ARCHITECTURE DESIGN

INSURANCE PREMIUM PREDICTION

Barnalikka Pradhan, Punith B C
iNeuron

Document Version Control

Date Issued	Version	Description	Author
29.09.2022	V1.0	Initial Architecture Design-V1.0	Barnalikka Pradhan, Punith B C

Contents

Abstract

Insurance premiums have become a necessity in today's time owing to the rise in people being aware about their individual health situation and wanting to buy an insurance premium. In this project, machine learning algorithms play a vital role in using data to be able to predict the cost of the insurance premiums which make it easier for the individual to come up with a plan while purchasing one.

1. Introduction

1.1 What is Architecture Design?

This document gives a vivid description of the structural steps that have been followed while coding, for Insurance Premium Prediction. Each module has been described clearly so that one can follow the steps of coding and do it accordingly from this document.

1.2 Scope

The scope of this web application is that it gives an estimate of the cost which the individual will have to bear for purchasing the insurance premiums. This further gives them the clarity with which they can choose their insurance plans, learn about the scope of benefits included in it, the plans cost sharing requirements and the extent of coverage.

1.3 Constraints

Very few physiological features around 6 attributes have been used to predict the insurance premiums of the individuals.

2. Technical Specification

2.1 Dataset Overview

The dataset contains data of 1338 individuals. The main objective is to estimate expenses which is our dependent variable using attributes like age, sex, bmi, children, smoker, region which form our independent variables.

```
In [5]:  1 cleaned = pd.read_csv("/Users/barnalikkapradhan/Downloads/INTERNSHIP/insurance.csv")
         2 cleaned
```

Out[5]:

	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

As can be seen there are various datatypes ranging from 3 categorical variables which are named as objects and 4 numerical variables of which 2 are integer and 2 are float.

```
In [3]:  1 cleaned.info()
```

```
<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: 57.6+ KB
```

The important statistics like mean, standard deviation, median, count of values and maximum value, etc. are shown below for numerical attributes.

```
In [5]:  1 cleaned.describe()
```

Out[5]:

	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

2.2 Predicting Disease

The system displays input boxes to be filled of physiological features like age, sex, bmi, children, smoker, region.
Once the user has input their information and clicked on submit button the system should be able to display the predicted cost of the person's individual premium.

2.3 Logging

Logging is done for every activity performed by the user.

- The System identifies at what step logging required.
- The System should be able to log each and every system flow.
- System should not be hung even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

2.4 Database

System needs to store every request in the database in such a way that it becomes easy to retrain the model as well.
Here, Cassandra database has been used, to store each and every data, given by the user or received on request to the database.

2.5 Deployment

The entire code has first been pushed into GitHub Repository, which has then been deployed from GitHub into the AWS cloud plateform.

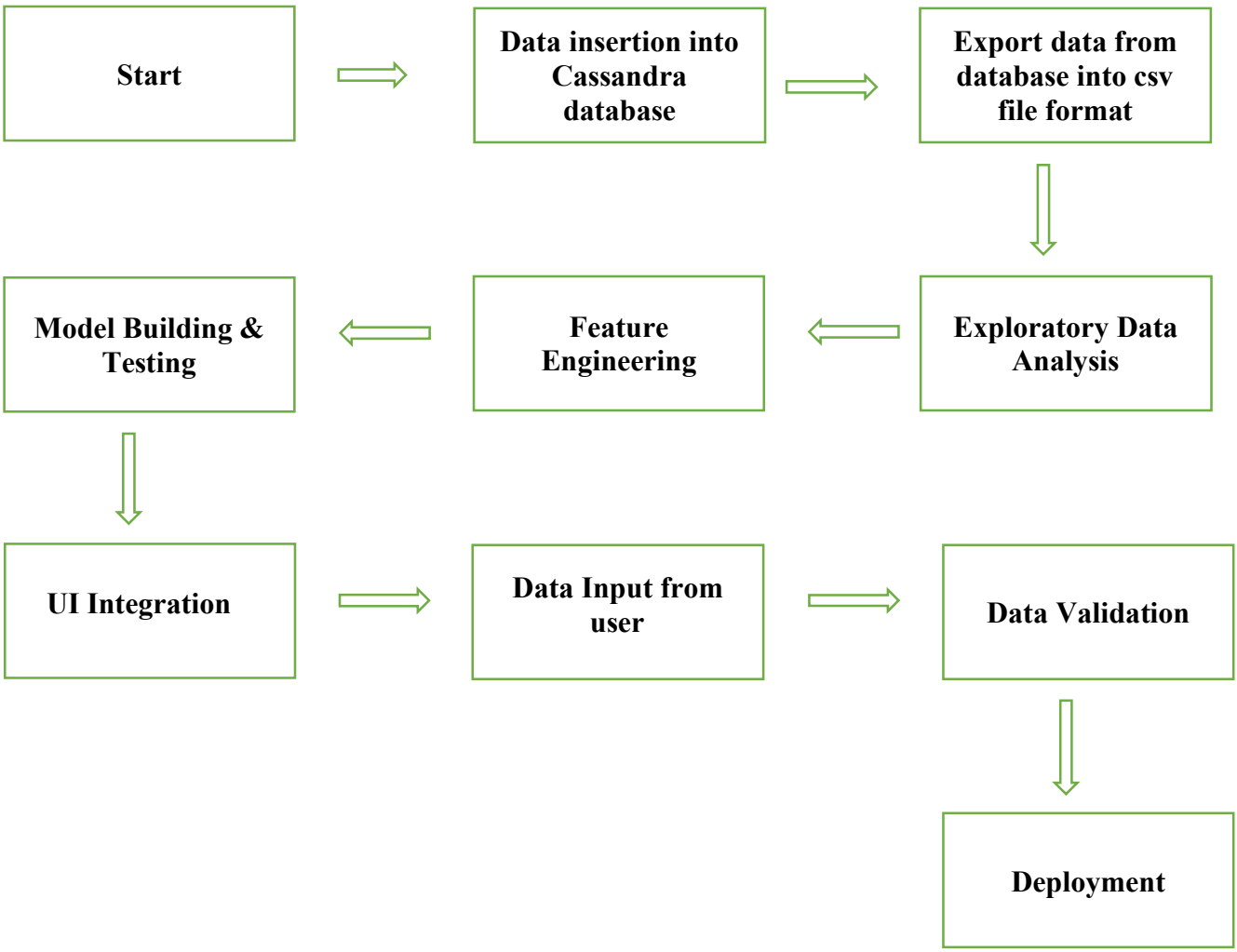
3. Technology stack

Front end development	HTML/CSS
Backend development	Python/Flask
Database	Cassandra
Deployment	AWS

4. Proposed Solution

Firstly, an Exploratory data analysis, helps show the relation between the attributes to estimate the cost of expenses. The machine learning algorithm choosen with the highest accuracy will be used for predicting the cost of insurance premium once the user inputs their individual details in the input boxes and get immediate results through the web application.

5. Architecture



5.1 Data Description

Dataset source: https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction
This dataset is stores in .csv file format.

5.2 Data Insertion into Database

Cassandra Database which is an open source database has the ability to handle large volumes of data with ease, is being used in this scenario, for data table creation and insertion of the insurance premium dataset into the table.

5.3 Export Data from Database

The data that is stored in Cassandra database is exported as a CSV file to be used for Data Pre-processing and Model Training.

5.4 Exploratory Data Analysis

This step of data visualization helps to showcase the relationship between independent and dependent variables and get more insights on the data via uni-variate, bi-variate & multi-variate analysis.

5.5 Feature Engineering

Here all the categorical variables have been encoded to turn them into numeric variables followed by standard scaling to get all the variables in a fixed range.

5.6 Model building & Testing

Once the Feature engineering step has been completed, the three machine learning algorithms, i.e Decision Tree Regressor, Random Forest Regressor, Gradient Boost Regressor will be fine-tuned using Grid search CV to get better results, and the one with the highest accuracy score will be used for predicting the cost of insurance premium for the individual user.

5.7 UI Integration

Front end development will be created using HTML/CSS framework where the functionality and design of the app will be created in such a manner to ensure smooth entry of data by the user.

5.8 Data Input from User

The physiological data from the user basically the age, sex, body mass index, no of children, smoker or non-smoker, region they belong to is collected.

5.9 Data Validation

The data provided by the user is then being processed by application.py file and validated.

5.10 Deployment

The project was deployed from GitHub into the AWS platform.

6. User Input / Output Workflow