Computational Intelligence and Decision Making

Lab Work - 1

Ravinthiran Partheepan

Master's in Artificial Intelligence

Semester: 1 / Year: 1

Kaunas University of Technology

About the Task

- Problem: based on the given data of historical real estate transactions create the decision-making model (DMM) which aims to predict prices of new real estate objects.

  - Project workflow:

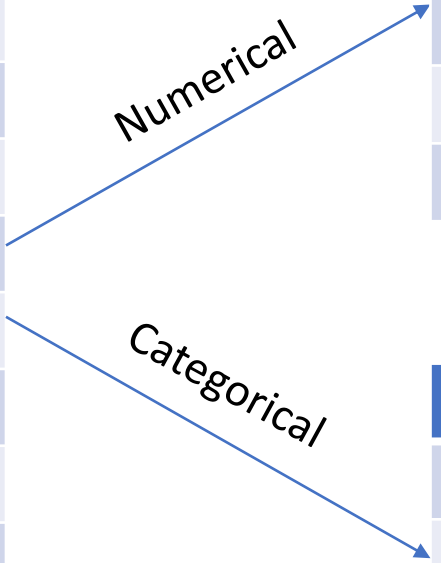
    - P1. Perform given data analysis and preprocessing
    - P2. Implement K-Nearest Neighbors (KNN), Decision tree (DT), and random forest (RF) algorithms (You cannot use library functions for these algorithms)
    - P3. Use implemented algorithms to create DMM for the given problem and evaluate the results.
    - P4. Use “scikit-learn” (or other) library functions for the same algorithms and evaluate the results.
    - P5. Write conclusions.

Libraries Used

- Pandas
- Numpy
- Sklearn

About Dataset

Features	Data Type
Id	Int64
LotFrontage	Float64
LotArea	Int64
Street	Object
Neighorhood	Object
YearBuilt	Int64
YearRemodAdd	Int64
CentralAir	Object
PavedDrive	Object
SaleCondition	Object
SalePrice	Int64



Features	Type
Id (Dropped)	Numerical
LotArea	Numerical
YearBuilt	Numerical
YearRemodAdd	Numerical
SalePrice	Numerical

Features	Type
Street	String
Centrail Air	Boolean
Paved Drive	Boolean
SaleCondition	String

Data Quality Test – Missing Value Imputation

Features	Missing Values
Id	0
LotFrontage	173
LotArea	0
Street	0
Neighorhood	0
YearBuilt	0
YearRemodAdd	0
CentralAir	0
PavedDrive	0
SaleCondition	0
SalePrice	0

Median Imputation

$$\frac{(n-1)+(n+1)}{2}$$

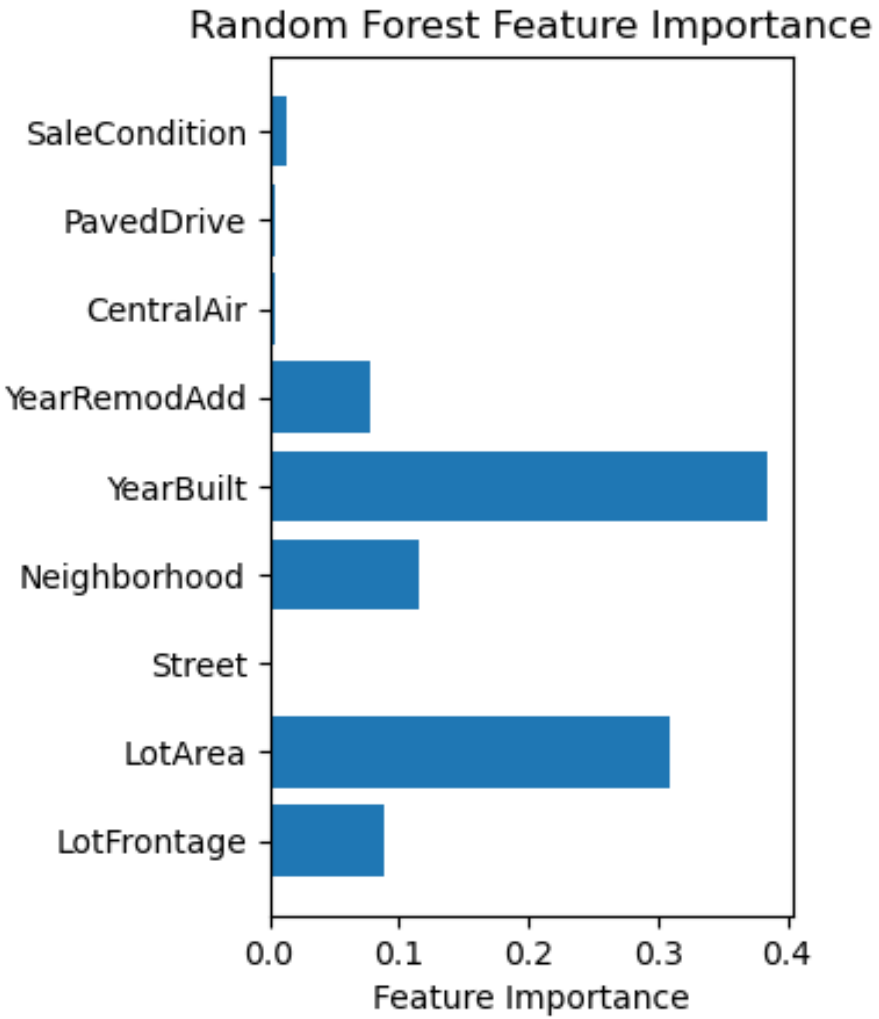
Alternative:
- Mean

$$\mu = \frac{\sum elements_i}{n}$$

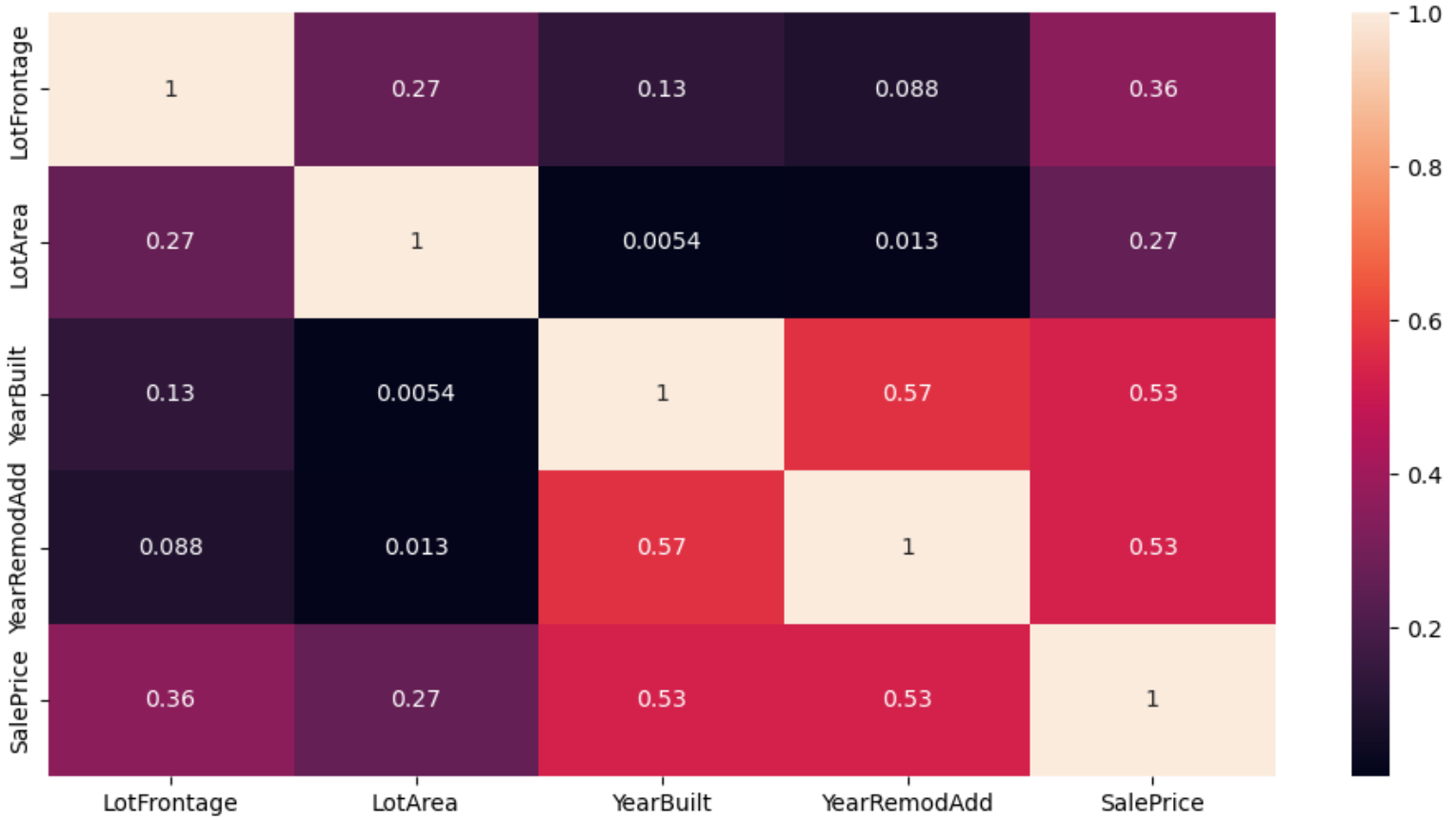
Features	Missing Values
Id	0
LotFrontage	0
LotArea	0
Street	0
Neighorhood	0
YearBuilt	0
YearRemodAdd	0
CentralAir	0
PavedDrive	0
SaleCondition	0
SalePrice	0

Feature Selection

- Feature importance was evaluated using the random forest approach or we can also any tree-based approach for feature selection such as Decision tree, etc.,
- Based on the visualization, we can interpret the features [YearBuilt, Neighborhood, LotArea, LotFrontage] carries a lot of weight which influence the SalePrice of a hosuing.



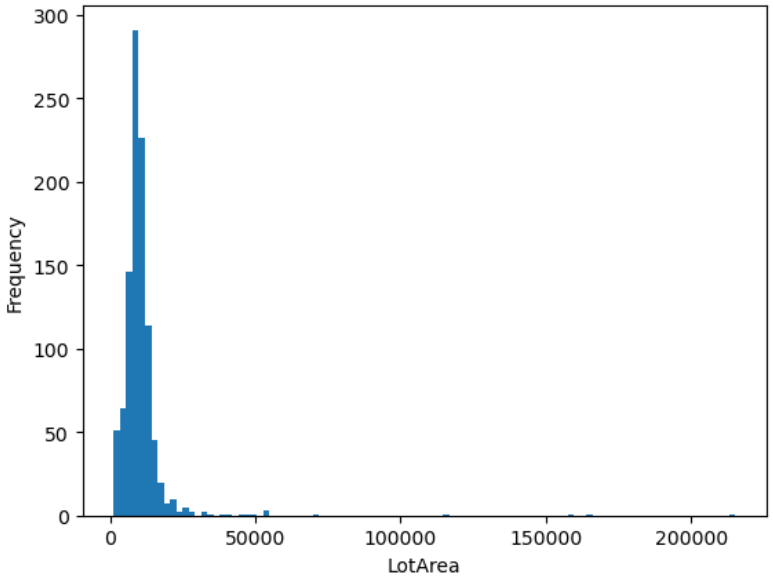
Correlation Analysis

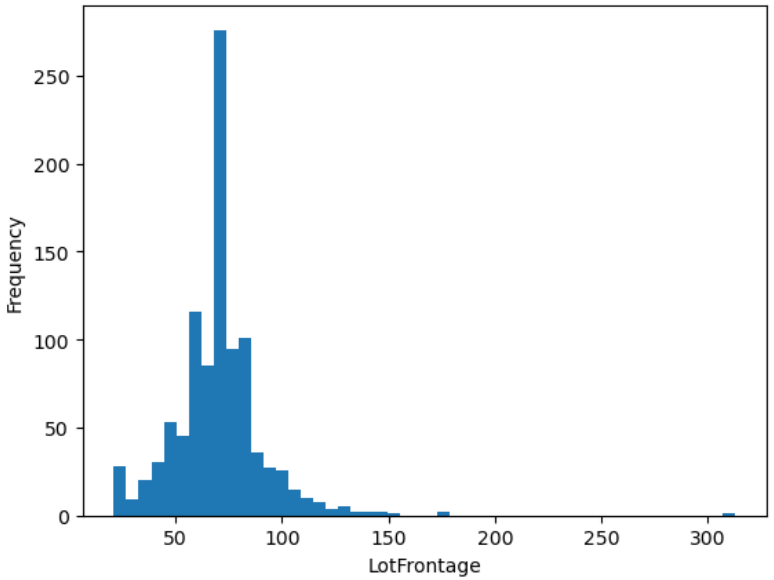


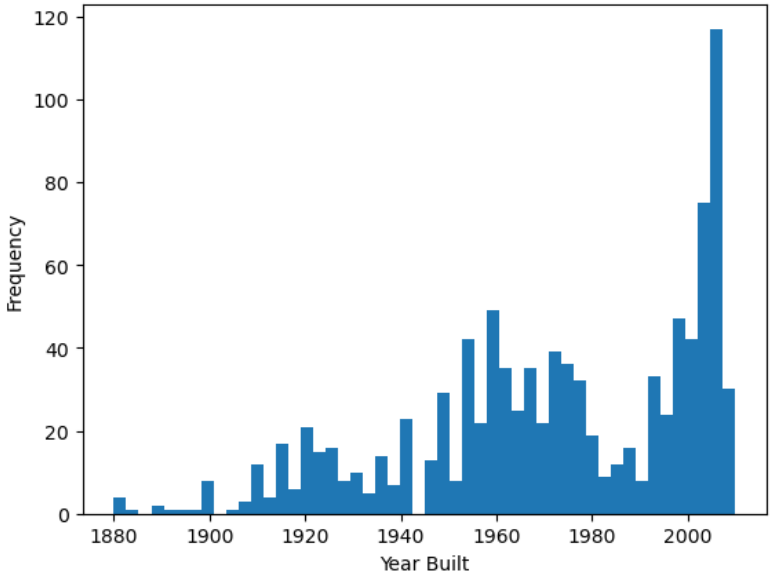
Interpretation:

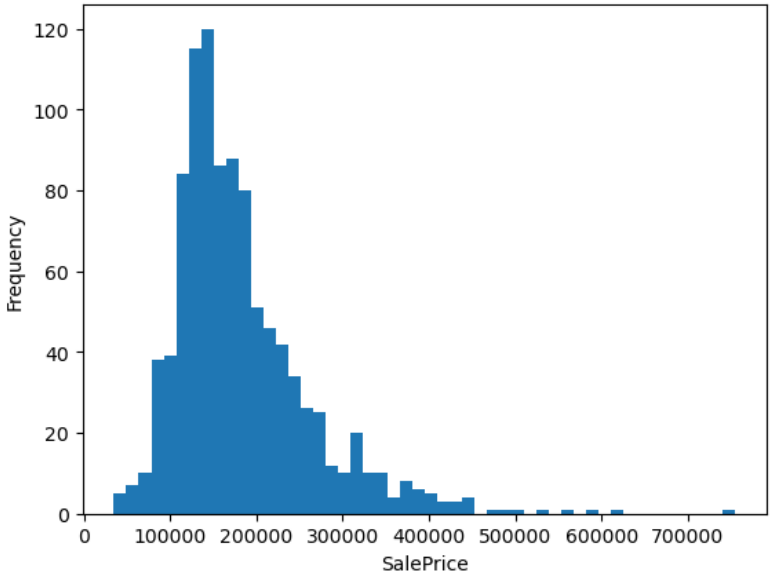
- Based on the visualization, we can interpret the features in the diagonal position are strongly correlated with each other and it is represented with the value 1.
- The features which are positively correlated with each other was indicated in darker color
- Since, there are no features which are represented in negative magnitude so we can say most of the features are correlated with each other.

Data Distribution





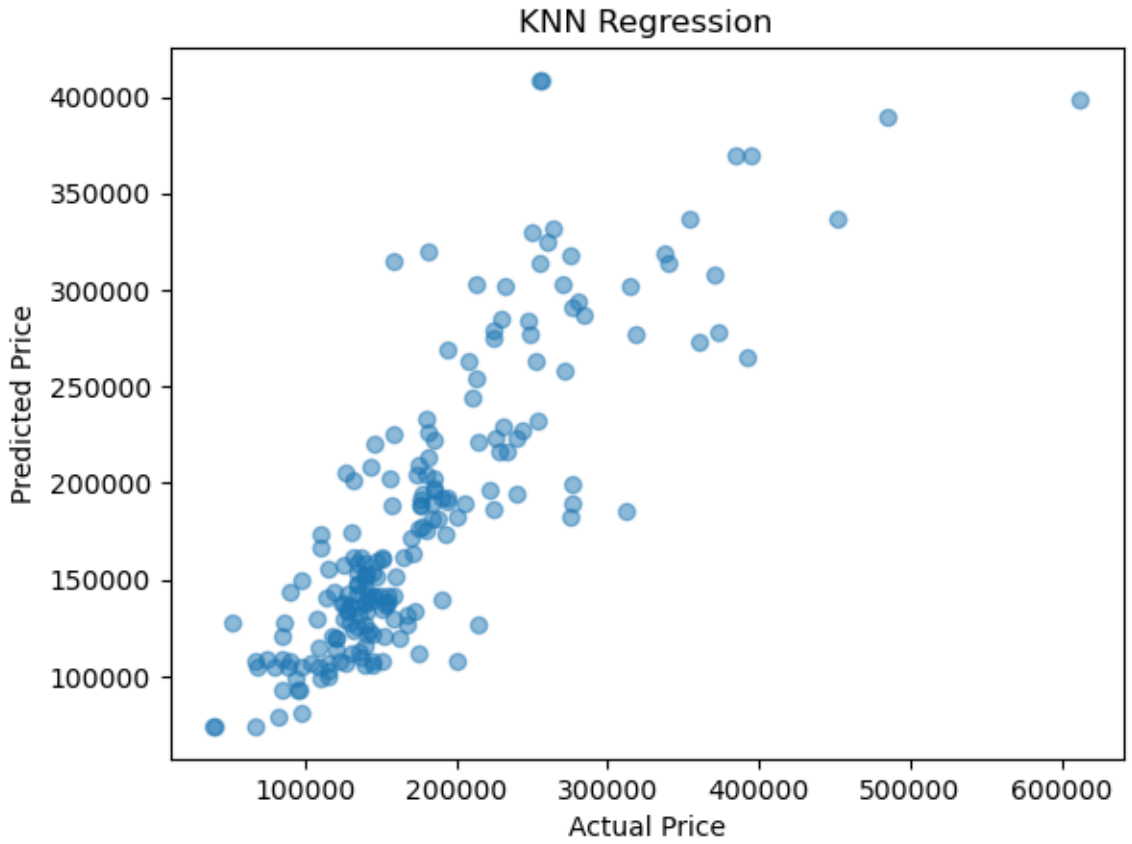




Interpretation:

- Since most of important features follows the Gaussian Distribution.
- In SalePrice feature, there is a presence of kurtosis at the end of the distribution tail.
- [2] Fatness at the end of the tail is usually called Leptokurtic which indicates there is a presence of outliers in smaller scale.

K-Nearest Neighbors



Metrics Used:

- Euclidean Distance = $\sqrt{\sum_{i=1}^{K}(x_i - y_i)^2}$
- Manhattan Distance = $\sum_{i=1}^{K}|x_i - y_i|$
- Minkowski Distance = $(\sum_{i=1}^{K}(|x_i - y_i|)^n)^{1/n}$

Result Interpretation:

- Based on the visualization, we can interpret the KNN regressor model has predicted that actual value is somewhat closer to the predicted value.
- For short distance, 4 nearest neighbors was considered which can be represented as K=4

Decision Tree Regressor

Metrics Used:

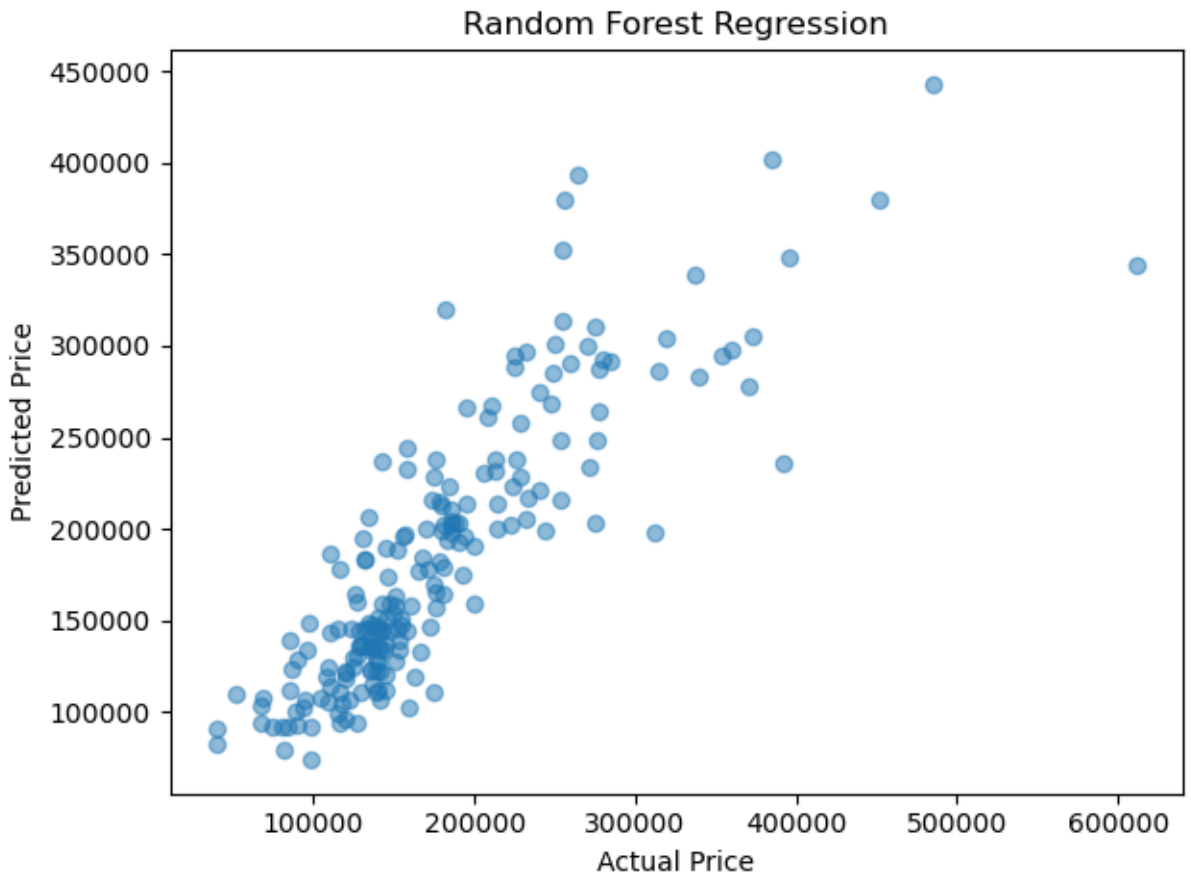
- Entropy (ΔS) = $-p * log_2(p) - (1-p) * log_2(1-p)$
- Information Gain =$\Delta V[Feature] = \left(\frac{S_i}{S}\right) * \Delta S$
- Gini Impurity = $1 - \sum_{i=1}^{n}(P)^2$
- MSE (Regression) = $\left(\frac{1}{\Delta S}\right) * \sum(y_i - \hat{y})^2$

Result Interpretation:

- Based on the visualization, we can interpret the Decision Tree regressor model has predicted that actual value is not really closer to the predicted value.

Decision Tree Hyperparameter optimal choice: DecisionTreeRegressor(max_depth=10, min_samples_leaf=4, random_state=42)

Random Forest Regressor



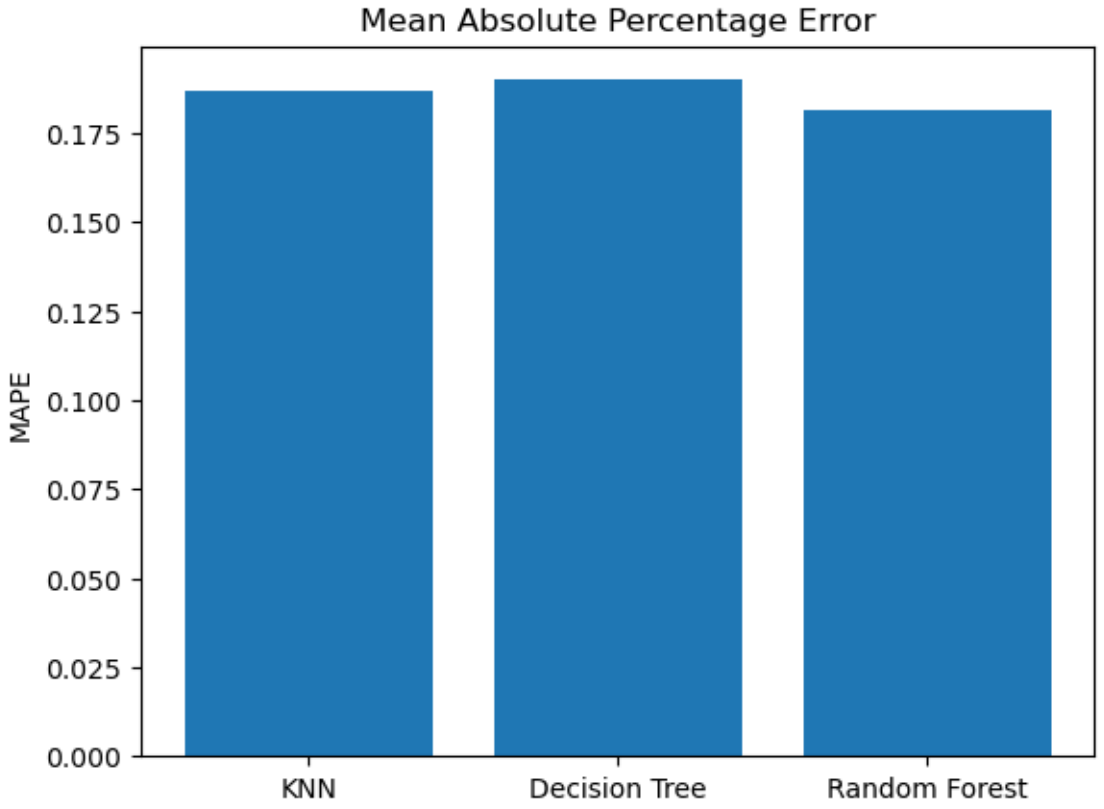
Metrics Used:

- Entropy (ΔS) = $-p * log_2(p) - (1-p) * log_2(1-p)$
- Information Gain =$\Delta V[Feature] = \left(\frac{S_i}{S}\right) * \Delta S$
- Gini Impurity = $1 - \sum_{i=1}^{n}(P)^2$
- MSE (Regression) = $\left(\frac{1}{\Delta S}\right) * \sum(y_i - \hat{y})^2$

Result Interpretation:

- Based on the visualization, we can interpret the Random Forest model has predicted that actual value is closer to the predicted value.
- Random Forest Hyperparameter optimal choice: RandomForestRegressor(n_estimators=150, random_state=42)

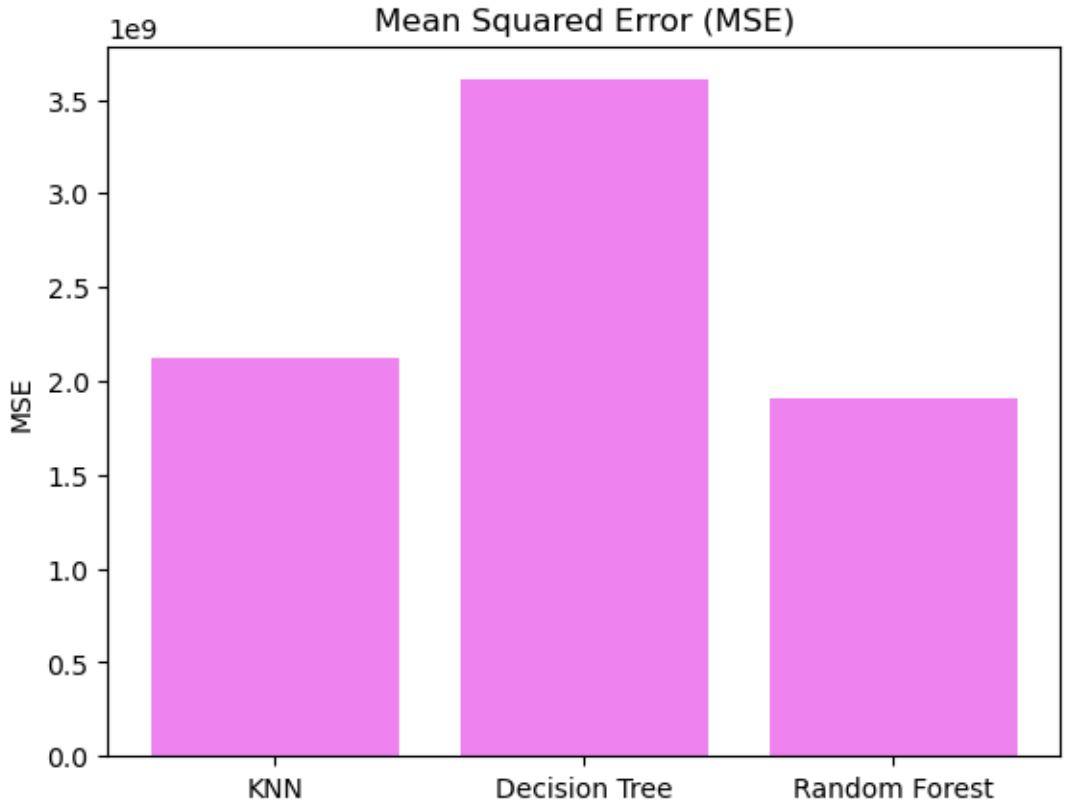
Evaluation Metrics

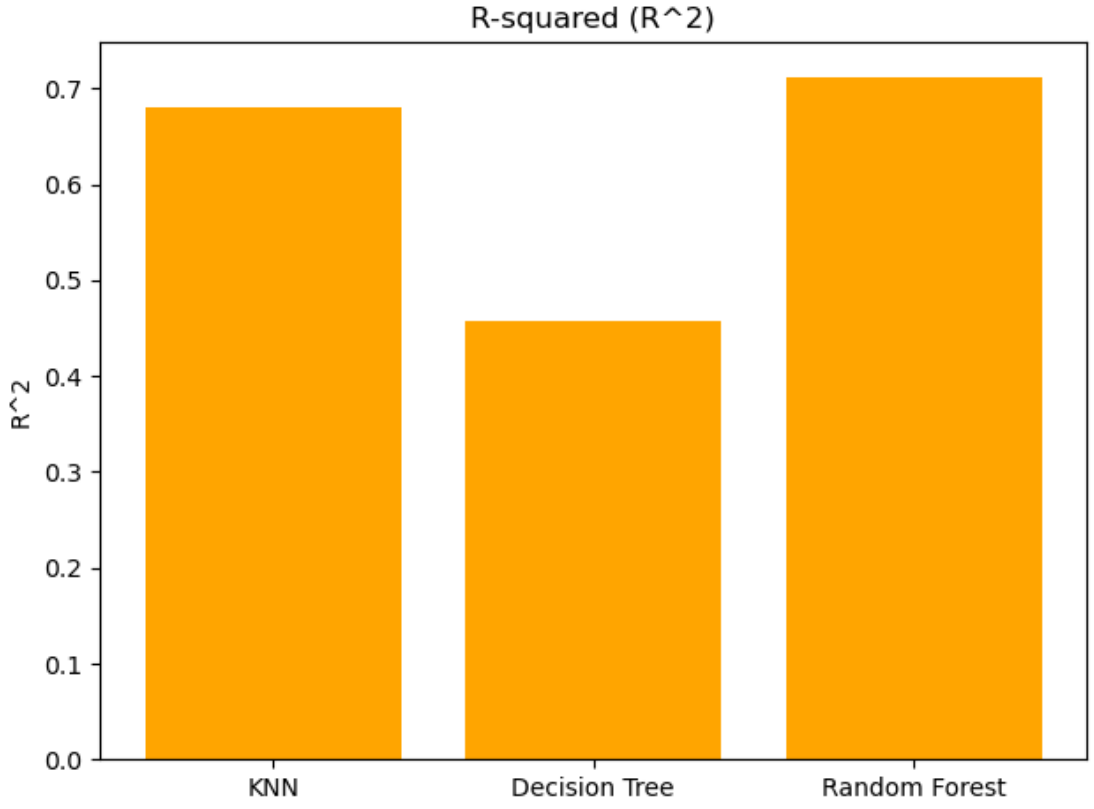


$$MAE = \left(\frac{1}{n}\right) * \sum |Actual\ - Predicted|$$

$$MAE = \left(\frac{1}{n}\right) * \sum \left|\frac{Actual\ - Predicted}{Actual}\right| * 100$$

Evaluation Metrics





$$MAE = \left(\frac{1}{n}\right) * \sum (Actual\ - Predicted)^2$$

$$R^2 = \left(\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n * ((\sum x^2)(\sum y)^2)\ * ((\sum y^2)(\sum y)^2)}}\right)$$

Conclusion

- Based on the results, Random Forest model has the lower Mean Squared Error rate (Squared difference between the actual value and the predicted value) with the value of 1.8 compared to Decision Tree and KNN Models.
- From R-Squared evaluation metric perspective, Random Forest produced 0.7 units compared to KNN = 0.67, and Decision Tree = 0.41 which indicates [1] Random Forest model analyzed there is a high correlation between features which explains there is small difference between the actual value and fitted value or predicted value.

- Repository: https://github.com/ravinthiranpartheepan1407/ML-Tasks/tree/main/CI-DM

Additional Slides:

- KNN: https://assets.super.so/7d26cd67-f43c-4085-8f4c-2a4594e5dd30/files/1bc699b1-2f3b-4566-a5f0-830c694fbfc6.pdf
- Decision Tree: https://assets.super.so/7d26cd67-f43c-4085-8f4c-2a4594e5dd30/files/9b9f66b2-cc7f-42df-9492-35dd30df29b0.pdf
- Website: https://ravinthiranpartheepanwiki.super.site/
- Future Work Consideration: Comparive study between sklearn vs Vertez (Link: https://pypi.org/project/vertezml/)

Reference

- [1] https://statisticsbyjim.com/regression/interpret-r-squared-regression/
- [2] https://www.scribbr.com/statistics/kurtosis/

Thank you!

Would you like to ask me any questions related to this task?