Name: Sushruth Danivasa Sridhar

Assignment – 2

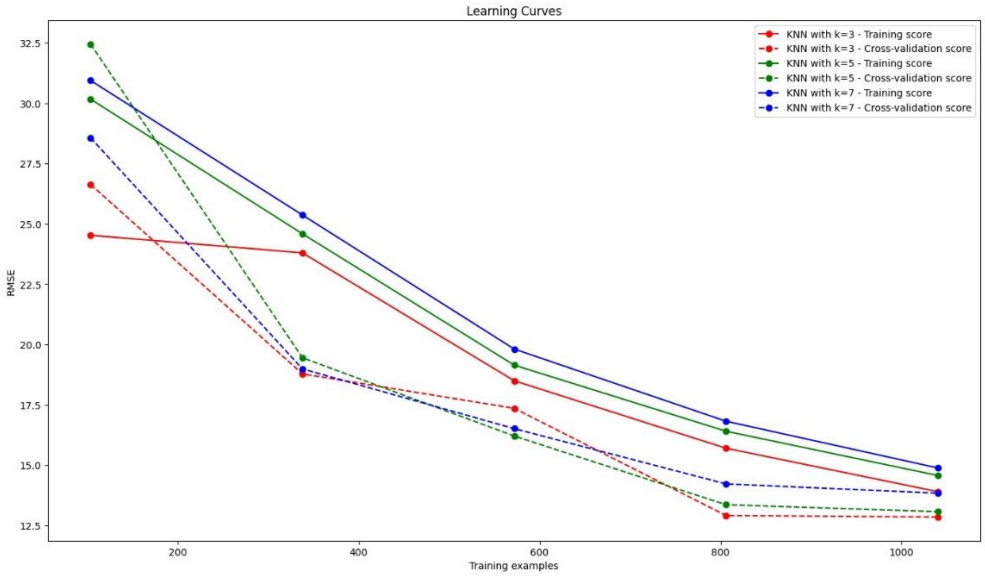
1. **A brief description of the dataset (what is the task, what are the features and the target)**

   For this assignment I have chosen socmob dataset with ID: 44987. This dataset focuses on social mobility, specifically examining how sons' occupations relate to their fathers' jobs. It explores the influence of a father's occupation, race, and family structure on the occupational outcomes of their sons. This dataset was collected for the survey titled "Occupational Change in a Generation II.

   The primary task associated with this dataset is likely to be a regression or analysis task, where the goal is to understand or predict the counts for sons' current occupation based on the other attributes. This involve exploring the relationship between a father's occupation and the son's current occupation, considering factors like race and family structure. The dataset contains 6 features in total, with 2 numeric features (counts_for_sons_current_occupation and counts_for_sons_first_occupation) and 4 nominal (categorical) features (fathers_occupation, sons_occupation, family_structure and race).
   The target in this dataset is counts_for_sons_current_occupation.

2. **Results for Task 1 in the form of graph and table.**

   For task 1, I have taken 3 k values which are 3, 5, 7.



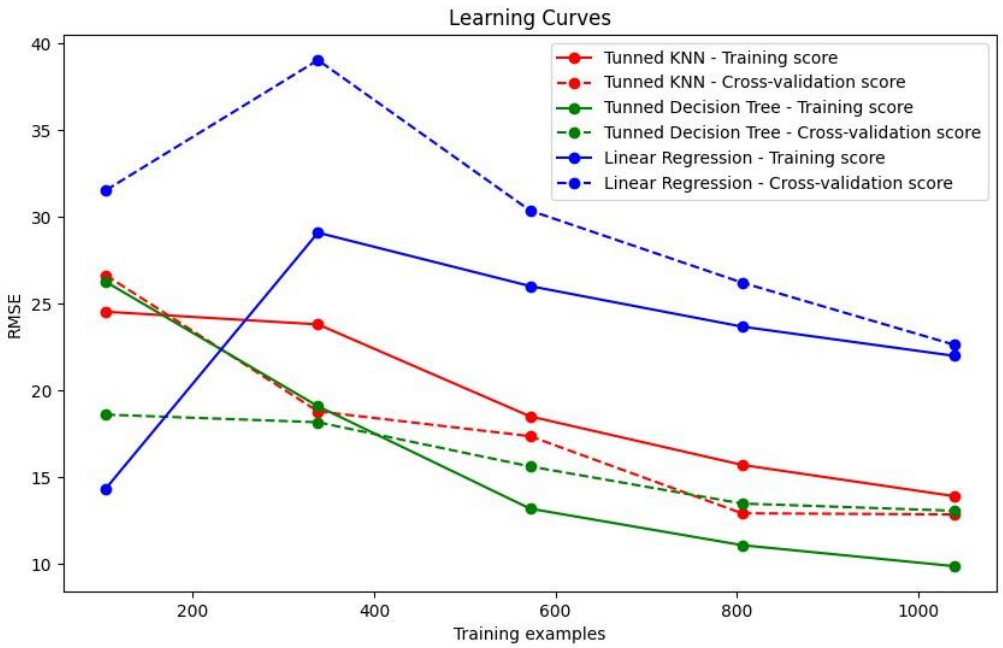
Task 1 Results:

k	Train Size	RMSE (Training)	RMSE (Test)
3	1040	13.900027	12.847619
5	1040	14.570214	13.070376
7	1040	14.882397	13.839265

3. Discussion on the results of Task 1.

Upon examining the outcomes of Task 1, in which I tested with various k values for the K-Nearest Neighbors regressor, a number of revelations became apparent. The model performed the best on the test set for k=3, indicating that it is the most successful at capturing the relationship in the data. The greater difference between the RMSEs for the test and training, however, can point to an overfitting inclination. Test RMSE gradually increased when k was increased to 5 and 7, which is a sign of underfitting, a condition in which the model's generalization starts to decrease. This trend is confirmed by the fact that as k grows, the difference between test and training scores gets less. Consequently, k=3 appears to offer a sweet spot among the evaluated k values, balancing the trade-off between model complexity and generalization ability.

4. Results for Task 2 in the form of graph and table.



Task 2 Results:

Model	Train Size	RMSE (Training)	RMSE (Test)
Best KNN	1040	13.900027	12.847619
Best Decision Tree	1040	9.865067	13.058267
Linear Regression	1040	21.985114	22.625101

5. Discussion on the results of Task 2.

Tunned KNN: With the lowest RMSE on the test set, the Best KNN model demonstrates a respectable generalization capability following hyperparameter adjustment. This shows that the adjusted model successfully and non-overfittingly reflects the underlying trends in the data.

Tunned Decision Tree: While showing an outstanding fit on the training set, the Best Decision Tree's performance on the test set is not appreciably superior to the Best KNN model. This might point to the Decision Tree's tendency toward overfitting, which could be lessened by pruning or additional parameter adjustment.

Linear Regression: With high RMSE values on both the training and test sets, the Linear Regression model performs worse than the other models. This suggests that there may be non-linear relationships in the dataset that the linear regression model is unable to identify.

In summary, the KNN model provides the best balance between performance and complexity, the Decision Tree model exhibits overfitting, and the Linear Regression model's incapacity to adequately represent the complexity of the dataset points to the need for more advanced feature engineering or modelling approaches.