Case Study Summary:

Here given set of attributes employees data and need to classify or build a model trained model which can predict the employee performance based on factors as inputs. **It is Classification problem since here we need to classify employee performance i.e. scale of 2,3, or 4.**

Steps that involved

1)Getting the data and load the data & do Cleaning of Data, Exploration of Data to make it consistent for Analysis.(Exploratory Data Analysis.ipynb)

2) Explore the data, and understand what the data is all about and feature selection.
Performance of the model depends.
Choice of Algorithm
Feature Selection
Feature Creation
Model selection

Feature selection is also known as variable selection. (Clearly documented in Cleaning of Data & Exploration of Data.ipynb).

3)Choosing of Algorithm **(Clearly documented in IABAAC project employee performance with 94% with RandomforestClassifier.ipynb**

what features do you have? What are their nature, what values can they have, what are their distributions?
For example, tree-based models are exceptionally good with categorical features.

What kind of analysis do we need?
If we are going to make a deep, detailed analysis of the learned dependency, you will need a highly interpretable model, say, linear/polynomial regression else can go with black box algorithms like Neural network.

so here selected tree based models since here some predictor variables are categorical variables.

Next important reason is other algorithms prone to outliers , Reduction in overfitting and Less variance

so chosen tree based model ie Random forest which is simple and easy to interpret as well

4) Several algorithms were used to find the best fit. Here are the two best fits:

- DecisiontreeRegressor: 90.80%
- RandomForestRegressor: 93.75%

Tried with some other Regressor algorithms like SVR,MLPRegressor but accuracy not achieved as expected.

- So finally concluded the project by taking random forests algorithm with 93.75% Since Random forests overcome several problems with decision trees since single decision tree may over fit the data, including:

- Reduction in overfitting

- Less variance

So in almost all cases, random forests are more accurate than decision trees