**Introduction**:

This study investigates the potential correlation between IMDb user ratings and the popularity of films on Ekşi Sözlük, a widely used Turkish social platform. Two main data sources were used for this analysis. The first one was obtained from (https://datasets.imdbws.com), IMDb's official datasets. Particularly, `title.basics.tsv.gz` was used to extract fundamental information such as film titles, release years, and types, while `title.ratings.tsv.gz` provided user ratings. The second was collected (https://eksisozluk.com), which includes entries that discussions related to films from the users. The number of entries for each film on Ekşi Sözlük was collected to represent the level of public interest. By combining these two datasets, this study investigates whether IMDb ratings are associated with a film’s popularity, as measured by entry counts on Ekşi Sözlük.

**EDA Results and Interpretations**

The exploratory data analysis was conducted across four distinct time periods: 2008–2011, 2012–2015, 2016–2019, and 2020–2024. In each period, the relationship between IMDb scores and Ekşi Sözlük entry counts was visualized and analyzed using scatter plots and linear regression trend lines. Pearson correlation coefficients and p-values were also computed to quantify the strength and statistical significance of the relationships.

**🔹 2008–2011**

* **Pearson Correlation**: 0.346
* **P-value**: 0.0489
* metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

  Yapay zeka tarafından oluşturulan içerik yanlış olabilir.This period shows a moderate positive correlation between the number of entries and IMDb scores. The p-value is just below the 0.05 threshold, indicating statistical significance. This suggests that films with higher IMDb scores tended to attract more release attention on Ekşi Sözlük during this period.

**🔹2012–2015**

* **Pearson Correlation**: 0.313
* **P-value**: 0.0626

metin, çizgi, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

* A moderate positive correlation is again observed, but the p-value is slightly above 0.05, indicating marginal non-significance. This suggests a possible trend, though not statistically confirmed, between higher IMDb scores and increased entry counts.

**🔹 2016–2019**

* **Pearson Correlation**: 0.335
* **P-value**: 0.0878

çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

* The correlation remains positive but relatively weak and statistically insignificant. The dispersion of points indicates that other factors may have played a larger role in entry frequency during this period.

**🔹 2020–2024**

* **Pearson Correlation**: 0.152
* **P-value**: 0.4135

metin, çizgi, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

* This period exhibits the weakest correlation among all four. The high p-value indicates that there is no statistically significant relationship between IMDb scores and popularity on Ekşi Sözlük during recent years. Increased social media fragmentation or changing audience behavior might explain this pattern.

**Hypothesis Testing**

To assess whether there is a statistically significant link between IMDb scores entry counts on Ekşi Sözlük, Pearson correlation analysis was performed for each time period. The results indicate a moderate positive correlation for the years 2008–2011 (r = 0.346, p = 0.0489), which is statistically significant at the 0.05 level. However, in the later periods, although weak to moderate correlations were observed, the p-values exceeded 0.05, suggesting no significant linear relationship in those cases. Finally, the hypothesis that higher IMDb scores are associated with greater entry counts is only supported for the 2008–2011 interval.

**Machine Learning Methods**

In this study, two machine learning models were applied to investigate whether Ekşi Sözlük entry counts can be used to predict IMDb scores: **K-Nearest Neighbors (KNN) Regressor** and **Random Forest Regressor**. The analysis was conducted for four different periods: 2008–2011, 2012–2015, 2016–2019, and 2020–2024. Each model was tested with different parameter settings — for KNN, various values of k (1, 3, and 5) were used, while for Random Forest, combinations of n\_estimators (number of trees) and max\_depth (tree depth) were evaluated. The performance of each model was assessed using two standard regression metrics: **Mean Squared Error (MSE)** and **R² Score**.

For the 2008–2011 period, both models exhibited poor performance with negative R² values. This indicates that the models failed to capture a meaningful relationship between entry counts and IMDb ratings. Although Random Forest with 100 trees and a maximum depth of 10 provided the best result, its R² value remained below zero (−0.2409).

The 2012–2015 dataset showed slight improvement. KNN with k=3 achieved an R² of 0.195, outperforming all Random Forest configurations, which reached up to approximately 0.1369.

The best results were obtained in the 2016–2019 period. Here, KNN with k=1 yielded the highest R² value (0.6216), suggesting a relatively strong nonlinear relationship. Random Forest also performed reasonably well in this period, with its best configuration achieving an R² of 0.3762.

However, in the 2020–2024 dataset, both models failed to produce satisfactory predictions. All R² values were negative again, with the best Random Forest result reaching only −0.7144. This may indicate that entry counts during this period were not a reliable indicator of film ratings, potentially due to irregular user behavior or data inconsistencies caused by external factors such as the COVID-19 pandemic.

In conclusion, **KNN Regressor performed better in years where nonlinear relationships were more dominant**, while **Random Forest provided more stable results across different configurations**. The analysis demonstrates that the effectiveness of machine learning models in predicting IMDb scores from Ekşi Sözlük data is highly dependent on the characteristics and quality of the dataset for each specific period.

**Challenges Encountered**

During the data collection process, several challenges were encountered, particularly in relation to title matching between IMDb and Ekşi Sözlük. For instance, the film *Roma* by Alfonso Cuarón is listed on Ekşi Sözlük as *"Roma (Alfonso Cuarón filmi)"*, which differs from its standard IMDb title. This inconsistency required manual review and adjustment to ensure accurate mapping of titles across platforms.

Additionally, certain films had to be excluded due to ambiguity in keyword searches. For example, *Her* (2013) was not included in the dataset because the term “her” generated a wide range of irrelevant results, making it impractical to isolate entries specific to the film.

Another major challenge was the process of extracting data from Ekşi Sözlük. The platform does not offer an open API and has structural limitations that make automated scraping difficult. To overcome this, we had to implement browser automation tools and install additional extensions to simulate user behavior and retrieve entry counts accurately.

**General Insight**

Across all four periods, a weak-to-moderate positive correlation is observed between IMDb ratings and entry counts. However, the strength and significance of this relationship declines over time, suggesting that IMDb scores have become less predictive of Ekşi Sözlük engagement in recent years.