Department of Computer Science and Engineering (Data Science)

Subject: Applied Data Science (DJ19DSL703)

Experiment - 7

(Modelling and Optimisation Trade-off)

Name: Sameep Karia
SAP ID: 60009200082

Aim: To optimize the model.

Theory:

Hyper parameter and parameter tuning are crucial processes in machine learning that help optimize the performance of a model. They involve adjusting the settings of a machine learning algorithm to improve its ability to make accurate predictions. The 'K' in KNN can be regarded as the parameter whereas number of leaf nodes, number of branches, depth of a tree are hyper parameters of a Decision Tree or Random Forest.

1. Parameters:
Parameters are the internal settings or coefficients learned by the machine learning model during training. These values are adjusted automatically through the training process, and they determine how the model transforms input data into predictions. In a simple linear regression model, for example, the parameters are the slope and intercept of the regression line. Parameters are learned from the data and are specific to the model and its training data.

2. Hyper parameters:
Hyper parameters are settings or configurations that are not learned from the data but are set by the machine learning engineer or data scientist before training begins.
These settings control the behaviour of the learning algorithm itself and can significantly impact the model's performance.

Examples of hyper parameters include the learning rate in gradient descent, the number of hidden layers and neurons in a neural network, or the regularization strength in a support vector machine.
Hyper parameters are set by trial and error or by using techniques like grid search or random search to find the best values for a specific problem.

Parameter Tuning:
Parameter tuning involves adjusting the internal parameters of a machine learning model to optimize its performance on a specific dataset. This is typically done automatically during the training process, as the model learns the best values for its parameters to minimize a predefined loss function.
For example, in a neural network, the weights and biases of the neurons are adjusted during training to minimize the prediction error. Parameter tuning is specific to the model and is intrinsic to the training process.

Hyper parameter Tuning:
Hyper parameter tuning, also known as hyper parameter optimization, focuses on finding the best settings for hyper parameters to improve a model's generalization and performance. Hyper parameter

tuning is an iterative process that typically involves searching through a range of hyper parameter values to find the combination that produces the best results on a validation dataset. Common techniques for hyper parameter tuning include grid search, random search, Bayesian optimization, and automated tools like scikit-learn's GridSearchCV or hyperopt.
Hyper parameter tuning is crucial for optimizing the performance of a machine learning model on a specific problem. In summary, parameters are internal settings learned by the model during training, while hyper parameters are external settings configured before training. Parameter tuning focuses on optimizing internal settings, while hyper parameter tuning is about finding the best external settings to improve model performance. Both processes are essential for developing effective machine learning models.

Lab Assignment:

- Students need to try various permutations and combinations for getting the right set of parameters and hyper parameter values.
- Post this your model is ready for deployment. Create a pickle file for your model such that it will take the input and produce the model response as the output. This should be then converted to an API using Flask/FastAPI frameworks.
- The API will have an endpoint (say /run-model) which will be a POST request. It will take the data records for which the prediction is to be made as input, run the model in the background and the API response will be the output from the model. For example, I can provide an email body as input to the API and it will return a response if the content can be regarded as spam or not. Or I can even provide multiple input fields for a house description and the endpoint will provide the estimated house price in response.

```
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
import pandas as pd

df = pd.read_csv('/content/car_evaluation.csv')
df.head()
```

	vhigh	vhigh.1	2	2.1	small	low	unacc
0	vhigh	vhigh	2	2	small	med	unacc
1	vhigh	vhigh	2	2	small	high	unacc
2	vhigh	vhigh	2	2	med	low	unacc
3	vhigh	vhigh	2	2	med	med	unacc
4	vhigh	vhigh	2	2	med	high	unacc

```
df.shape
```

```
(1727, 7)
```

```
df.describe()
```

	vhigh	vhigh.1	2	2.1	small	low	unacc
count	1727	1727	1727	1727	1727	1727	1727
unique	4	4	4	3	3	3	4
top	high	high	3	4	med	med	unacc
freq	432	432	432	576	576	576	1209

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1727 entries, 0 to 1726
Data columns (total 7 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   vhigh    1727 non-null   object
 1   vhigh.1  1727 non-null   object
 2   2        1727 non-null   object
 3   2.1      1727 non-null   object
 4   small    1727 non-null   object
 5   low      1727 non-null   object
 6   unacc    1727 non-null   object
dtypes: object(7)
memory usage: 94.6+ KB
```

```
col_names = ['buying', 'maint', 'doors', 'persons', 'lug_boot', 'safety', 'class']

df.columns = col_names

col_names
```

```
['buying', 'maint', 'doors', 'persons', 'lug_boot', 'safety', 'class']
```

```
df.isnull().sum()
```

```
buying      0
maint       0
doors       0
persons     0
lug_boot    0
safety      0
class       0
dtype: int64
```

```
df['class'].value_counts()
```

```
unacc    1209
acc       384
```

```
     good       69
     vgood      65
     Name: class, dtype: int64
```

```
X=df.drop(['class'],axis=1)
```

```
y=df['class']
```

```
from sklearn.preprocessing import OneHotEncoder
```

```
encoder = OneHotEncoder(sparse=False, drop='first')
categorical_columns = X.select_dtypes(include=['object']).columns
X_encoded = encoder.fit_transform(X[categorical_columns])
column_names = encoder.get_feature_names_out(categorical_columns)
X_encoded = pd.DataFrame(X_encoded, columns=column_names)
X = pd.concat([X.drop(columns=categorical_columns), X_encoded], axis=1)

# 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)

# Define your machine learning model
model = RandomForestClassifier()

# Define a grid of hyperparameters to search
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

# Create GridSearchCV object
grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=5, scoring='accuracy')

# Fit the grid search to the data
grid_search.fit(X_train, y_train)

# Get the best hyperparameters from the grid search
best_params = grid_search.best_params_
print("Best Hyperparameters:", best_params)

# Train the model with the best hyperparameters on the entire training set
best_model = RandomForestClassifier(**best_params)
best_model.fit(X_train, y_train)

# Evaluate the model's performance on the test set
accuracy = best_model.score(X_test, y_test)
print("Model Accuracy:", accuracy)
```

```
     /usr/local/lib/python3.10/dist-packages/sklearn/preprocessing/_encoders.py:868: FutureWarning: `sparse` was renamed to `sparse_output` i
       warnings.warn(
     Best Hyperparameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
     Model Accuracy: 0.8815028901734104
```