

# Project:

# Insurance Premium Predictor

By-Mr.Aniket Dumbre

Prepared By

Mr.Aniket Dumbre

Approved By

Mr.Sunny Savita

<b>1. Abstract.....</b>	<b>3</b>
<b>2. Introduction.....</b>	<b>3</b>
2.1 What is Architecture Design? .....	3
2.2 Scope.....	3
<b>3. Data Description.....</b>	<b>4</b>
<b>4. Architecture.....</b>	<b>5</b>
<b>5. Architecture Description .....</b>	<b>6</b>
5.1.Data Insertion into Database .....	6
5.2.Training pipeline.....	6
5.2.1. Data Ingestion .....	6
5.2.2. Data Validation .....	6
5.2.3. Data Transformation .....	7
5.2.4. Model Training .....	7
5.2.5. Model Evaluation.....	8
5.2.6. Model Pusher .....	8
5.3.Data for prediction.....	8
5.4.Prediction Pipeline.....	8
5.4.1. Data Ingestion.....	8
5.4.2. Data validation.....	8
5.4.3. Data Transformation.....	9
5.4.4. Model Prediction.....	9
5.5.Deployment.....	9
<b>6. Document Version Control.....</b>	<b>9</b>

Prepared By

Approved By

Mr.Aniket Dumbre

Mr.Sunny Savita

## 1. 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.

## 2. Introduction

### 2.1 What is Architecture Design?

The goal of Architecture Design (AD) is 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 specification. It describes the modules so that the programmer can directly code the program from the document.

### 2.2 Scope

Architecture Design (AD) 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.

Prepared By	Approved By
Mr.Aniket Dumbre	Mr.Sunny Savita

### 3. Data Description

Dataset have 3 numerical features (age, bmi, children etc), 3 categorical features (sex, smoker, and region) and target feature 'expenses' as numerical feature.

**age** – age of individuals in years

**sex** – Male/Female

**no. of children** – no. of children person have

**bmi** – body mass index

**smoker** – whether person smokes or not

**region** – 4 region

**expenses** – expenses of person on health insurance

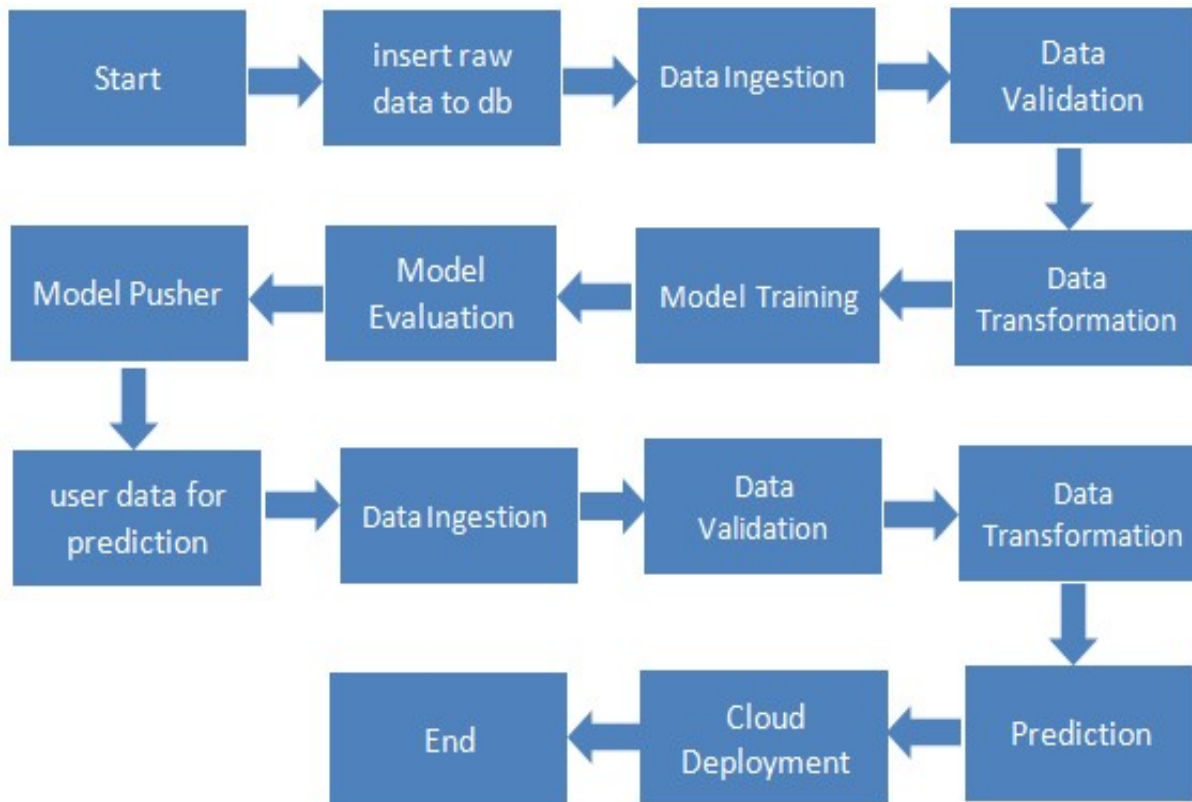
Prepared By

Approved By

Mr.Aniket Dumbre

Mr.Sunny Savita

#### 4. Architecture



Prepared By

Mr.Aniket Dumbre

Approved By

Mr.Sunny Savita

## 5. Architecture Description

### 5.1 Data Insertion into Database

- Use Mongo atlas database for storing our data which is NoSQL database.
- Read the dataframe file and convert it to json format and then dumb it to Mongoddb.
- Database Creation and connection –  
Create a database with the name “ipp” and collection name “insurance”.  
If the database and collection is already present, then the new database and collection are not created and new files are inserted in the already present collection.

### 5.2 Training Pipeline

#### 5.2.1 Data Ingestion

- Import collection data as pandas dataframe.
- Replace na values with NAN if present in dataset.
- Save updated dataset to ‘feature\_store ’ folder.
- Split dataset into train and test data and save it into ‘dataset’ folder.
- Save these folders in ‘artifact’ folder.

#### 5.2.2 Data Validation

- **Null values-** Drop the columns which have missing values more than specified threshold.
- **Data-type** – If required change data type of columns.
- **No of columns** -Validate the number of columns present in the files.
- **Data Drift**-Check whether new dataset is having major data drift or not.  
Save the report.yml in artifact folder.

Prepared By	Approved By
Mr.Aniket Dumbre	Mr.Sunny Savita

### 5.2.3 Data Transformation

Pipeline used for data pre processing. Create two pipelines for transforming input and target features.

- **One hot encoding** – Encode categorical columns using one hot encoder.
- **Handling missing values**- Use imputation techniques to impute missing values in columns eg. Simple imputer, KNN imputer etc.
- **Feature Scaling**- Try Standard scaler, min max scalar or robust scaler for scaling.
- **Outlier handling**-Use inter-quantile range method to handle outlier/robust scaler.
- **Imbalanced data**- Use SMOTETomek technique to handle imbalanced data.
- Save input feature transformation as transformer.pkl, target feature as transformer\_target.pkl, train data as train.npz and test data as test.npz in artifact folder.

### 5.2.4 Model Training

- There are two approaches for model training
  1. Using single algorithm
  2. Multimodal system
- We adapted multimodal approach where model is being trained on multiple algorithms and selected one which gives high accuracy after hyper parameter tuning.
- **Hyper parameter tuning:**

Hyper parameter tuning was done by using Grid search algorithm.
- Test model for under fitting or over fitting.
- Save best trained model in pickle format artifact folder as model.pkl.

Prepared By	Approved By
Mr.Aniket Dumbre	Mr.Sunny Savita

### 5.2.5 Model Evaluation

While re-training check whether current model performance with previous model.

### 5.2.6 Model Pusher

Save trained model and transformed feature files in saved\_models directory which will be used during model evaluation.

## 5.3 Data for prediction

Dump user data in db for prediction using AWS S3 bucket for batch prediction or use can predict insurance premium using flask app.

## 5.4 Prediction Pipeline

Apache airflow is used to monitor ML model performance. Once data is received from user, prediction pipeline gets triggered using github actions and model gets retrained and best model is selected for prediction.

### 5.4.1 Data Ingestion

- Import data frame file from db storage.
- Replace na values with NAN if present in dataset.

### 5.4.2 Data Validation

- **Null values**- Drop the columns which have missing values more than specified threshold.
- **Data-type** – If required change data type of columns.
- **No of columns** -Validate the number of columns present in the files.
- **Data Drift**-Check whether new dataset is having major data drift or not. Save the report.yml in artifact folder.

Prepared By	Approved By
Mr.Aniket Dumbre	Mr.Sunny Savita



#### 5.4.4 Prediction

Predict the target feature and saved it in S3 bucket in .csv format. Also user can predict insurance premium using flask app.

#### 5.5 Deployment

Deploy model to the AWS. Batch prediction file stored in S3 bucket.

### 6 Document change log

Sr. No.	Rev. No.	Rev. Date	Comment	Author
1	00	10-11-22	Initial release	Aniket Dumbre

Prepared By

Mr.Aniket Dumbre

Approved By

Mr.Sunny Savita