پروژه SVM

ROC:

ROC (Receiver Operating Characteristic) curve is a graphical representation that illustrates the performance of a binary classification model across all possible thresholds. It plots the true positive rate (TPR) against the false positive rate (FPR) for different threshold values.

Components of ROC Curve:

1. True Positive Rate (TPR) or Recall:
    - TPR, also known as recall or sensitivity, is the ratio of correctly predicted positive instances (spam messages correctly identified as spam) to all actual positive instances (total number of spam messages).
    - FN(false negatives are the number of spam messages that were incorrectly identified as not spam)

$$TPR = \frac{TP}{TP + FN}$$

    - TPR measures the proportion of actual positive instances that are correctly predicted by the model as positive.

2. False Positive Rate (FPR):

- FPR is the ratio of incorrectly predicted negative instances (ham messages incorrectly identified as spam) to all actual negative instances (total number of ham messages).
- TN(the number of correctly identified ham messages, or not spam)

$$FPR = \frac{FP}{FP + TN}$$

- FPR measures the proportion of actual negative instances that are incorrectly predicted by the model as positive.

Interpreting the ROC Curve:

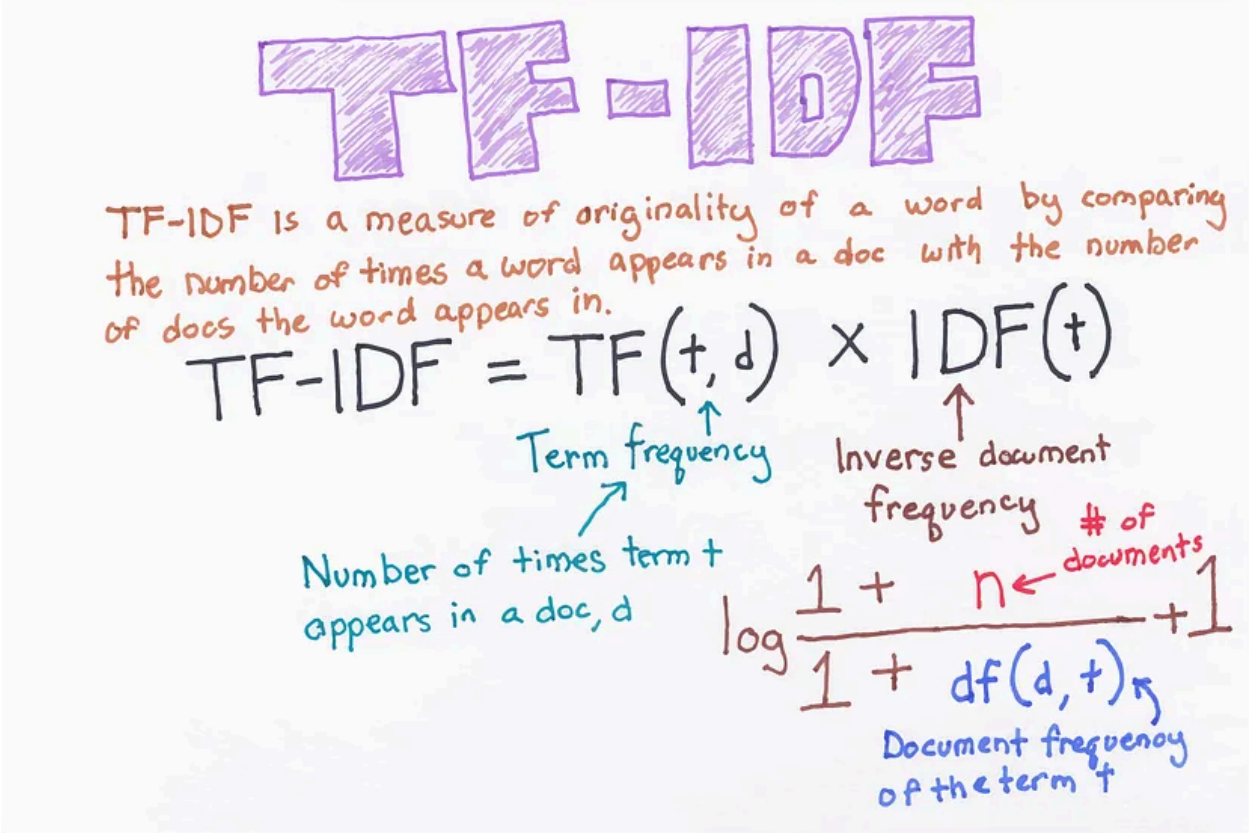
- The ROC curve plots TPR against FPR at various threshold settings.
- It helps visualize the trade-off between sensitivity (TPR) and specificity (1 - FPR).
- A diagonal line from (0,0) to (1,1) represents random guessing, where the model is as good as flipping a coin.
- The top-left corner (0,1) represents an ideal scenario where the model achieves high TPR with low FPR across all thresholds.

In this, the ROC curve is plotted to visualize the performance of the SVM model in distinguishing between 'ham' and 'spam' SMS messages based on the predicted probabilities. The area under this curve (ROC AUC) is also calculated to quantitatively evaluate the model's performance.

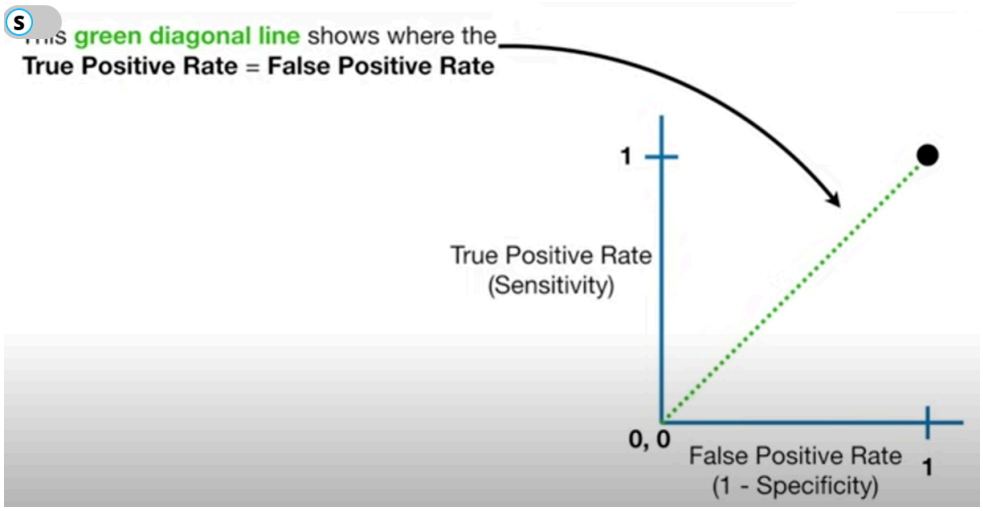
معیار خوب بودن نمودار ROC

- **خط مورب (Diagonal Line)**: اگر منحنی ROC بر روی خط مورب (یعنی خط ۴۵ درجه) قرار داشته باشد، عملکرد مدل به صورت تصادفی است و مدل نمی‌تواند بین کلاس‌ها تفاوت قائل شود.
- **گوشه بالا-چپ**: هر چه منحنی ROC بیشتر به گوشه بالا-چپ نزدیک باشد، مدل عملکرد بهتری دارد. این به این معناست که مدل دارای TPR بالا و FPR پایین است.
- **مساحت زیر منحنی (AUC - Area Under the Curve)**: یکی از معیارهای مهم برای ارزیابی عملکرد مدل، مساحت زیر منحنی ROC (AUC) است. مقدار AUC بین 0.5 و 1 متغیر است. مقدار 0.5 نشان‌دهنده عملکرد تصادفی و مقدار 1 نشان‌دهنده عملکرد ایده‌آل است.

TF-IDF:



کد این بخش داده‌های متنی را به ماتریس ویژگی‌های TF-IDF تبدیل می‌کند که می‌تواند به عنوان ورودی به مدل‌های یادگیری ماشین (مانند SVM) داده شود. این روش به مدل کمک می‌کند تا با استفاده از ویژگی‌های عددی مهم، داده‌های متنی را دسته‌بندی کند.



Q1)

معیارهای ارزیابی:

گزارش دسته‌بندی (classification_report) یک تابع ارائه شده توسط sklearn (scikit-learn) است که یک گزارش جامع از عملکرد مدل دسته‌بندی تولید می‌کند. این گزارش چندین معیار مهم مانند دقت (precision)، بازیابی (recall)، نمره F1 (F1-score)، و حمایت (support) برای هر کلاس را در یک قالب جدولی نمایش می‌دهد.

سطرها (کلاس‌ها):

دقت (ستون 1):

دقت نسبت پیش‌بینی‌های مثبت درست به کل پیش‌بینی‌های مثبت انجام شده توسط مدل است.

فرمول: $\text{Precision} = \frac{TP}{TP+FP}$

دقت بالا نشان می‌دهد که وقتی مدل نتیجه‌ای مثبت پیش‌بینی می‌کند، معمولاً صحیح است.

بازیابی یا همان sensitivity list:

بازیابی، همچنین به عنوان حساسیت یا نرخ مثبت درست شناخته می‌شود، نسبت پیش‌بینی‌های مثبت درست به کل مثبت‌های واقعی در داده‌ها را اندازه‌گیری می‌کند.

فرمول: $\text{Recall} = \frac{TP}{TP+FN}$

بازیابی بالا نشان می‌دهد که مدل اکثر مثبت‌های واقعی را به درستی تشخیص می‌دهد.

نمره F1 (ستون 3):

نمره F1 میانگین هارمونیک دقت و بازیابی است. این معیار تعادل بین دقت و بازیابی را فراهم می‌کند.

فرمول: $\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

نمره F1 بهترین مقدار خود را در 1 (دقت و بازیابی کامل) و بدترین مقدار خود را در 0 می‌گیرد.

حمایت (ستون 4):

حمایت تعداد وقوع‌های واقعی هر کلاس در مجموعه داده مشخص است. این معیار تعداد وقوع‌های هر کلاس

در

true(y را اندازه‌گیری می‌کند.

ستون‌ها (معیارهای کلی):

دقت، بازیابی، نمره F1، حمایت:

این معیارها برای هر کلاس به صورت جداگانه محاسبه می‌شوند (در این مورد، برای '0' (ham) و '1' (spam)).

هر معیار در ستون مربوطه برای هر کلاس نمایش داده می‌شود.

آخرین سطر گزارش (معمولاً با عنوان 'macro avg') میانگین معیارها در همه کلاس‌ها را ارائه می‌دهد، که به تعداد موارد واقعی برای هر کلاس وزنی داده می‌شود.

تفسیر:

دقت: مدل چقدر دقیق است وقتی که یک کلاس خاص را پیش‌بینی می‌کند.

بازیابی: مدل چقدر خوب موارد یک کلاس خاص را تشخیص می‌دهد.

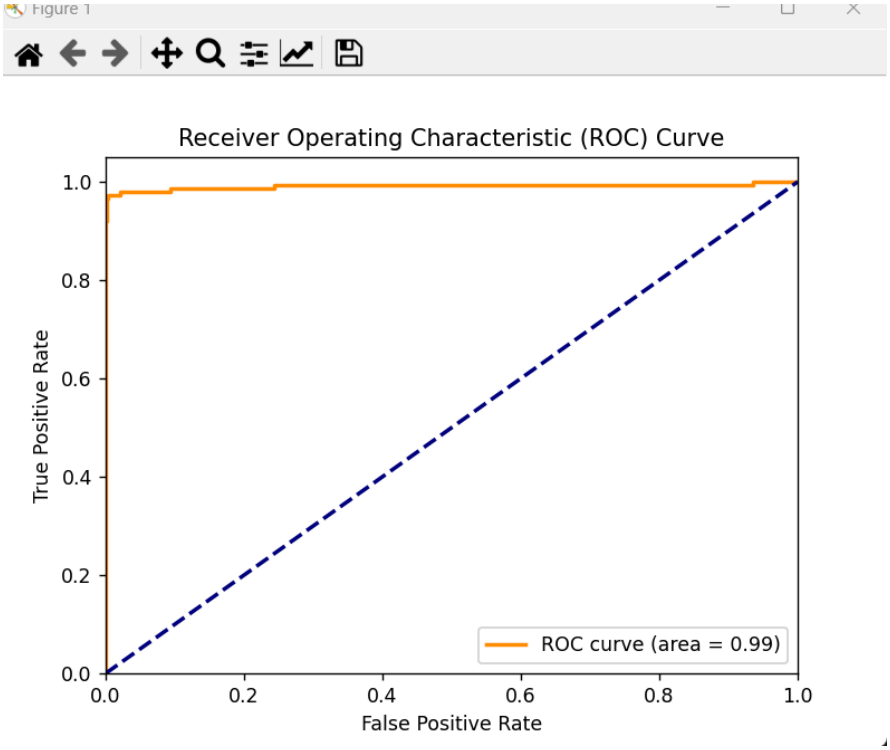
نمره F1: میانگین هارمونیک دقت و بازیابی، که یک معیار واحد برای ارزیابی مدل فراهم می‌کند.

حمایت: تعداد وقوع‌های هر کلاس در مجموعه داده.

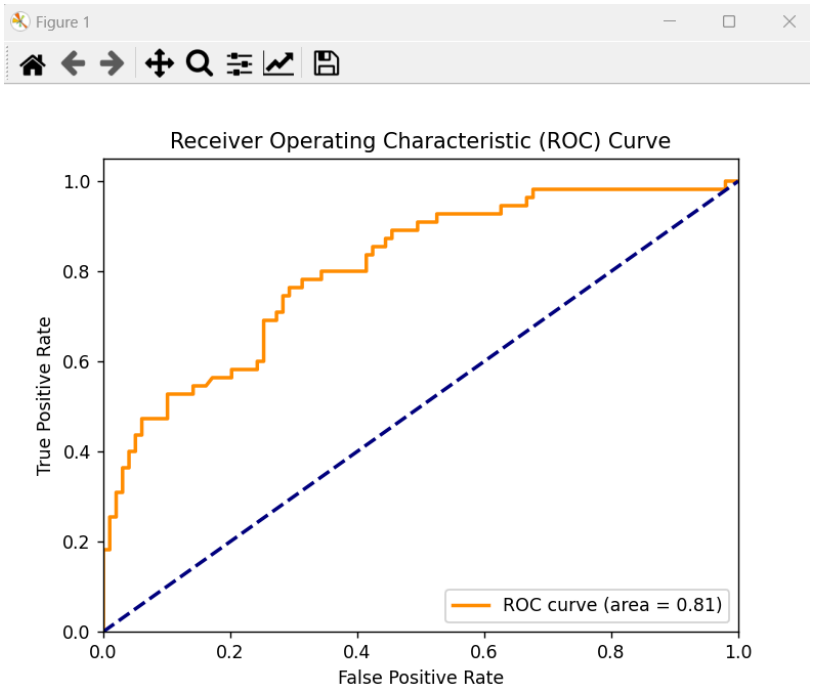
```
Accuracy: 0.9919282511210762
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       966
           1       0.99      0.95      0.97       149

    accuracy                           0.99      1115
   macro avg       0.99      0.97      0.98      1115
weighted avg       0.99      0.99      0.99      1115
```

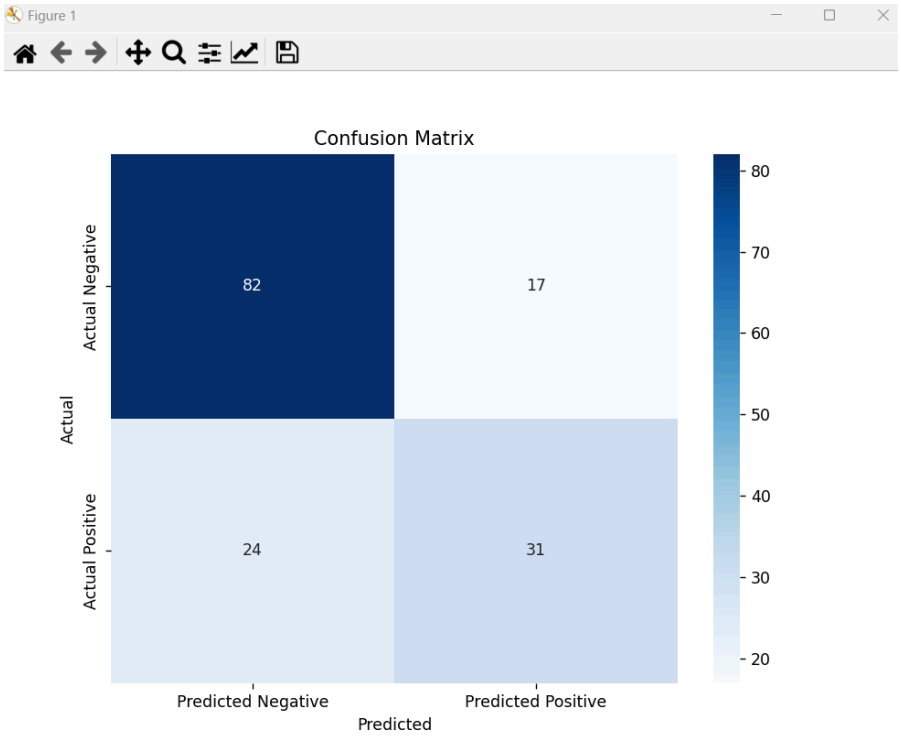
```
True Positives (TP): 141
False Negatives (FN): 8
False Positives (FP): 1
True Negatives (TN): 965
                  Actual Positive (Spam)  Actual Negative (Ham)
Predicted Positive                    141                      1
Predicted Negative                      8                    965
```

یاسمن توکلی: 98522175



Q2)






```
Desktop\TenthSemester\AI\SVM\q2\q2.py "
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigreeFunction  Age  Outcome
0            6      148             72             35        0  33.6                     0.627   50        1
1            1       85             66             29        0  26.6                     0.351   31        0
2            8      183             64              0        0  23.3                     0.672   32        1
3            1       89             66             23       94  28.1                     0.167   21        0
4            0      137             40             35      168  43.1                     2.288   33        1
Accuracy: 0.7337662337662337
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.83      0.80        99
           1       0.65      0.56      0.60        55

    accuracy                           0.73       154
   macro avg       0.71      0.70      0.70       154
weighted avg       0.73      0.73      0.73       154

True Positives (TP): 31
False Negatives (FN): 24
False Positives (FP): 17
True Negatives (TN): 82
                  Actual Positive  Actual Negative
Predicted Positive              31              17
Predicted Negative              24              82
Cross-validation scores: [0.77235772 0.80487805 0.73170732 0.75609756 0.77868852]
Mean cross-validation score: 0.7687458349993335
```

کد این بخش برای استانداردسازی داده‌های عددی مورد استفاده قرار می‌گیرد. استانداردسازی یا نرمال‌سازی یکی از مراحل پیش‌پردازش داده‌ها است که باعث می‌شود ویژگی‌های داده‌ها دارای مقیاس و توزیع مشابهی باشند.

تسریع در همگرایی: در الگوریتم‌های یادگیری ماشین که مبتنی بر بهینه‌سازی هستند (مانند رگرسیون خطی یا شبکه‌های عصبی)، استانداردسازی داده‌ها می‌تواند به تسریع در فرآیند همگرایی کمک کند.

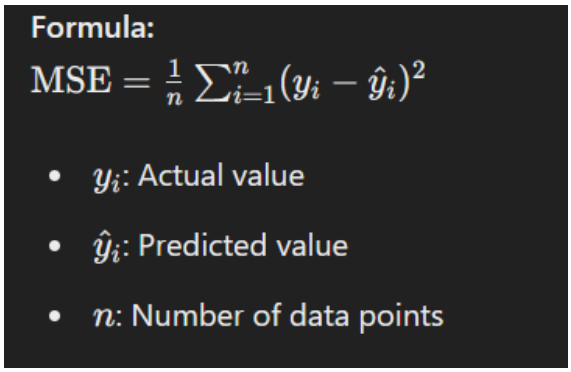
بهبود عملکرد مدل: برخی الگوریتم‌ها مانند SVM و KNN به مقیاس داده‌ها حساس هستند و استانداردسازی می‌تواند به بهبود عملکرد آن‌ها کمک کند.

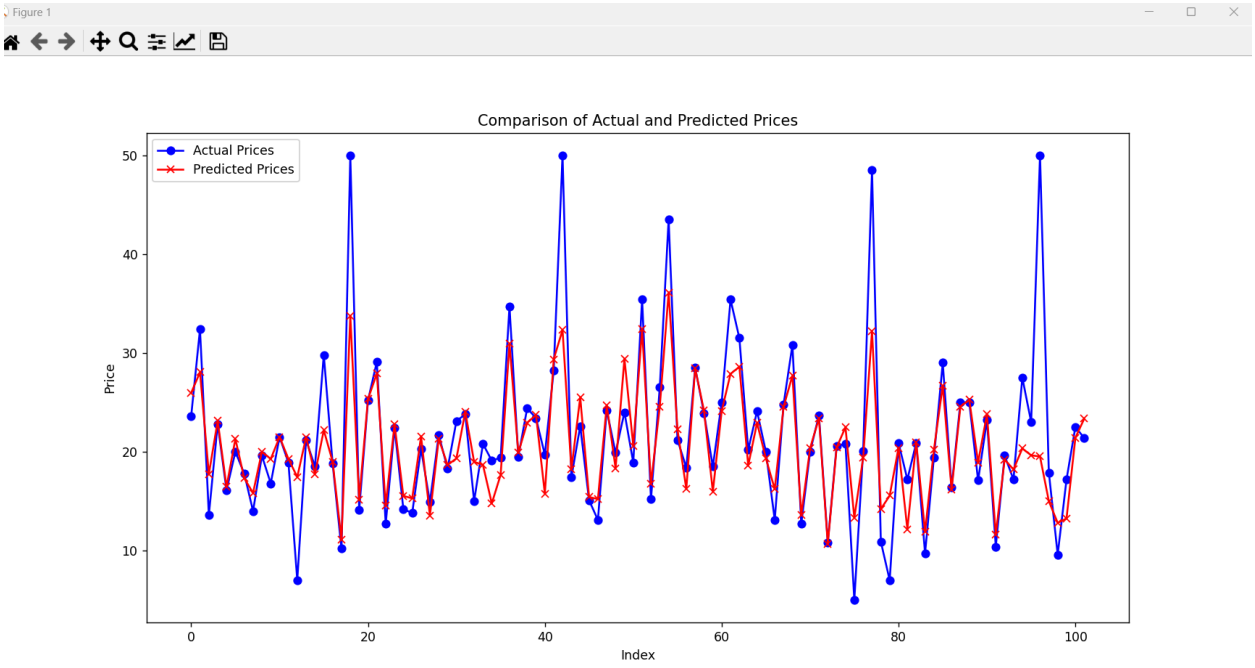
پیشگیری از اولویت دادن به ویژگی‌های با مقیاس بزرگتر: ویژگی‌هایی که مقیاس بزرگتری دارند ممکن است به‌طور غیرمنصفانه‌ای تاثیر بیشتری در مدل داشته باشند. استانداردسازی این مشکل را رفع می‌کند.

Q3)

Mean Squared Error (MSE):

میانگین مربعات خطا (MSE) یک معیار رایج است که برای ارزیابی عملکرد یک مدل رگرسیون استفاده می شود. میانگین مربعات خطاها را اندازه گیری می کند، جایی که خطا تفاوت بین مقدار واقعی و مقدار پیش بینی شده است.






```
      crim    zn  indus  chas    nox     rm   age     dis  rad  tax  ptratio       b  lstat  medv
0  0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296     15.3  396.90   4.98  24.0
1  0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242     17.8  396.90   9.14  21.6
2  0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242     17.8  392.83   4.03  34.7
3  0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222     18.7  394.63   2.94  33.4
4  0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222     18.7  396.90   5.33  36.2
Mean Squared Error: 25.668539678396044
R-squared: 0.6499766059760035
Cross-validation MSE scores: [26.03579636 39.61659168 38.2396779  32.19280936 25.70582526]
Mean cross-validation MSE score: 32.3581401119639
```

R-squared:

R-squared (ضریب تعیین) معیاری آماری است که نسبت واریانس متغیر وابسته را که از روی متغیرهای مستقل قابل پیش بینی است، نشان می دهد.

محدوده: مقادیر مربع R از 0 تا 1 متغیر است.

0: مدل هیچ یک از تغییرپذیری داده های پاسخ را حول میانگین آن توضیح نمی دهد.

1: مدل تمام تغییرپذیری داده های پاسخ را حول میانگین آن توضیح می دهد.

مقادیر R-squared بالاتر به طور کلی نشان دهنده تناسب بهتر مدل با داده ها است.

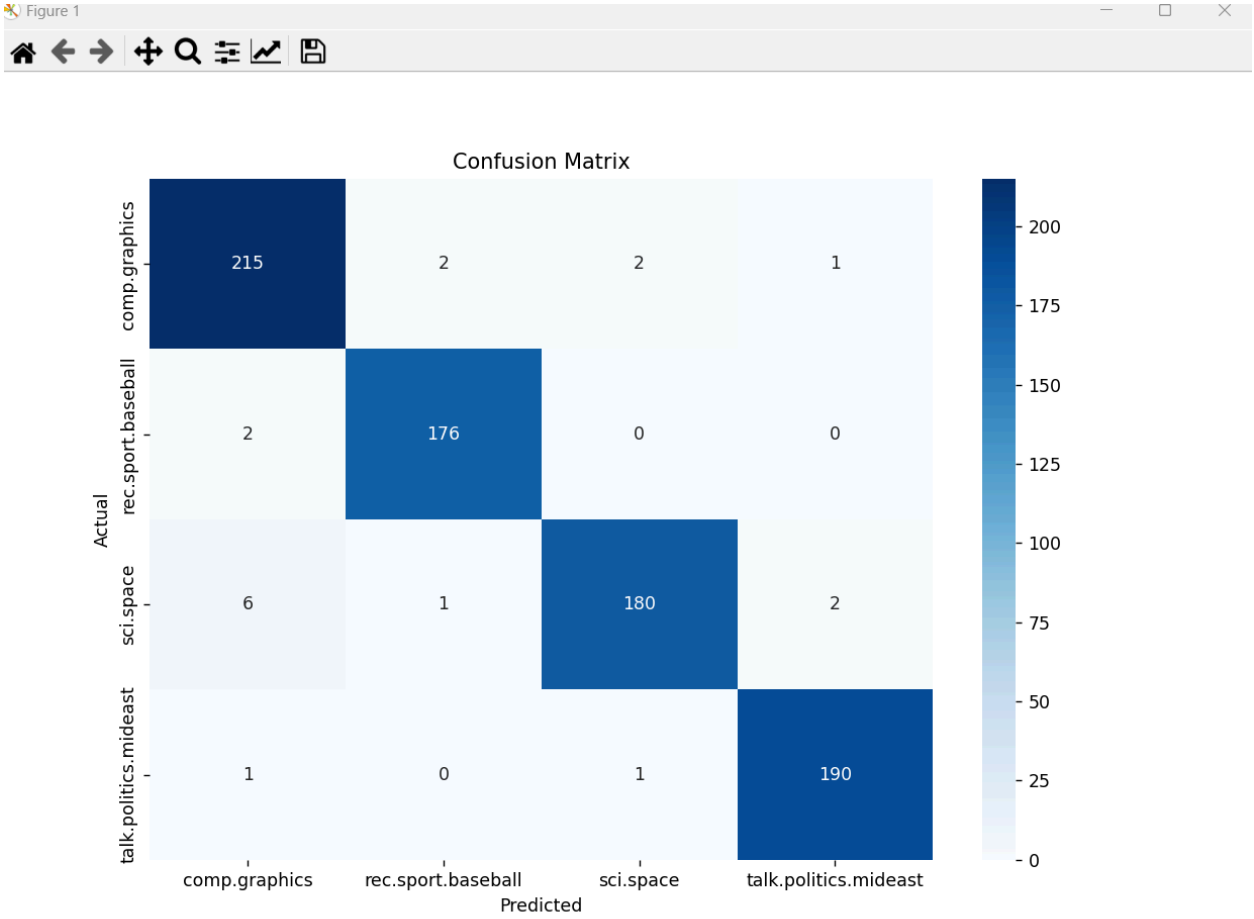
با این حال، R-squared بالا به تنهایی به معنای خوب بودن مدل نیست. بررسی سایر معیارها مانند MSE، امتیازهای اعتبارسنجی متقابل و تطبیق بیش از حد بالقوه ضروری است.

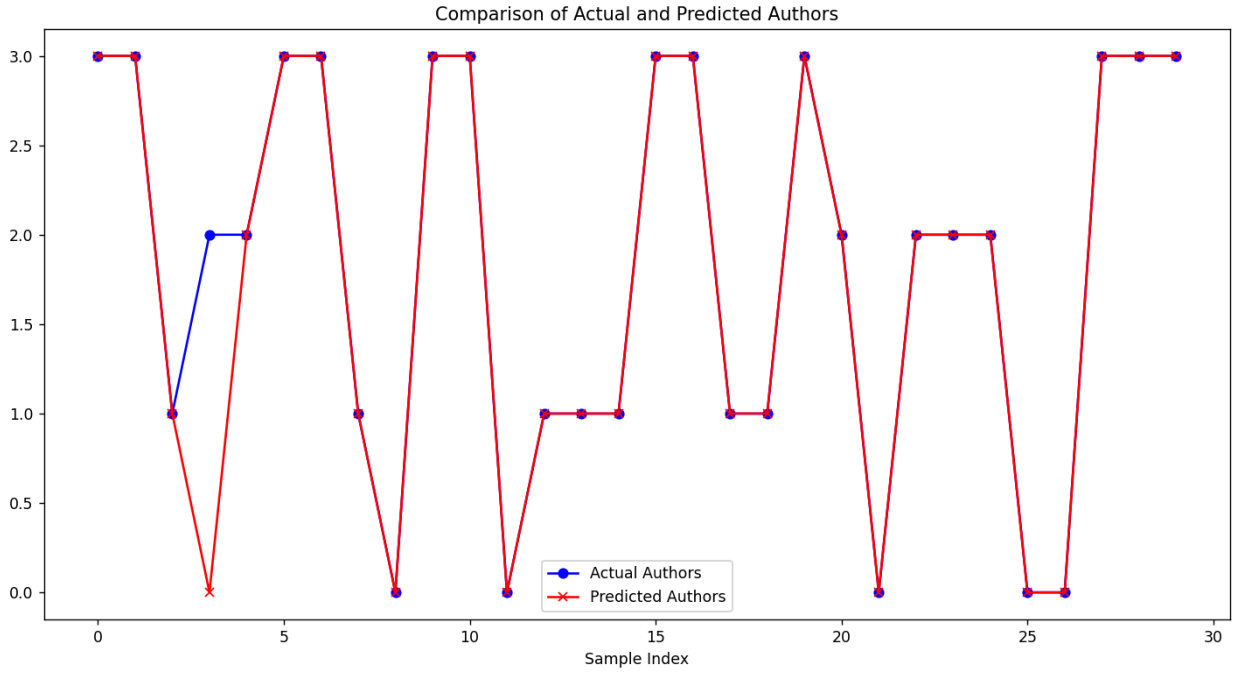
```
# Cross-validation scores
scores = cross_val_score(model, X_train_scaled, y_train,
                         cv=5, scoring='neg_mean_squared_error')
```

cv=5: تعداد تاخوردگی‌های (folds) اعتبارسنجی متقابل. در اینجا از 5-fold cross-validation استفاده می‌شود، یعنی داده‌ها به 5 بخش تقسیم می‌شوند و مدل 5 بار آموزش داده می‌شود، هر بار یک بخش به عنوان داده‌های اعتبارسنجی و چهار بخش دیگر به عنوان داده‌های آموزشی استفاده می‌شوند.

Q4)

یاسمن توکلی: 98522175






```
Accuracy: 0.9768934531450578
Classification Report:
                      precision    recall  f1-score   support

       comp.graphics       0.96      0.98      0.97       220
   rec.sport.baseball       0.98      0.99      0.99       178
           sci.space       0.98      0.95      0.97       189
talk.politics.mideast       0.98      0.99      0.99       192

            accuracy                           0.98       779
           macro avg       0.98      0.98      0.98       779
        weighted avg       0.98      0.98      0.98       779
```

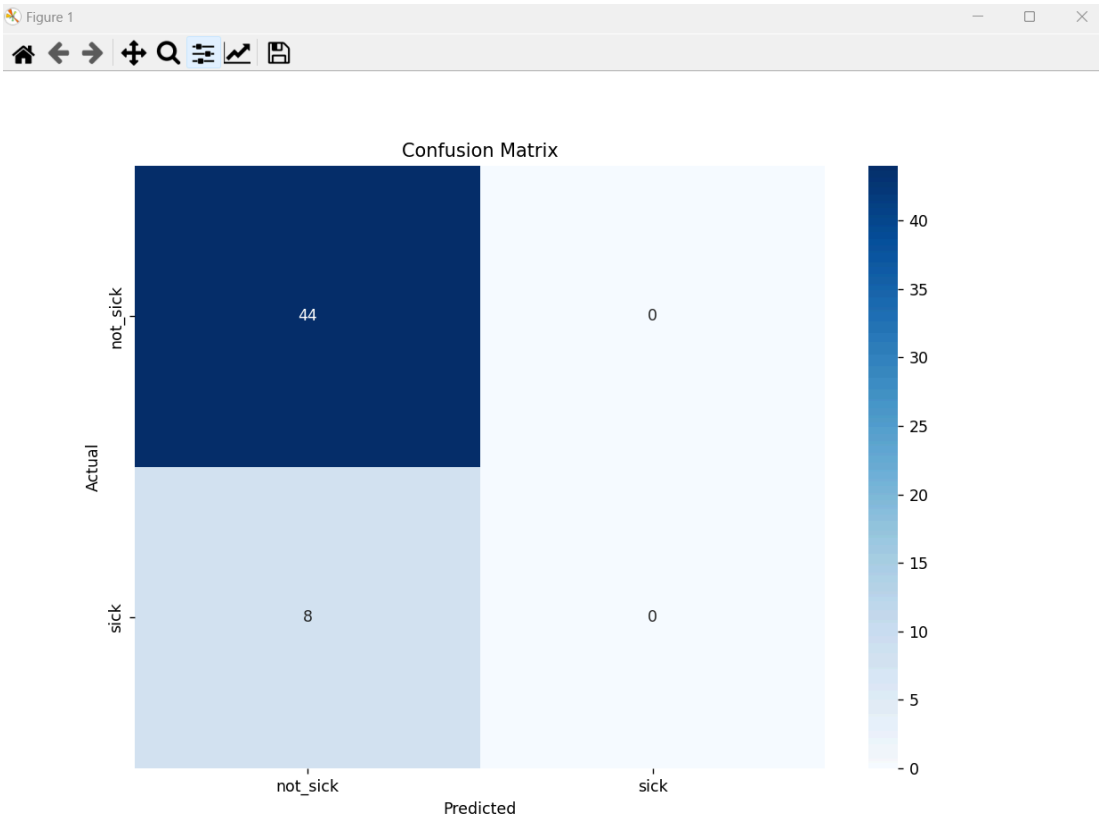
Q5)

برای این قسمت ما از این دیتابیس استفاده میکنیم که پارامترهای مختلفی مثل جنسیت، فشارخون و غیره برای افراد مختلف دارد.

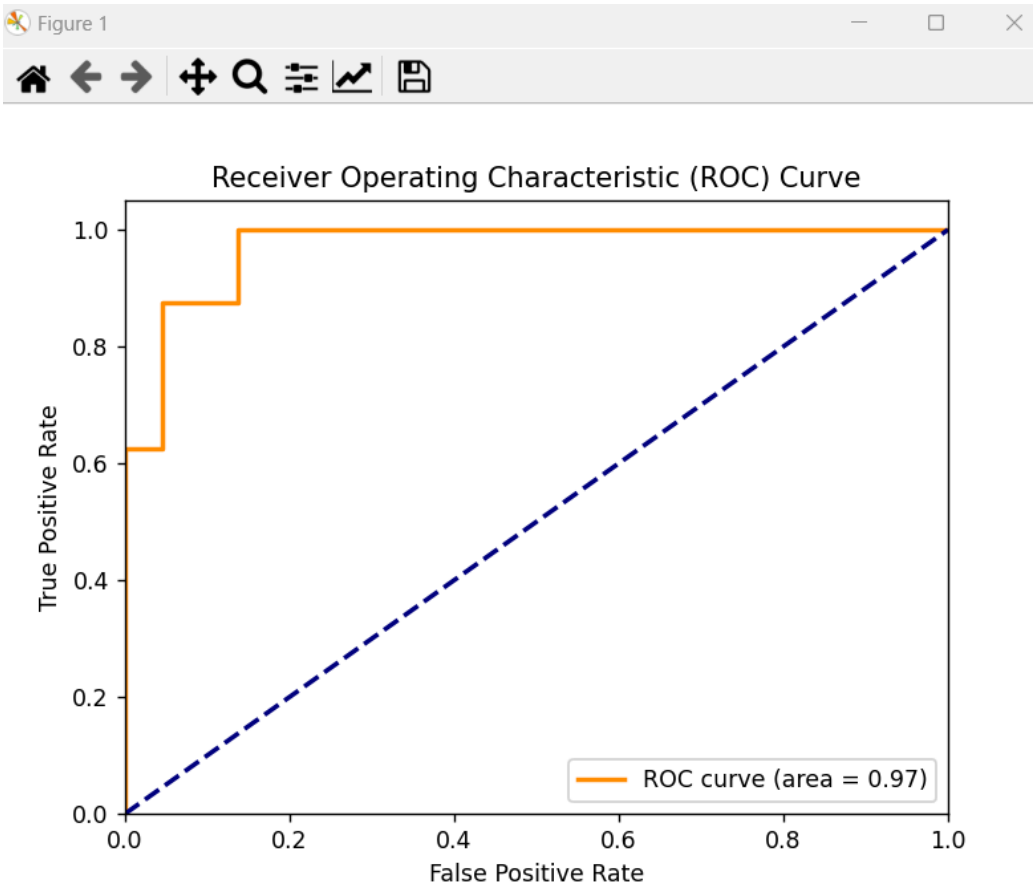
مجموعه داده NHANES شامل طیف وسیعی از متغیرهای مرتبط با سلامتی است، اما پارامتر واحدی ندارد که به طور مستقیم نشان دهد که آیا فرد بیمار است یا خیر. با این حال، می‌توانید از پارامترهای خاصی به عنوان پروکسی برای طبقه‌بندی افراد به عنوان بیمار یا غیر بر اساس شرایط یا معیارهای سلامت خاص استفاده کنید.

به عنوان مثال، می توانید از خوانش فشار خون، سطح کلسترول یا سایر شاخص های سلامتی برای ایجاد یک طبقه بندی باینری استفاده کنیم.

فرض کنیم می خواهیم افراد را بر اساس فشار خون بالا طبقه بندی کنیم. ما می توانیم از خوانش فشار خون سیستولیک (BPXSY1) و دیاستولیک (BPXDI1) برای طبقه بندی افراد به عنوان پرفشاری خون (بیمار) یا غیر پرفشاری خون (غیر بیمار) استفاده کنیم. و برهمین اساس ماتریس درهم ریختگی و نمودار ROC را بکشیم.
بقیه قسمت های کد نیز مشابه قبل می باشد.



یاسمن توکلی: 98522175



	precision	recall	f1-score	support
not_sick	0.85	1.00	0.92	44
sick	0.00	0.00	0.00	8
accuracy			0.85	52
macro avg	0.42	0.50	0.46	52
weighted avg	0.72	0.85	0.78	52

Confusion Matrix:
[[44 0]
[8 0]]

تفاوت کرنل خطی و کرنل RBF در SVM:

1. کرنل خطی (Linear Kernel):

- تعریف: کرنل خطی ساده‌ترین نوع کرنل است که برای داده‌هایی استفاده می‌شود که به‌طور خطی قابل تفکیک هستند.
- مزایا:
  - سادگی و سرعت بالا.
  - مناسب برای داده‌های با ابعاد بالا (بسیار زیاد ویژگی‌ها).
  - قابل تفسیر بودن مدل.
- معایب:
  - ناتوانی در مدل‌سازی روابط غیرخطی بین داده‌ها.
- مناسب برای:
  - داده‌هایی که مرز تفکیک آنها خطی است یا تقریباً خطی است.
  - مثال‌ها: دیتاست‌های با ویژگی‌های زیادی که به‌طور طبیعی خطی قابل تفکیک هستند.

2. کرنل RBF (Radial Basis Function):

- تعریف: کرنل RBF یا گوسی برای داده‌هایی استفاده می‌شود که به‌طور غیرخطی قابل تفکیک هستند.
- مزایا:
  - توانایی مدل‌سازی روابط پیچیده و غیرخطی.
  - انعطاف‌پذیری بالا در تشخیص مرزهای پیچیده بین کلاس‌ها.
- معایب:
  - زمان محاسباتی بیشتر به دلیل پیچیدگی بالاتر.
  - نیاز به تنظیم دقیق پارامترها (به‌خصوص γ\gammaγ و CCC).
- مناسب برای:
  - داده‌هایی که مرز تفکیک آنها پیچیده و غیرخطی است.
  - مثال‌ها: دیتاست‌های پیچیده با الگوهای غیرخطی مانند تصاویر، داده‌های بیولوژیکی و غیره.

یاسمن توکلی: 98522175

انتخاب کرنل مناسب:

- کرنل خطی:
  - زمانی که داده‌ها به‌طور خطی قابل تفکیک هستند یا تقریباً خطی هستند.
  - زمانی که تعداد ویژگی‌ها بسیار زیاد است و مرز تفکیک ساده است.
- کرنل RBF:
  - زمانی که داده‌ها به‌طور پیچیده و غیرخطی قابل تفکیک هستند.
  - زمانی که الگوهای پیچیده‌تری در داده‌ها وجود دارد که نیاز به مدل‌سازی روابط غیرخطی دارد.

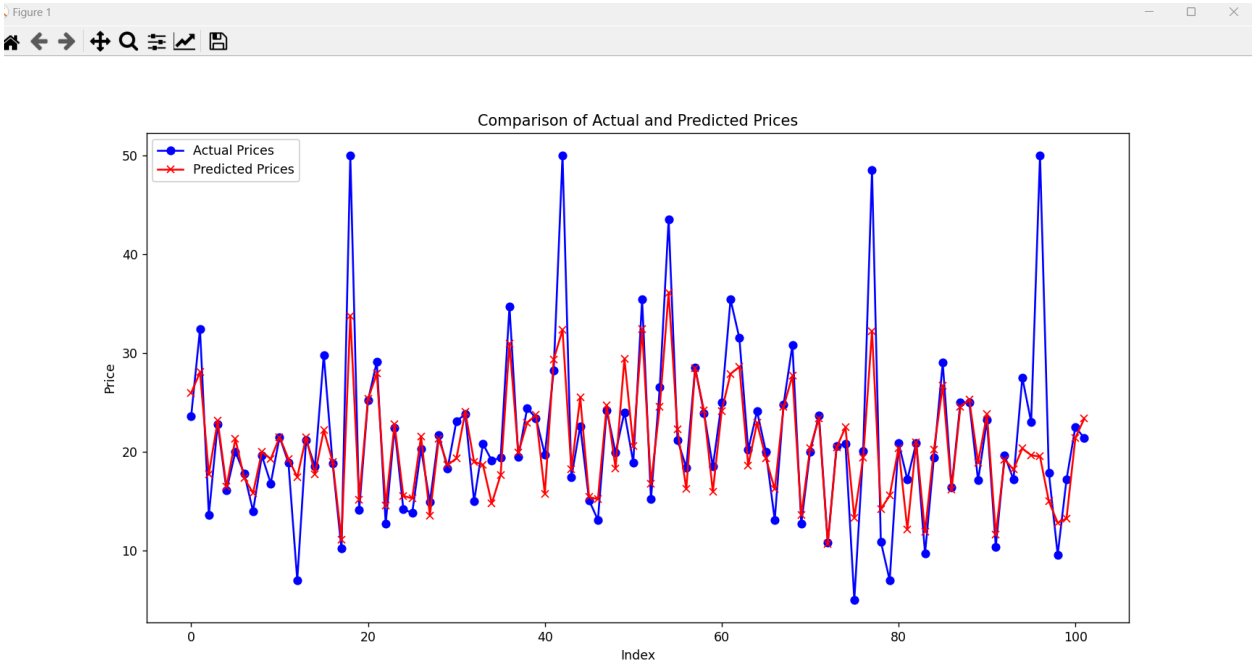
مثال‌ها:

- دیتاست‌های ساده و خطی:
  - دیتاست‌هایی مانند Iris که مرز تفکیک بین کلاس‌ها نسبتاً خطی است.
  - دیتاست‌هایی با تعداد ویژگی‌های زیاد (ابعاد بالا).
- دیتاست‌های پیچیده و غیرخطی:
  - دیتاست‌های تصویری مانند MNIST که شامل داده‌های پیچیده و غیرخطی هستند.
  - دیتاست‌های بیولوژیکی مانند داده‌های ژنتیکی یا پزشکی که شامل روابط پیچیده هستند.

برای مثال دقت سوال 3 با کرنل خطی کمی پایین تر از کرنل RBF است:

کرنل RBF:

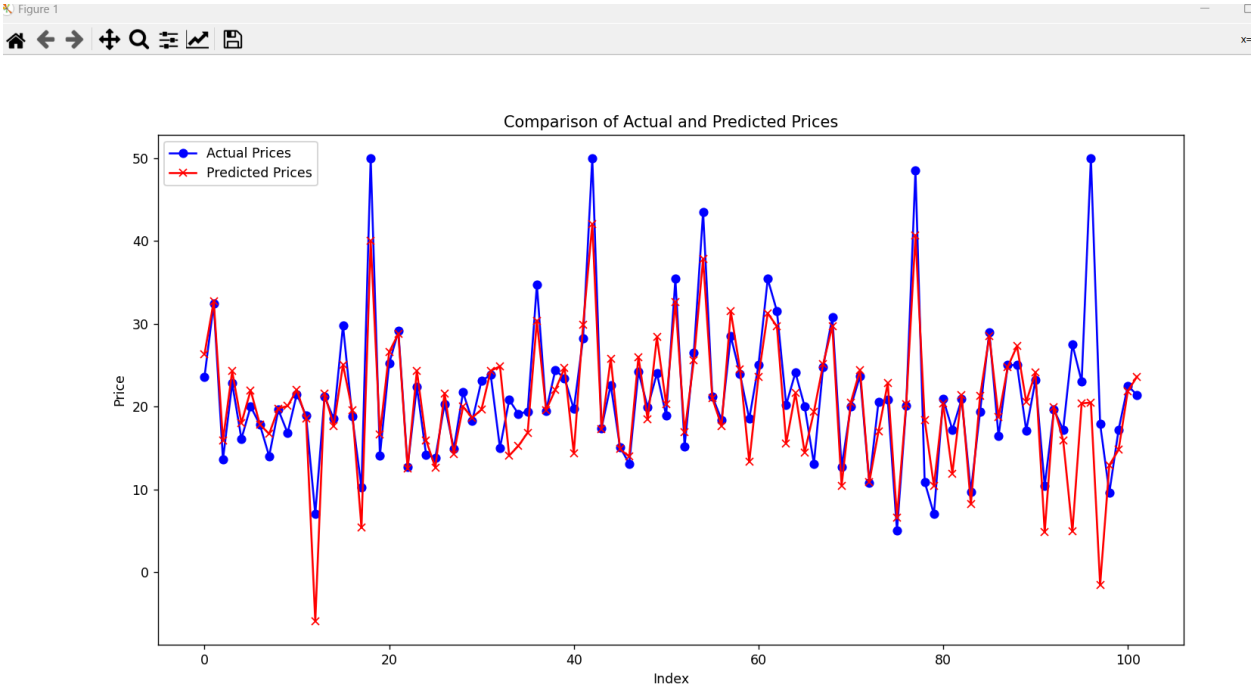
یاسمن توکلی: 98522175




```
      crim    zn  indus  chas    nox     rm   age     dis  rad  tax  ptratio       b  lstat  medv
0  0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296     15.3  396.90   4.98  24.0
1  0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242     17.8  396.90   9.14  21.6
2  0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242     17.8  392.83   4.03  34.7
3  0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222     18.7  394.63   2.94  33.4
4  0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222     18.7  396.90   5.33  36.2
Mean Squared Error: 25.668539678396044
R-squared: 0.6499766059760035
Cross-validation MSE scores: [26.03579636 39.61659168 38.2396779  32.19280936 25.70582526]
Mean cross-validation MSE score: 32.3581401119639
```

با کرنل خطی:

یاسمن توکلی: 98522175




```
Mean Squared Error: 28.91852267161847
R-squared: 0.6056589279132574
Cross-validation MSE scores: [17.94295415 37.46500652 31.94230479 20.78837304 23.77539099]
Mean cross-validation MSE score: 26.38280589816627
```