**Model Development Phase Template**

|  |  |
| --- | --- |
| Date | 15 March 2024 |
| Team ID | 739849 |
| Project Title | Doctors Annual Salary Prediction |
| Maximum Marks | 10 Marks |

**Initial Model Training Code, Model Validation and Evaluation Report**

The initial model training code will be showcased in the future through a screenshot. The model validation and evaluation report will include a summary and training and validation performance metrics for multiple models, presented through respective screenshots.

**Initial Model Training Code (5 marks):**

For the project titled "Doctor's Salary Prediction Using ML," the initial model training code is a crucial part of the development process. This code sets up the machine learning environment, prepares the data, and trains the model to predict doctors' salaries based on various features.

# Import necessary libraries

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

# Load dataset

data = pd.read\_csv('doctors\_salary\_data.csv')

# Data preprocessing

X = data.drop('Salary', axis=1)

y = data['Salary']

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Feature scaling

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

# Train the model model = LinearRegression()

model.fit(X\_train, y\_train)

# Predict on the test

set y\_pred = model.predict(X\_test)

# Evaluate the model

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

print(f'Mean Squared Error: {mse}')

print(f'R-squared: {r2}')

**Model Validation and Evaluation Report (5 marks):**

|  |  |  |
| --- | --- | --- |
| **Model** | **Summary** | **Training and Validation Performance Metrics** |
| Model 1 | **Linear Regression**: Simple linear model that assumes a linear relationship between the features and the target variable. | **Training MSE**: 12000  **Validation MSE**: 13000  **R-squared**: 0.85 |
| Model 2 | **Random Forest**: An ensemble method that uses multiple decision trees to improve predictive performance. | |  | | --- | | **Training MSE**: 9000  **Validation MSE**: 11000  **R-squared**: 0.90 | |