INSURANCE PREMIUM PREDICTION

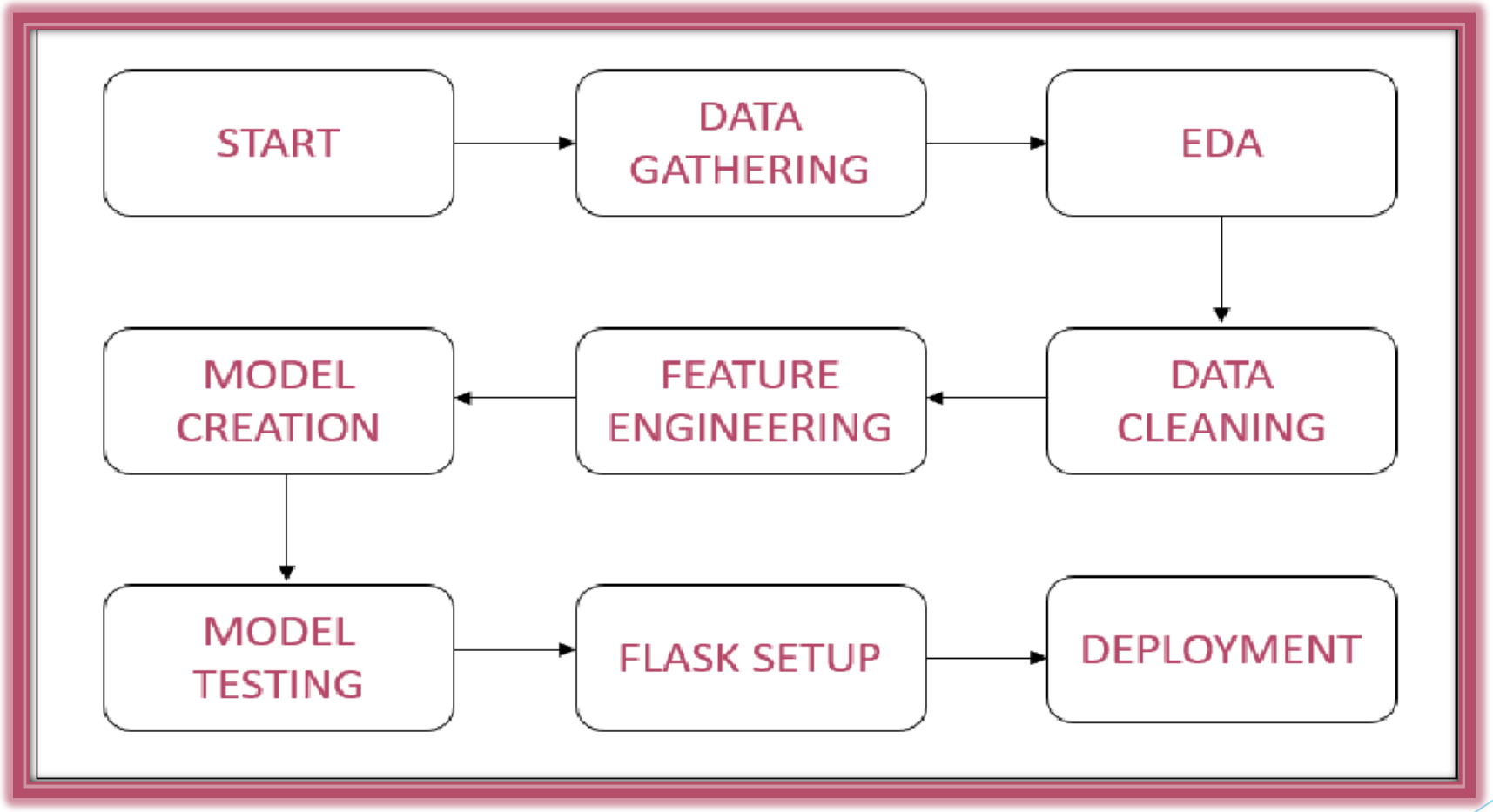
Objective :

- The goal of this project is to give an estimate of how much they need for their individual health situation and build a solution that should be able to predict the premium of the personnel for health insurance.

- **Benefits :**

  - Gets an idea about how much amount is required annually according to their health status.
  - This can help a person focus more on the health aspect of insurance.
  - Help in giving premium health insurance.

Architecture :



Data Collection and Data Validation :

- The dataset was taken from the Kaggle competition page.
- Data type of columns – Validating the data type of the columns if
  - wrong, then it was corrected.
- Null values in columns – Validating the column in the dataset have null values or missing information.
- Duplicate values in the dataset – Deleting duplicate records.

Model Training :

- **Data Pre-processing :**

- Performing EDA to get insights into the data like identifying distribution, numerical features, categorical features, outliers, missing values, duplicate data etc.

- Check any null values present in the dataset. If present, then imputes those null values.

- Converting Categorical features into Numerical Features.

- Scale down the data for better results.

- **Model Selection :**

- After pre-processing and model training, we find the best model for premium prediction. The model is trained on multiple regression algorithms like Linear Regression, Decision Trees, Random Forest, Gradient Boosting, and Grid Search CV for best parameters.

- After prediction, we will find the accuracy of those predictions using evaluation metrics like RMSE (Root mean squared error) and r2_score (R-squared).

- **Predictions :**

- Then all the trained models were used for validating the test set.

- We perform pre-processing techniques on it.

- The best RMSE and r2 score models were saved for developing API for the prediction of premium.

Q & A :

Q1) What is the source data?

- The source of the data is Kaggle. The data is in the form of a CSV file.

Q2) What was the type of the data?

- The data was a combination of categorical and numerical values.

Q3) What's the complete flow you followed in this project?

- Refer to the 3rd slide for a better understanding.

Q4) What techniques were you using for data pre-processing?

- Visualizing the relation of independent variables with each other and dependent variables.
- Checking distribution of Continuous variables.
- Checking for any null values in the dataset.
- Checking for duplicate values.
- Converting categorical data to numerical values.
- Scaling the data.

Q & A :

- Q5) How training was done or what models were used?
- ✓ Before training the model, the dataset is divided into a training set and a
- testing/validation set.
- ✓ The scaling was performed of training and validation set.
- ✓ The categorical columns were converted into numeric values.
- ✓ Algorithms like Linear Regression, Decision Trees, Random Forest, and Gradient Boosting were used for model training. Based on RMSE & r2_score, the GradientBoostingRegressor model was selected for Grid Search CV for best parameters after the hypertunning model was saved for Validation.

- Q6) How prediction was done?
- ✓ Based on the trained model, the prediction was performed. We also created an API interface for estimating the cost of the premium based on personal
- health information/status.

- **Q & A :**

- **Q7) What are the different stages of deployment?**
  - When the model is ready, we deploy it on the AWS platform.

THANK YOU