Introduction to Data Science (DS201.3)

- Coursework Report
- R K G Tharushika – 36508
- Tasks: Classification & Regression

Introduction

This report explains the methodology, implementation, and results of two data science tasks completed using Python and Jupyter Notebook. Task 01 focuses on a classification problem using the Breast Cancer dataset, while Task 02 focuses on a regression problem using the K-Drama dataset. All explanations in this report are directly based on the code implemented in the submitted notebooks Task 01.ipynb and Task 02.ipynb.

- Task 01: Classification
- Dataset: Breast Cancer Dataset

1. Dataset Appropriateness and Relevance

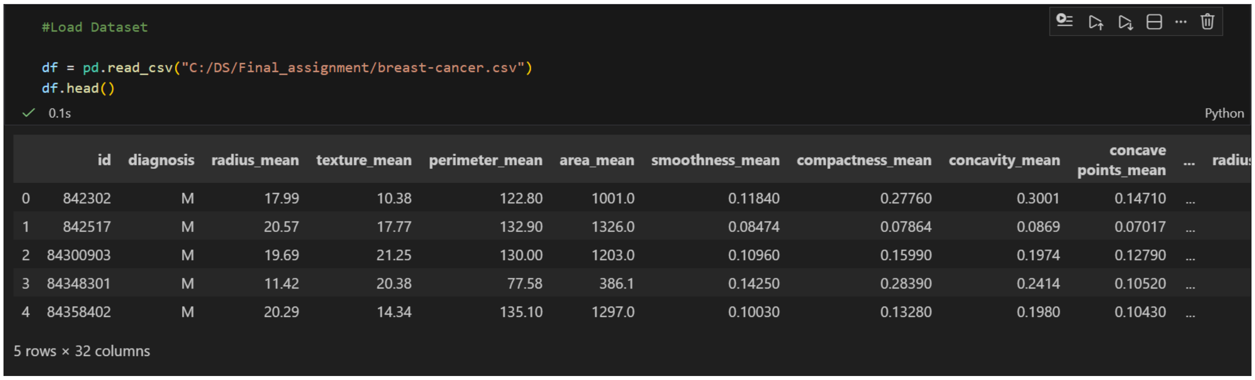
- The breast cancer dataset is well-suited for a classification task because the target variable, diagnosis, is categorical with two classes. It contains numerical features derived from medical imaging, which relate directly to tumor characteristics. This makes the dataset significant for a real-world healthcare issue where accurate classification is crucial for early diagnosis and treatment.

2. Methodology

Data Preprocessing

- First, we examined the dataset to understand its structure and features. We removed the id column as it does not help with prediction. We encoded the target variable, diagnosis, into numerical values, labeling benign tumors as 0 and malignant tumors as 1.

- Since the dataset had no missing values, no imputation was necessary.



```
#Load Dataset

df = pd.read_csv("C:/DS/Final_assignment/breast-cancer.csv")
df.head()
```

✓ 0.1s Python

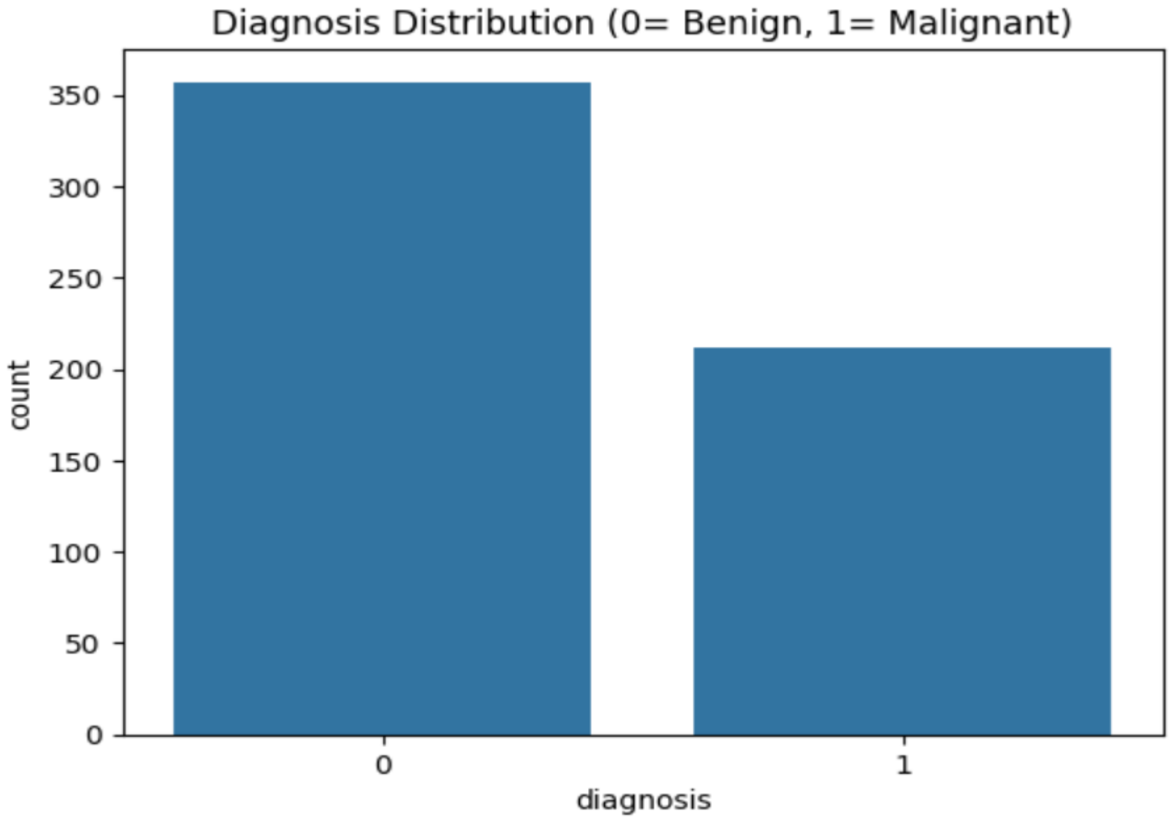
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	radiu
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

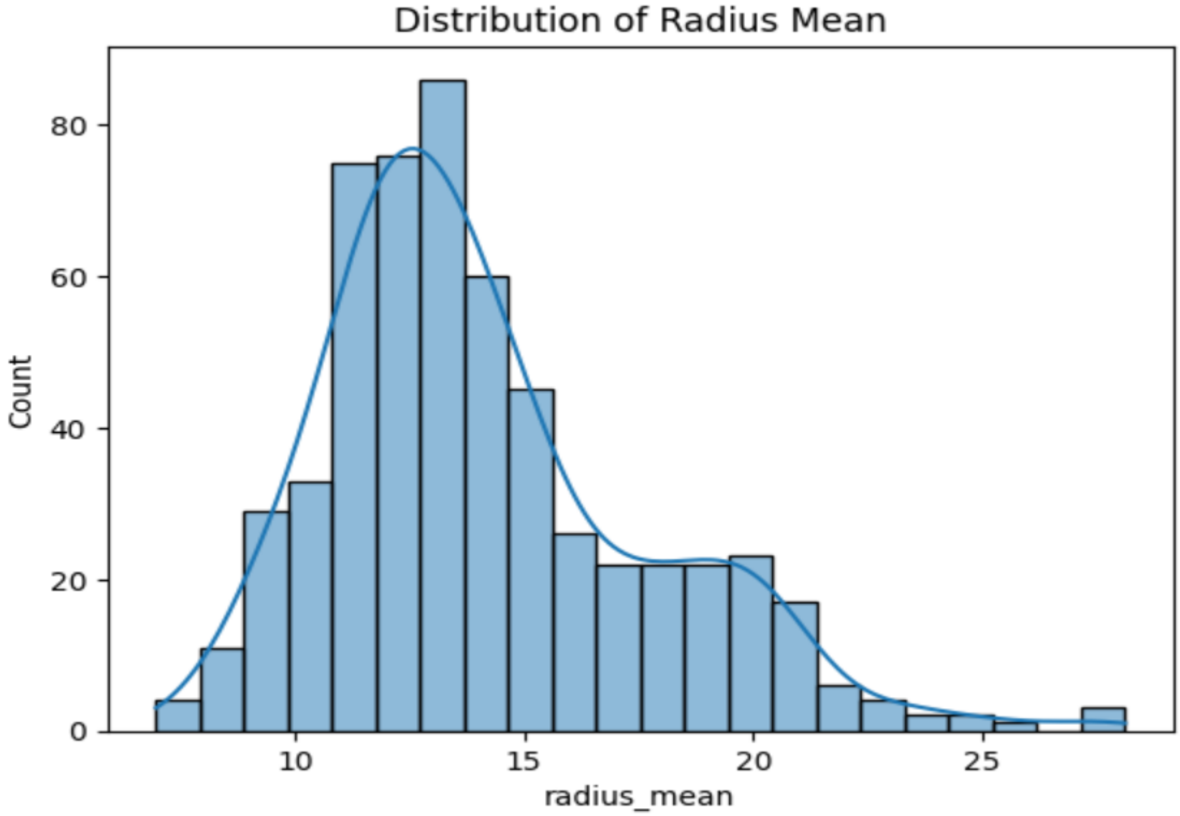
5 rows × 32 columns

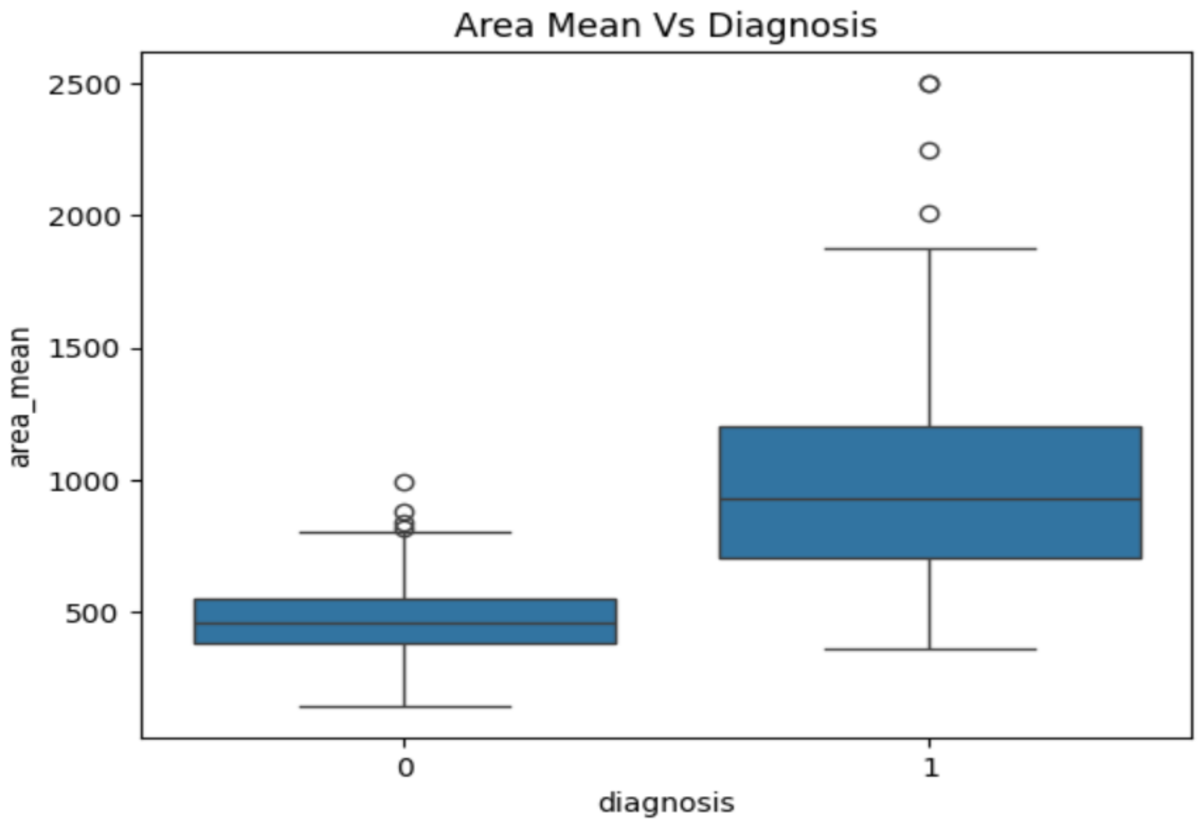
Exploratory Data Analysis(EDA)

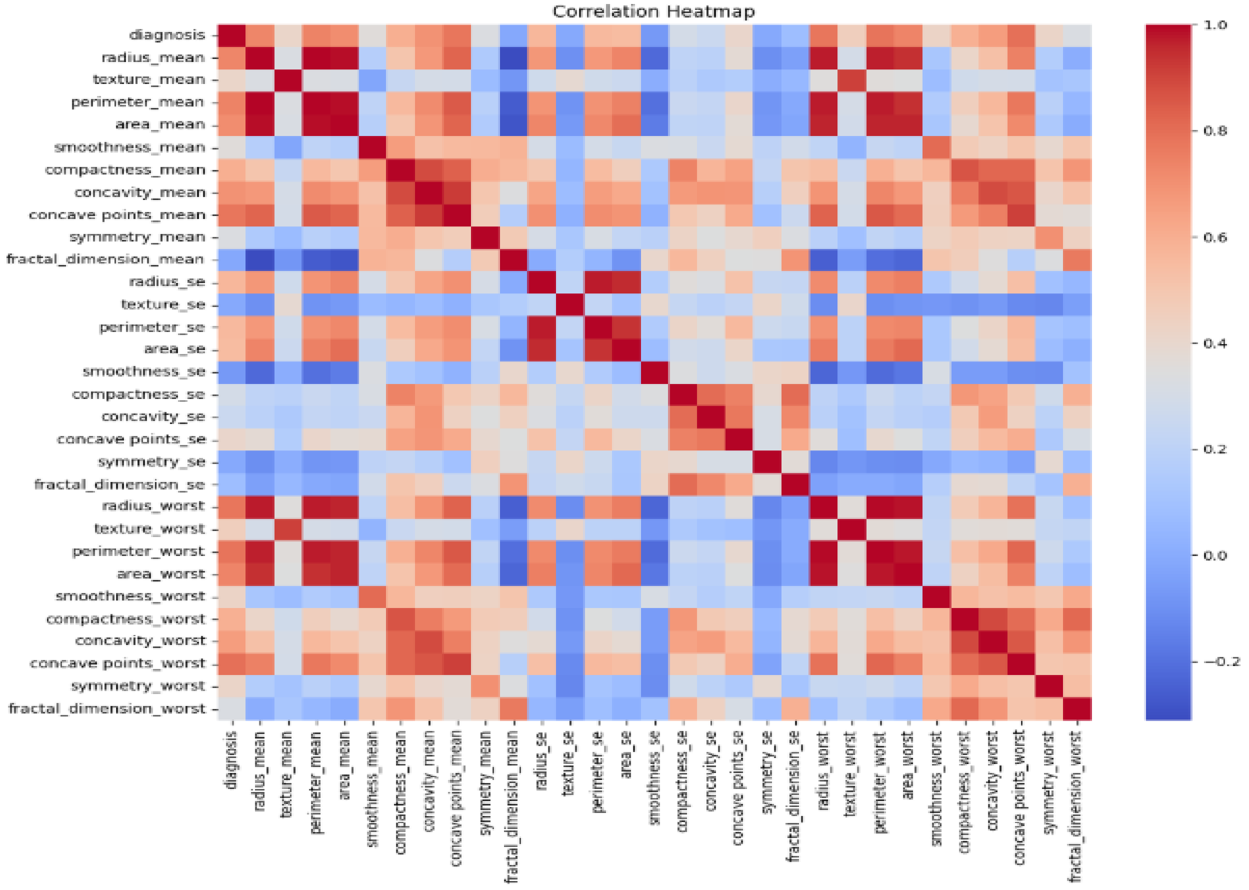
- We performed EDA to explore data distribution and relationships between features. Count plots visualized class balance, while histograms and boxplots examined feature distributions.

- A correlation heatmap helped identify strongly related features. The analysis showed that malignant tumors typically have higher radius and area values compared to benign tumors.


Diagnosis Distribution (0= Benign, 1= Malignant)
count
350
300
250
200
150
100
50
0
0
1
diagnosis

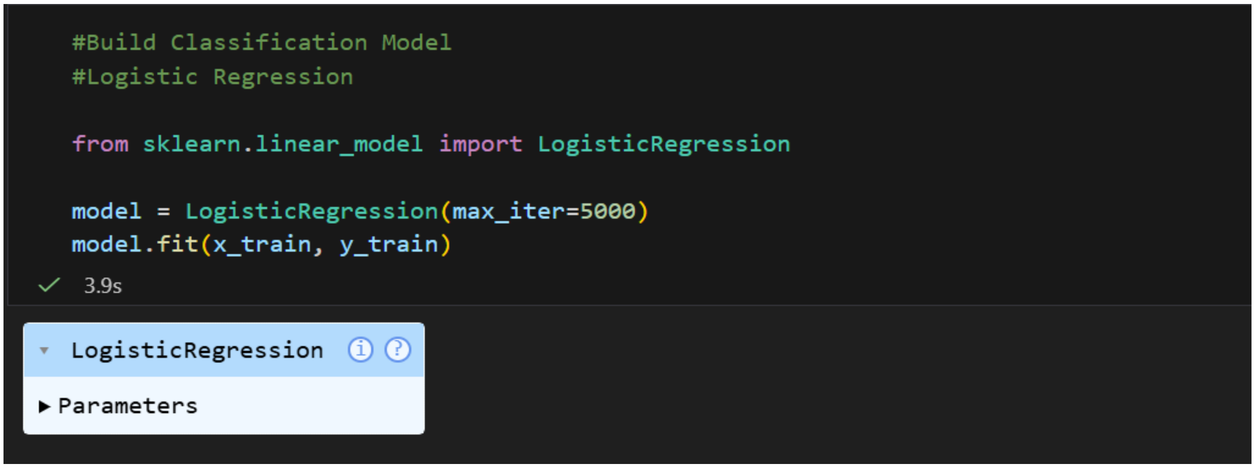

Distribution of Radius Mean
Count
80
60
40
20
0
10
15
20
25
radius_mean


Area Mean Vs Diagnosis
area_mean
2500
2000
1500
1000
500
0
1
diagnosis


Correlation Heatmap
diagnosis
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
symmetry_mean
fractal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave points_se
symmetry_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave points_worst
symmetry_worst
fractal_dimension_worst
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Model Selection

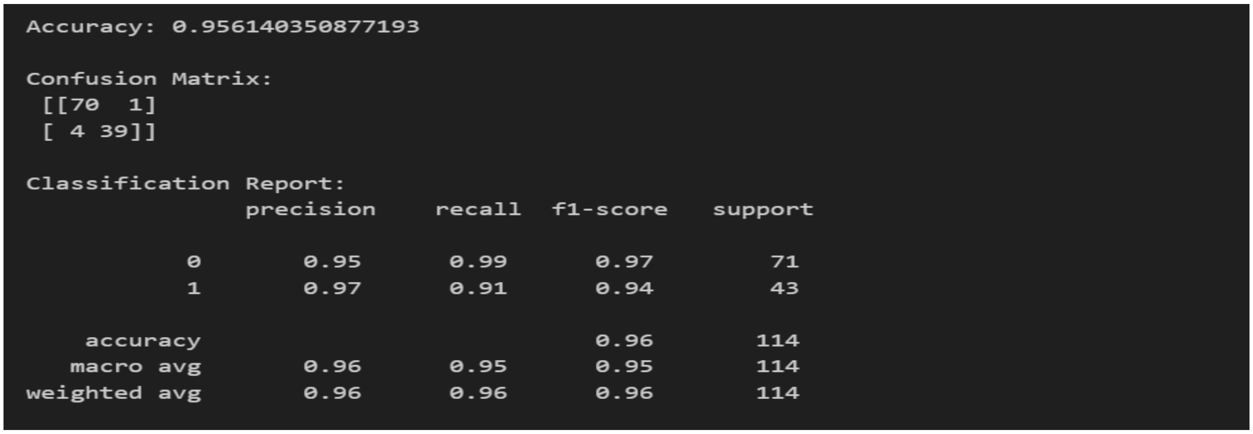
- We chose Logistic Regression as the classification algorithm since it fits well for binary classification problems and offers interpretable results. We split the dataset into 80% training data and 20% testing data before training the model.



```
#Build Classification Model
#Logistic Regression

from sklearn.linear_model import LogisticRegression

model = LogisticRegression(max_iter=5000)
model.fit(x_train, y_train)
```

3. Results and Interpretation

- We evaluated the logistic Regression model using accuracy, a confusion matrix, and a classification report. The model achieved high accuracy, including strong classification performance.
- The confusion matrix revealed a low number of false negatives, which is especially crucial in medical diagnosis. Precision, recall, and F1 scores were balanced across both classes, confirming the model’s reliability.



```
Accuracy: 0.956140350877193

Confusion Matrix:
 [[70  1]
 [ 4 39]]

Classification Report:
               precision    recall  f1-score   support

           0       0.95      0.99      0.97        71
           1       0.97      0.91      0.94        43

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

4. Conclusion(Task 01)

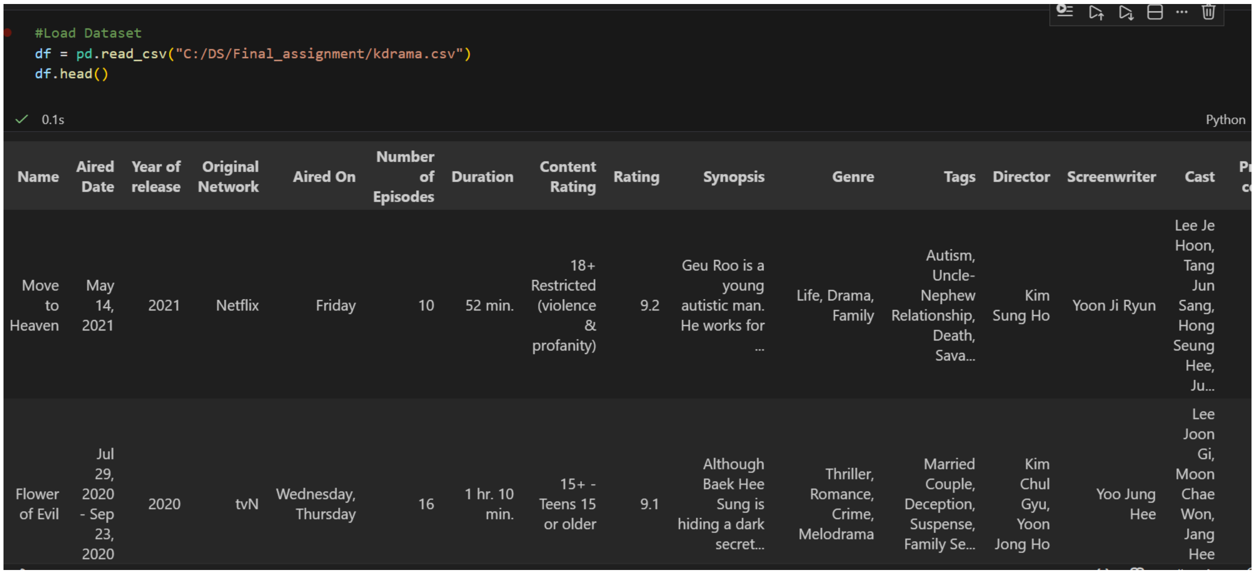
- The classification task successfully showed the use of Logistic Regression to differentiate between malignant and benign breast tumors.
- The results underscore the potential of machine learning models to assist medical professionals in early cancer detection. Future improvements could involve testing more sophisticated classifiers to further improve performance.

Task 02:Regression

Dataset: Korean Drama (K-drama) dataset

1. Dataset Appropriateness and Relevance

- The K–drama dataset is suitable for a regression task because the target variable, rating, is continuous. The dataset includes numerical features such as year of release, number of episodes, and episode duration, which serve as meaningful predictors for estimating audience ratings. This makes the dataset significant for understanding factors that affect viewer preferences.



```
#Load Dataset
df = pd.read_csv("C:/DS/Final_assignment/kdrama.csv")
df.head()
```

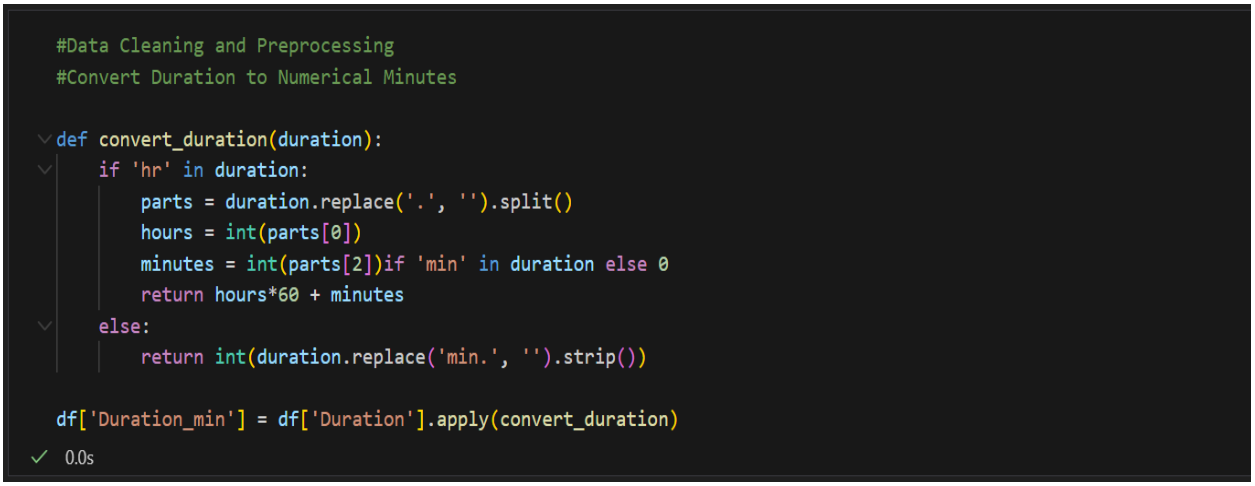
✓ 0.1s Python

Name	Aired Date	Year of release	Original Network	Aired On	Number of Episodes	Duration	Content Rating	Rating	Synopsis	Genre	Tags	Director	Screenwriter	Cast	Pr c
Move to Heaven	May 14, 2021	2021	Netflix	Friday	10	52 min.	18+ Restricted (violence & profanity)	9.2	Geu Roo is a young autistic man. He works for ...	Life, Drama, Family	Autism, Uncle-Nephew Relationship, Death, Sava...	Kim Sung Ho	Yoon Ji Ryun	Lee Je Hoon, Tang Jun Sang, Hong Seung Hee, Ju...	
Flower of Evil	Jul 29, 2020 - Sep 23, 2020	2020	tvN	Wednesday, Thursday	16	1 hr. 10 min.	15+ - Teens 15 or older	9.1	Although Baek Hee Sung is hiding a dark secret...	Thriller, Romance, Crime, Melodrama	Married Couple, Deception, Suspense, Family Se...	Kim Chul Gyu, Yoon Jong Ho	Yoo Jung Hee	Lee Joon Gi, Moon Chae Won, Jang Hee	

2. Methodology

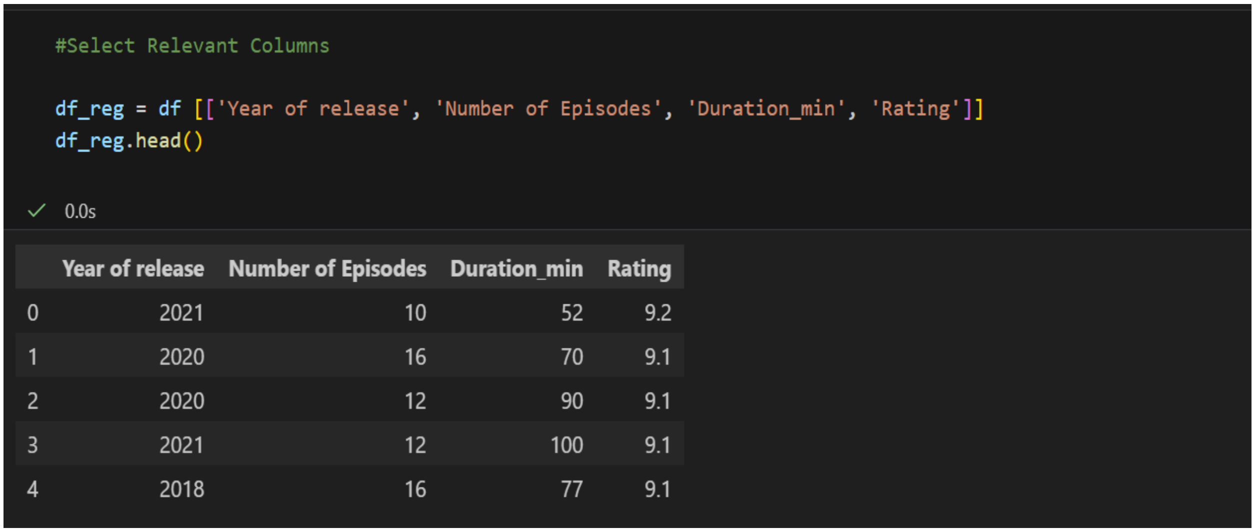
Data Preprocessing

- We inspected the dataset to find data quality issues. The duration column was originally in text format and was converted into numerical minutes to fit the regression analysis. We handled missing values in the duration column using mean imputation. We should use relevant numerical features for model training.



```python
#Data Cleaning and Preprocessing
#Convert Duration to Numerical Minutes

def convert_duration(duration):
    if 'hr' in duration:
        parts = duration.replace('.', '').split()
        hours = int(parts[0])
        minutes = int(parts[2])if 'min' in duration else 0
        return hours*60 + minutes
    else:
        return int(duration.replace('min.', '').strip())

df['Duration_min'] = df['Duration'].apply(convert_duration)
```

✓ 0.0s



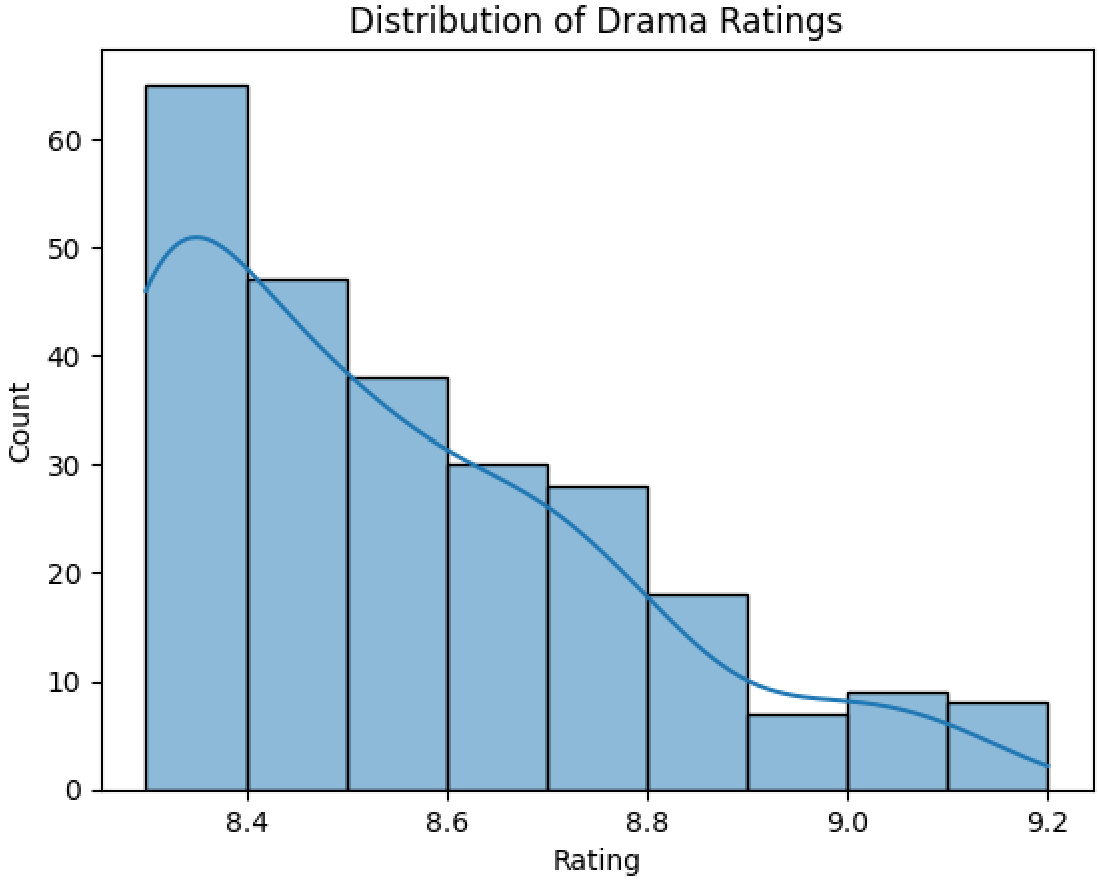
```python
#Select Relevant Columns

df_reg = df [['Year of release', 'Number of Episodes', 'Duration_min', 'Rating']]
df_reg.head()
```

✓ 0.0s

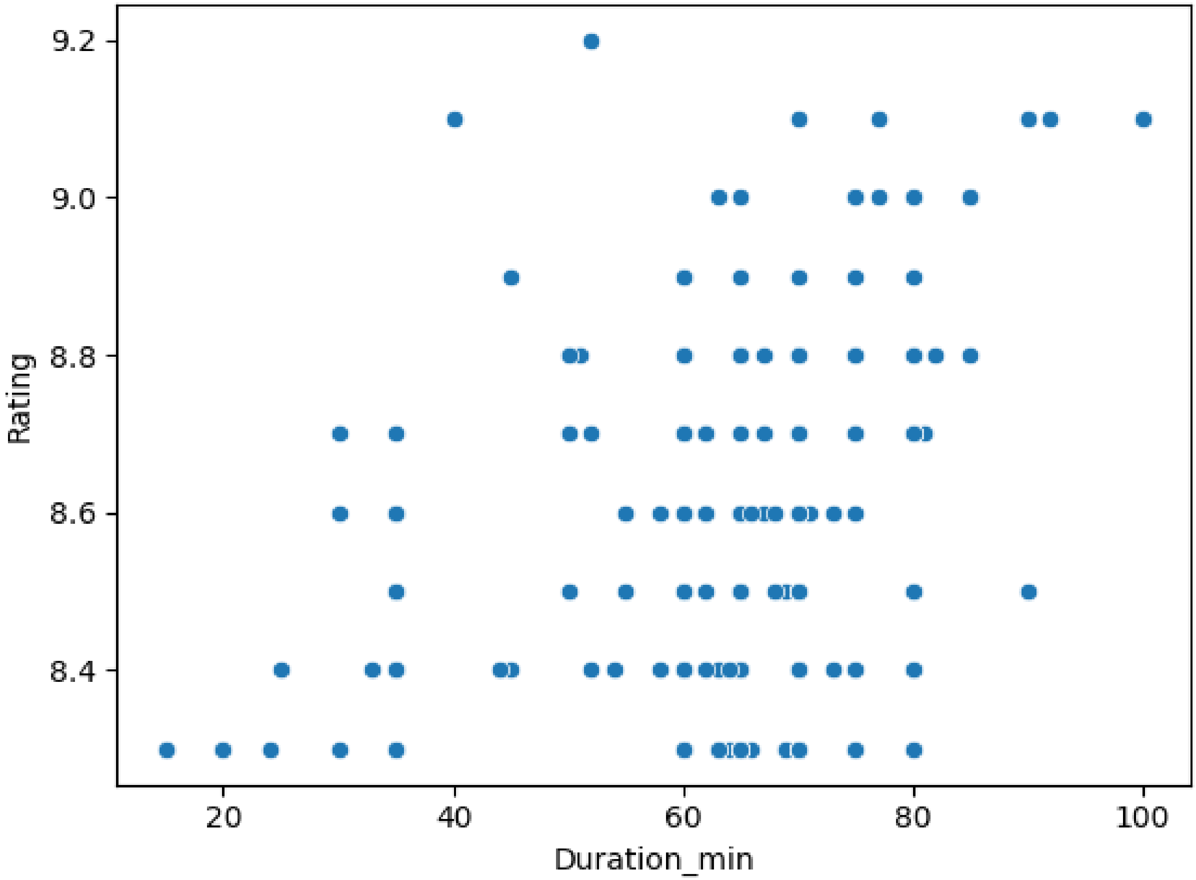
	Year of release	Number of Episodes	Duration_min	Rating
0	2021	10	52	9.2
1	2020	16	70	9.1
2	2020	12	90	9.1
3	2021	12	100	9.1
4	2018	16	77	9.1

Exploratory Data Analysis(EDA)

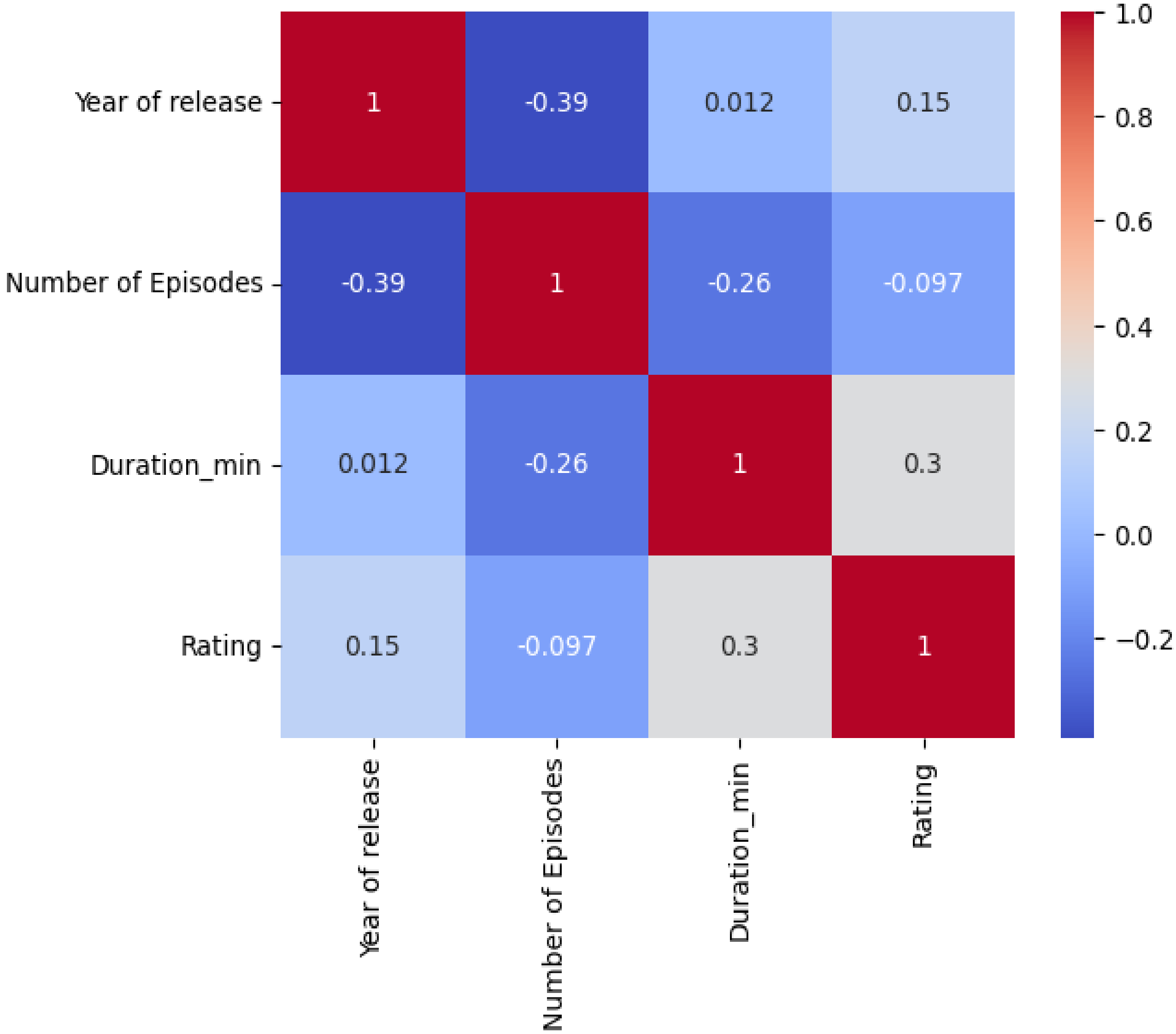
- Our EDA included histograms to analyze the distribution of ratings, scatter plots to visualize relationships between features, ratings, and a correlation heatmap to identify important variables. The analysis suggested that duration and year of release have a moderate influence on ratings, while the number of episodes has a weaker relationship.



- Scatter Plot

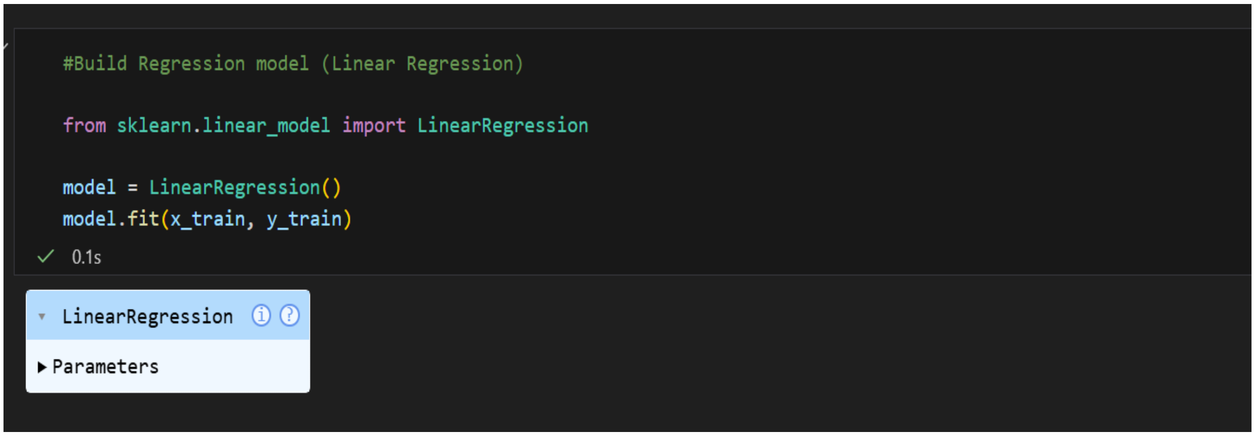


➢ Correlation Heatmap



Model Selection

- We selected Linear regression because it is a basic and interpretable model for predicting continuous values. We split the dataset into 80% training data and 20% testing data before training the model.

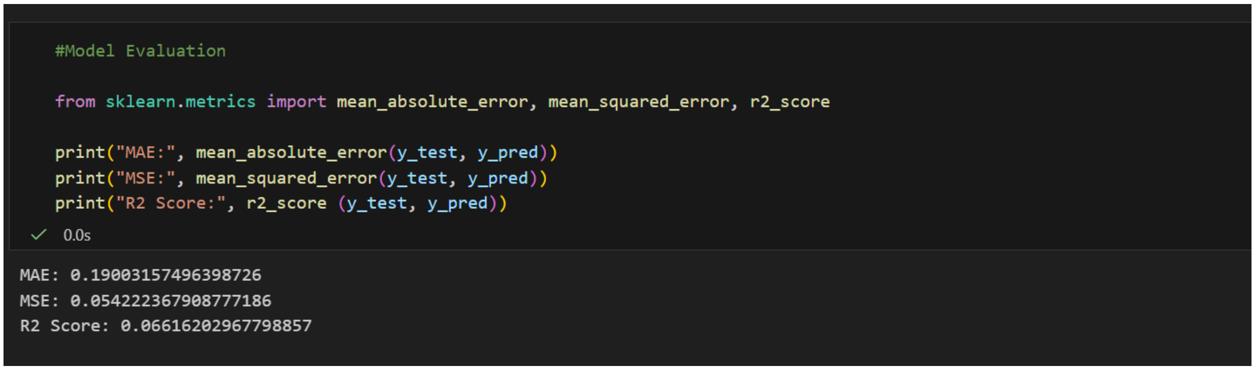


```
#Build Regression model (Linear Regression)

from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(x_train, y_train)
```

✓ 0.1s

3. Results and Interpretation

- We evaluated the regression model using Mean Absolute Error(MAE), Mean Squared Error(MSE), and the R2 Score. The result indicated that the predicted ratings were fairly close to the actual ratings. The R2 Score showed that a moderate portion of the variation in ratings was explained by the selected features. However, some variation remained unexplained, suggesting that qualitative factors like storyline and cast popularity also influence ratings.



```
#Model Evaluation

from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

print("MAE:", mean_absolute_error(y_test, y_pred))
print("MSE:", mean_squared_error(y_test, y_pred))
print("R2 Score:", r2_score (y_test, y_pred))
```

✓ 0.0s

```
MAE: 0.19003157496398726
MSE: 0.054222367908777186
R2 Score: 0.06616202967798857
```

4. Conclusion(Task 02)

- The regression task demonstrated the use of Linear regression to predict Korean Drama ratings based on numerical features. While the model provided reasonable predictions, incorporating additional qualitative variables or using more advanced regression techniques could further enhance accuracy.

Overall Conclusion

- This coursework effectively applied classification and regression techniques to two real-world datasets. The analysis followed a structured data science workflow, including preprocessing, EDA, model building, evaluation, and interpretation. The results illustrate how machine learning models can support decision-making in both healthcare and entertainment fields.