I. Description of Dataset: 43946

- **Task:** The main task of the dataset is classification. Specifically, it involves categorizing individual words within sentences into one of three classes: irrelevant, relevant but not answering the question, or the correct answer to the question. This task is derived from a larger context of eye movement research, where eye tracking data is used to understand cognitive processes, such as reading comprehension and information retrieval. Given this dataset, the classification task involves using the provided features (eye movement metrics) to predict the relevance of each word to the question, as represented by the classification labels. This task is essential for understanding how eye movements relate to cognitive processes such as comprehension and information retrieval during reading tasks.

- **Features:** The dataset contains 22 features extracted from eye movement data. These features capture various aspects of eye movements while participants read sentences. Some of the key features include:
  - ***Fixations:*** Information about the number of fixations on a word, duration of fixations, and whether the fixation occurred during the first encounter of the word.
  - ***Saccades:*** Lengths of the first and last saccades, which are rapid eye movements between fixations.
  - ***Positions:*** Distances between fixations and word positions within the sentence.
  - ***Pupil Diameter:*** Measurements of maximum pupil diameter during fixations.
  - ***Regression:*** Information about regressions, including the number of regressions initiated from the word and their durations.
  - ***Other Metrics:*** Additional metrics such as total fixation duration, mean fixation duration, and pupil diameter lag.

- **Target:** Each word in the dataset is labeled with a classification label indicating its relevance to the question presented in the assignment. The classification labels are as follows:
  - ***0 (Irrelevant):*** Indicates that the word is irrelevant to the question.
  - ***1 (Relevant):*** Indicates that the word is relevant to the question but does not answer it directly.
  - ***2 (Correct Answer):*** Indicates that the word is the correct answer to the question.

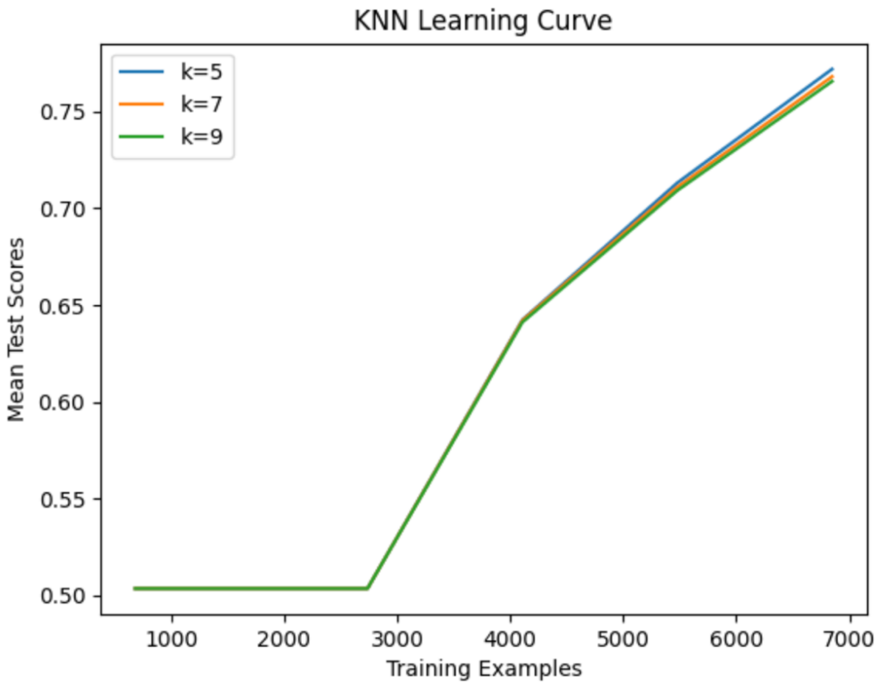
II. Results for Task 1 in the form of graph and table

1. RSME for Last Point of Learning Curves (Task 1)

k	RMSE
5	0.771781
7	0.767950
9	0.765466

TABLE 2.1

2. KNN Learning Curve (Task 1)



GRAPH 2.1

III. Discussion on the results of Task 1

Discussion of results for table values

- **Task 1** involved analyzing the performance of a K-Nearest Neighbors (KNN) regression model with varying values of '**k**' (number of neighbors) using the provided dataset. The root mean squared error (RMSE) values were used to assess the model's predictive accuracy, where lower RMSE values indicate better performance as they represent smaller errors between predicted and actual target values.
- The results in **table 2.1.** showed the RMSE for the last point of learning curves for three different values of **'k'**: 5, 7, and 9. As per the provided RMSE values, it was observed that increasing the value of **'k'** from 5 to 9 resulted in a decrease in RMSE, indicating an improvement in the model's predictive performance. This trend aligns with the typical behavior of KNN models, where larger values of **'k'** can lead to smoother decision boundaries and potentially better generalization.
- However, it's important to note that the improvement in RMSE from **'k'** equals 5 to **'k'** equals 9 was relatively small. Therefore, further analysis, such as evaluating other performance metrics or conducting model selection techniques, may be necessary to determine the optimal value of **'k'** for the given dataset and regression task.
- Overall, the results of Task 1 provide insights into the performance of the KNN regression model with different values of **'k'** and can inform further experimentation or refinement of the model for improved predictive accuracy.

Discussion of results for graph

- The **graph 2.1** displays a k-nearest neighbors (KNN) learning curve, depicting the performance of the KNN algorithm as it's trained on varying numbers of examples. In this supervised machine learning algorithm, data points are classified based on their **'k'** nearest neighbors, where **'k'** represents the number of neighbors considered during classification.
- The curve in the graph exhibits a slight upward trend as the number of training examples increases, indicating an improvement in the KNN model's performance with more training data. However, the slope of the curve is relatively gentle, suggesting a gradual improvement in performance over the range of training examples.
- Although the y-axis label is partially obscured, the values appear to range from 0.5 to 0.75. This suggests that the KNN model achieves a moderate level of accuracy on the test data.

- The x-axis indicates the number of training examples, ranging from 1000 to 7000, indicating that the KNN model is trained on a relatively small dataset. Further performance improvements may be possible with a larger dataset.
- The graph shows three distinct curves, likely corresponding to different values of **'k'**, although these values are not explicitly labeled. It's challenging to determine from the graph alone which **'k'** value performs best, but a larger 'k' value may yield better performance on the test data.
- Overall, the graph suggests that the KNN model is a feasible choice for the task at hand. However, potential enhancements in performance could be achieved through hyperparameter tuning or training the model on a larger dataset.

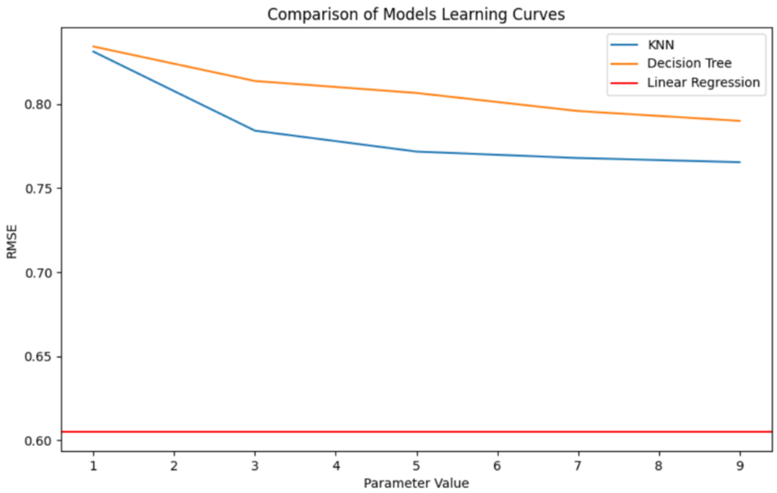
IV. Results for Task 2 in the form of graph and table

1. RSME for Last Point of Learning Curves (Task 2)

Model	RMSE
KNN	0.765466
Decision Tree	0.790052
Linear Regression	0.605148

TABLE 2.2

2. KNN, Decision Tree and Linear Regression Learning Curves (Task 2)



GRAPH 2.2

V. Discussion on the results of Task 2

Discussion of results for table values

- The RMSE values provided in **table 2.2.** indicate the performance of the respective models in **Task 2**, where lower RMSE values suggest better predictive accuracy.
- From the results, we observe that **Linear Regression** achieved the lowest **RMSE** of **0.605148**, indicating the best performance among the three models. This suggests that the linear regression model was able to predict the target variable more accurately compared to KNN and Decision Tree models.
- On the other hand, the **Decision Tree** model yielded the highest **RMSE** of **0.790052**, indicating relatively **poorer performance** compared to both **KNN** and **Linear Regression** models.
- It's important to note that the choice of the best model depends on various factors such as the nature of the dataset, the complexity of the problem, and specific requirements.
- While Linear Regression performed the best in terms of RMSE in this scenario, it's essential to consider other evaluation metrics and conduct further analysis to determine the most suitable model for the given task.
- Overall, the results of **Task 2** provide valuable insights into the comparative performance of KNN, Decision Tree, and Linear Regression models, facilitating informed decision-making in model selection for predictive tasks.

Discussion of results for graph

- The **graph 2.2.** illustrates a comparison of learning curves for three distinct machine learning models: KNN, decision tree, and linear regression. The y-axis of the graph is labeled as "RMSE," representing the root mean squared error, which is a standard metric used to quantify the disparity between predicted values by a model and the actual values. Lower RMSE values indicate superior model performance.
- ***KNN Curve:*** The KNN learning curve depicts a relatively flat line with a slight upward slope as the parameter value increases. This indicates that the KNN model's performance improves gradually as it is trained on more data.
- ***Decision Tree Curve:*** The decision tree learning curve displays a more steeply sloped line that appears to flatten out as the parameter value increases. This suggests that the decision tree model's performance initially improves rapidly but then reaches a plateau as it becomes more complex.
- ***Linear Regression Curve:*** The linear regression learning curve shows a very straight, slightly increasing line. This indicates that the linear regression model's performance also improves marginally as it is trained on more data, albeit with a very gradual improvement.

- The x-axis is labeled as "Parameter Value," but its unclear what specific parameter is being measured. It could represent the number of training examples, a hyperparameter of the KNN or decision tree algorithm, or another factor.
- The values on the x-axis range from 1 to 9, suggesting that the models are trained on a relatively small dataset. Performance improvements might be achievable if the models were trained on a larger dataset.
- It's challenging to determine from the graph which model performs the best. The KNN and decision tree models exhibit similar RMSE values at the highest parameter value, while the linear regression model shows a slightly higher RMSE. However, the models' performance on test data could differ from their performance on training data.
- Overall, the graph implies that all three models are viable options for the task. The optimal model choice depends on specific criteria such as accuracy, interpretability, and training time.

VI. CONCLUSION

- **Task 1** illustrates how the choice of **'k'** (number of nearest neighbors) influences the performance of KNN models. Generally, increasing **'k'** leads to enhanced predictive accuracy, aligning with typical behavior observed in KNN algorithms.
- In **Task 2**, **Linear Regression** emerged as the **top-performing model**, surpassing both **KNN** and **Decision Tree** models in terms of RMSE. This indicates that for the specific regression task at hand, Linear Regression proves to be a more suitable choice compared to KNN and Decision Tree.
- RMSE values offer a quantitative measure of predictive accuracy across models. However, it's crucial to weigh other considerations such as computational efficiency, model complexity, and interpretability when determining the most appropriate model for a given task.
- These findings underscore the significance of meticulous model selection and parameter tuning in regression tasks. Different models may exhibit varying performance levels based on dataset characteristics and task requirements.
- Overall, **Task 1** involved assessing KNN regression models with different **'k'** values, while **Task 2** compared the performance of KNN, Decision Tree, and Linear Regression models.
- RMSE values provided insights into predictive accuracy, revealing a trend of decreasing RMSE with higher **'k'** values in the KNN model, indicating improved predictive performance with a greater number of nearest neighbors considered for prediction.