Architecture Document
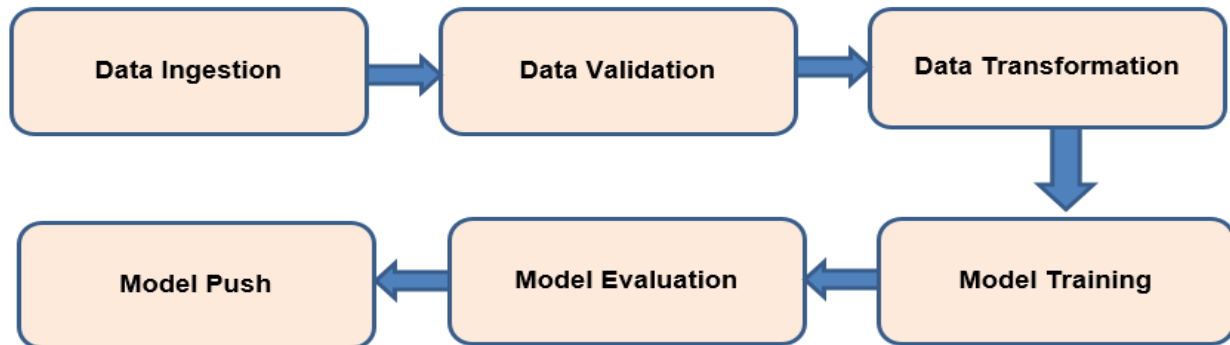
Adult Census Income Prediction
(With complete CI/CD pipelines)

Contents
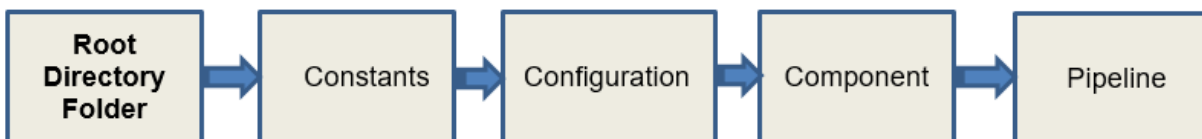
1. Architecture:
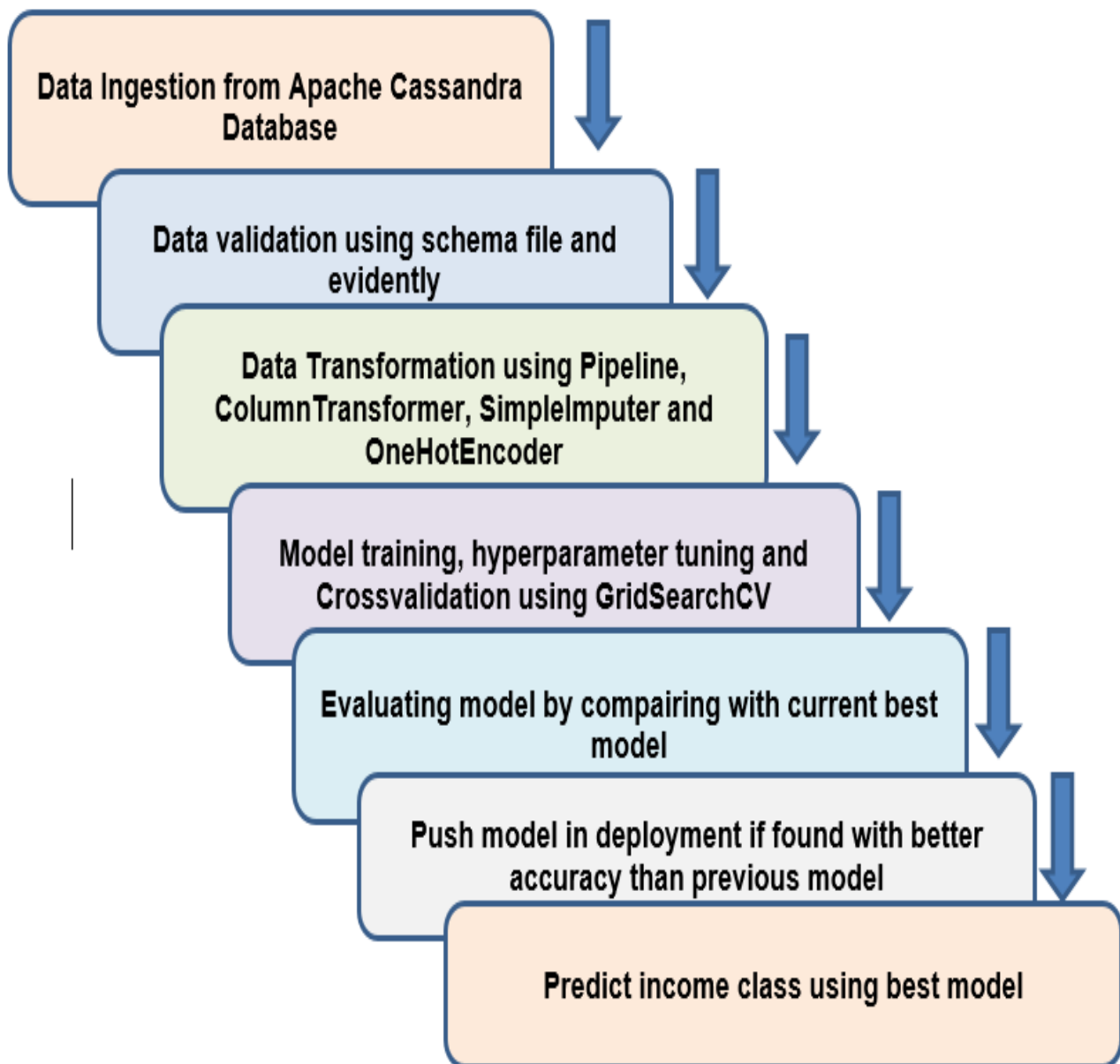
1.1 Components of Machine Learning Pipeline

1.2 Coding flow for building project structure

1.3 Complete pipeline flow

Data Ingestion from Apache Cassandra Database

Data validation using schema file and evidently

Data Transformation using Pipeline, ColumnTransformer, SimpleImputer and OneHotEncoder

Model training, hyperparameter tuning and Crossvalidation using GridSearchCV

Evaluating model by compairing with current best model

Push model in deployment if found with better accuracy than previous model

Predict income class using best model

2. Architecture Description:

1. **Data Description**:

    The dataset named Adult Census Income is available in kaggle and UCI repository. This data was extracted from the 1994 census bureau dataset by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The prediction task is to determine whether a person makes over $50K a year or not.

2. **Data Ingestion:**

    Loading data form the apache Cassandra database by establishing connection with database cluster. Ingested data will be in the form of csv file. Splitting file into train and test file using stratified split.

3. **Data Validation:**

    Validating data using schema.yaml file and generating report using evidently package. Checking for data drift.

4. **Data transformation:**

    Removing unnecessary columns from dataframe.Transforming data using various pipeline, columntransformer, simpleimputer and onehotencoder. Saving preprocessing object in pickle file.

5. **Model Training:**

    Training different model using transformed with crossvalidation and gridsearchCV.

6. **Model Evaluation:**

    Evaluating model performance using various performance metrics. Finding the best model by compairing different models.

7. **Model Push:**

    Deploying the model in production only if better model found. Prediction will be done using this model.

8. **Deployment on railway:**

    Deployment using github repository on railway cloud. Automating deployment using github actions. Then, app will automatically get updated when we commit any changes to main branch of github repository.

3. Application Interface