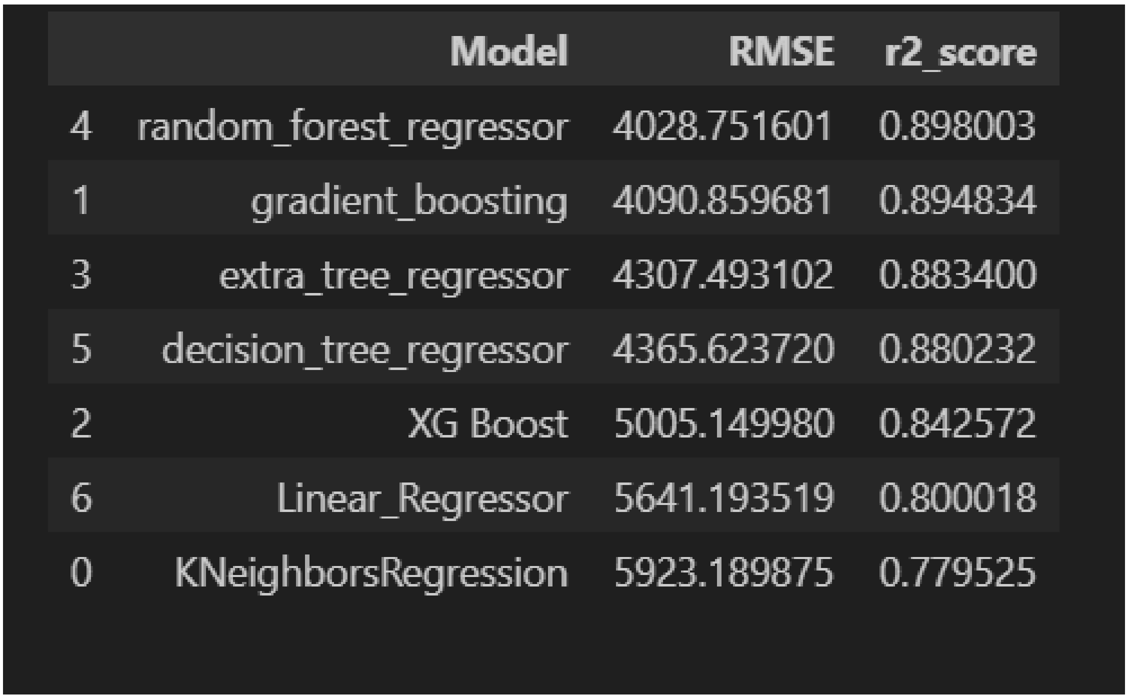
D. Visualized the relationship between the dependent and independent features. Also checked the relationship between independent features to get more insights **about the data.**

❖ Feature Engineering

Following the pre-processing step, a standard scalar operation is applied to scale down all the numeric features. Additionally, one-hot encoding is performed to convert categorical features into numerical representations. To streamline this process, a pipeline is created to efficiently execute the scaling of numeric features and the encoding of categorical features.

❖ Model Building

After doing all kinds of pre-processing operations mentioned above and performing scaling and encoding, the data set is passed through a pipeline to all the models. The following are the results:



	Model	RMSE	r2_score
4	random_forest_regressor	4028.751601	0.898003
1	gradient_boosting	4090.859681	0.894834
3	extra_tree_regressor	4307.493102	0.883400
5	decision_tree_regressor	4365.623720	0.880232
2	XG Boost	5005.149980	0.842572
6	Linear_Regressor	5641.193519	0.800018
0	KNeighborsRegression	5923.189875	0.779525

❖ Model Saving

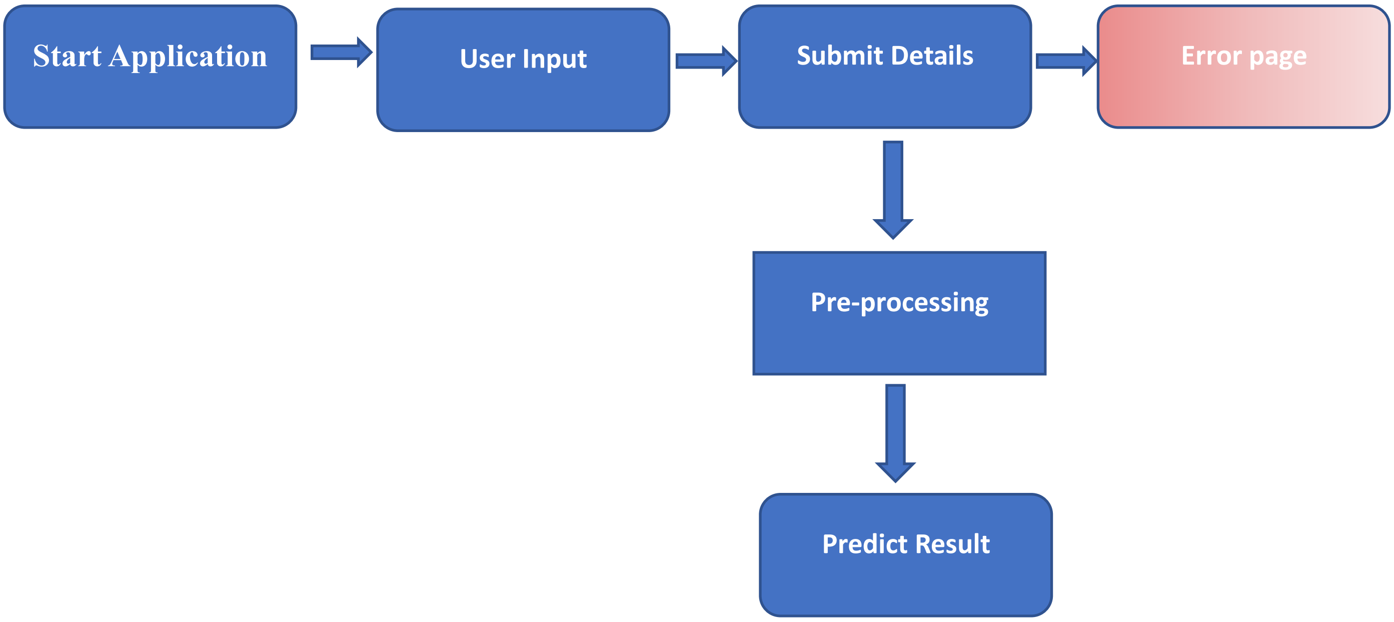
The model is saved using the pickle library in pickle format.

❖ 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. Whatever user will enter the data and then that data will be extracted by the model to estimate the premium of insurance, this is performed in this stage.

❖ **GITHUB**
The whole project directory will be pushed into the GitHub repository.

❖ Deployment
The project was deployed from GitHub into the Heroku platform.


Start Application
User Input
Submit Details
Error page
Pre-processing
Predict Result