STAGES INVOLVED IN DEVELOPING AN ML MODEL:

Developing a reliable and efficient machine learning model is a complex process that requires the completion of several essential phases. These stages play a critical role in ensuring that the model is capable of effectively addressing a specific problem and delivering accurate predictions.

Below are the eight stages that are involved in the development of a robust and optimized machine learning model:

1. **Data Collection**: The act of gathering data that is pertinent to the goals and objectives of your AI project is known as data collection. This process ultimately leads to the creation of a dataset, which is essentially a compilation of data that is ready to be utilized for training and input into a machine learning model.

2. **Data cleaning and processing**: This is a procedure of addressing incorrect, corrupted, inaccurately formatted, duplicate, or incomplete data present in a dataset. It involves either correcting or removing such data to ensure the accuracy and integrity of the dataset.

3. **Model selection:** In machine learning, selecting the appropriate model or algorithm for a particular problem is referred to as model selection. This involves assessing multiple machine learning models based on various criteria, including accuracy, complexity, and generalization performance, and ultimately choosing the model that demonstrates the highest level of performance.

4. **Training and testing the model:** The selected models are trained with 70 percent of the data and tested with the remaining 30 percent of the data. The data is split into train and test sets to evaluate how well the model performs. The train set is used to fit the model, and the statistics of the train set are known. The test set is solely used for predictions. This could prevent the occurrence of ‘overfitting’, which is crucial in machine learning, as it describes an undesirable behavior of the model where it produces accurate predictions for the training data but fails to generalize to new, unseen data.

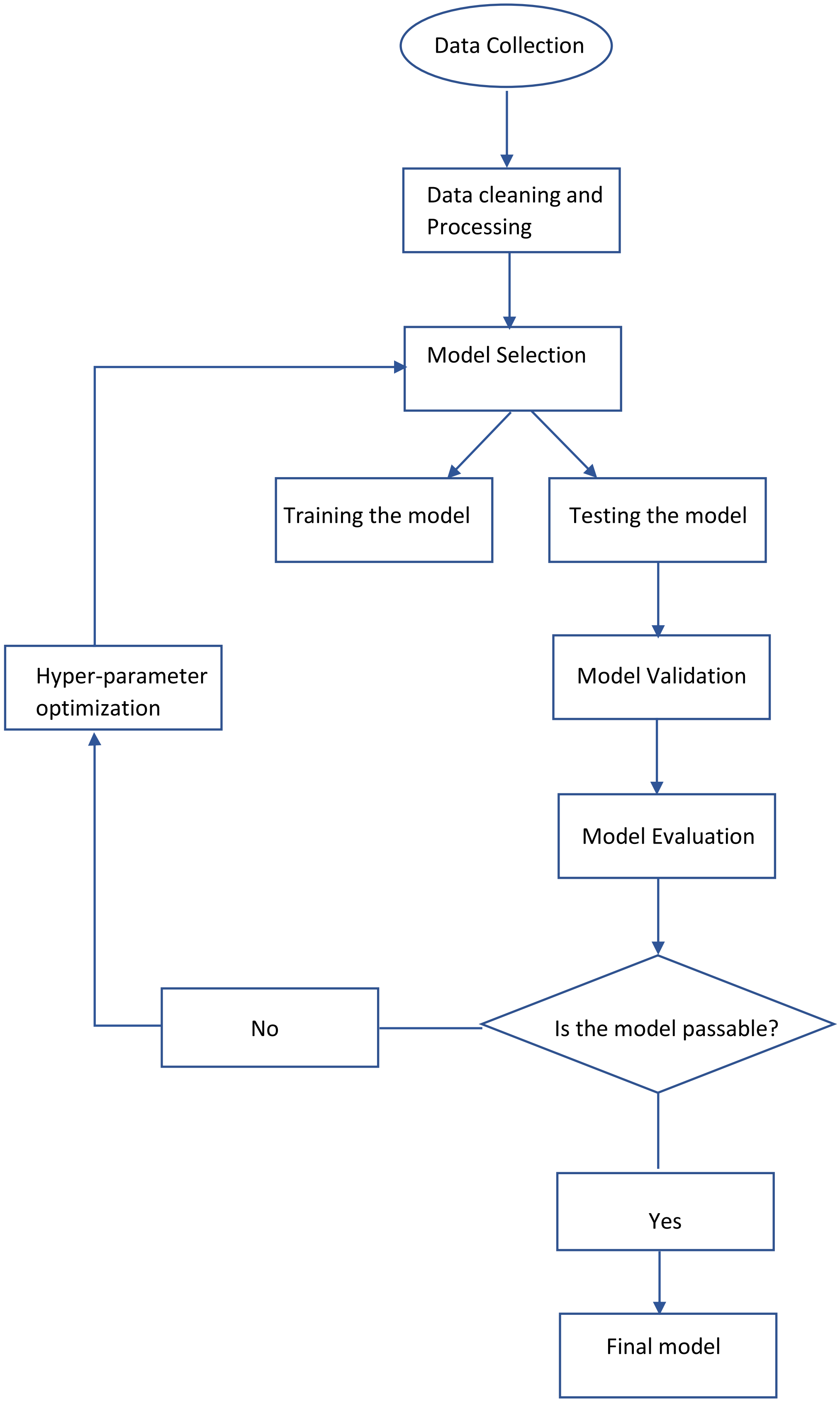
5. **Model validation:** To validate and test the performance of the model, resampling methods such as K-Fold cross validation, Leave-One-Out Cross Validation. These methods examine how the model performs on some testing data that was not used to fit or train the model.

6. **Model Evaluation**: In order to compare and analyze the performance of the machine learning models, various evaluation metrics are examined, that give prediction error rates and correlation between the variables. Part of them is Correlation coefficient (R2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

7. **Hyper-parameter optimization:** Hyperparameter optimization in machine learning involves identifying the optimal combination of pre-defined settings, or hyperparameters, for a given model to achieve peak performance. These settings, such as learning rate and regularization strength, cannot be learned from the data and must be defined beforehand.

8. **Final model determination:** The final ML model is determined by evaluating different models, optimizing the parameters, and assessing the performance metrics.

Stages involved in a developing an effective model.



PERFORMANCE METRICS IN MACHINE LEARNING

In order to compare and analyze the performance of the machine learning models, various prediction error rates are examined. Some of them are:

- **Correlation coefficient (R^2):** In machine learning, the correlation coefficient is a statistical measure that gauges the strength and direction of the relationship between two variables (true value of the data and the predicted value of the model). It ranges from -1 to 1, with -1 representing a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation. A high positive or negative correlation coefficient between features and the target variable can make them valuable predictors in a model. However, having highly correlated features may lead to overfitting or reduced model interpretability. Thus, when selecting features for a machine learning model, it's important to take the correlation coefficient into account.

- **Root mean square error (RMSE):** It is a widely used performance metric in machine learning for evaluating the accuracy of regression models. It calculates the difference between predicted and actual values by taking the square root of the average of the squared differences between them. RMSE is expressed in the same units as the target variable and can be used to measure the average distance between predicted and actual values. RMSE is commonly applied to machine learning tasks such as predicting housing or stock prices where the objective is to minimize the difference between predicted and actual values. A lower RMSE indicates better accuracy, whereas a higher RMSE indicates lower accuracy. However, RMSE has limitations such as sensitivity to outliers and the inability to provide information on the direction of the error. Consequently, RMSE is typically used in combination with other performance metrics to assess the overall model performance.

- **Mean absolute error (MAE)**: It measures the average absolute difference between predicted and actual values in a dataset. MAE is calculated by taking the average of the absolute differences between predicted and actual values. Unlike the RMSE, the MAE is not influenced by outliers, as it takes the absolute value of the difference between predicted and actual values. MAE is expressed in the same units as the target variable, and a lower MAE indicates better accuracy. MAE is often used in machine learning tasks such as predicting house prices or estimating demand for a product. However, like RMSE, it does not provide information on the direction of the error, and it may not be as sensitive to larger errors as RMSE. Therefore, it is commonly used in conjunction with other performance metrics to evaluate the overall accuracy of a model.

- **Mean absolute percentage error (MAPE):** It measures the average percentage difference between the predicted and actual values in a dataset. MAPE is calculated by taking the absolute difference between predicted and actual values, dividing it by the actual value, and then multiplying by 100 to express the result as a percentage. The average of these percentage differences is then calculated to obtain the MAPE. MAPE is expressed as a percentage and provides an indication of the size of the error relative to the actual value. A lower MAPE indicates better accuracy, with 0 indicating a perfect prediction. However, like other error metrics, MAPE has some limitations, such as being sensitive to outliers and not providing information on the direction of the error. As a result, it is commonly employed along with other metrics to assess the overall performance of a model.