A brief description of dataset:

The dataset consists of 53940 instances and has 9 features. Out of 9 features, 3 of them are nominal, and 6 of them are numeric. The objective is to forecast the price of diamonds based on the nine features. There were no missing values in the data. More information about the data can be found here: https://www.openml.org/d/42225. The list of features is as follow:

- price price in US dollars (\$326--\$18,823)
- carat weight of the diamond (0.2--5.01)
- cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- color diamond colour, from J (worst) to D (best)
- clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- x length in mm (0--10.74)
- y width in mm (0--58.9)
- z depth in mm (0--31.8)
- depth total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
- table width of top of diamond relative to widest point (43--95)

The data before handling the categorical has 9 columns as follow:

```
 #   Column   Non-Null Count  Dtype

---  ------   --------------  -----
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  category
 2   color    53940 non-null  category
 3   clarity  53940 non-null  category
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   x        53940 non-null  float64
 7   y        53940 non-null  float64
```

After handling the categorical features, the number of columns increased to 26 as follow :

```
#   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   encoder__cut_Fair        53940 non-null  float64
 1   encoder__cut_Good        53940 non-null  float64
 2   encoder__cut_Ideal       53940 non-null  float64
 3   encoder__cut_Premium     53940 non-null  float64
 4   encoder__cut_Very Good   53940 non-null  float64
 5   encoder__color_D         53940 non-null  float64
 6   encoder__color_E         53940 non-null  float64
 7   encoder__color_F         53940 non-null  float64
 8   encoder__color_G         53940 non-null  float64
 9   encoder__color_H         53940 non-null  float64
```

```
10  encoder__color_I        53940 non-null  float64
11  encoder__color_J        53940 non-null  float64
12  encoder__clarity_I1     53940 non-null  float64
13  encoder__clarity_IF     53940 non-null  float64
14  encoder__clarity_SI1    53940 non-null  float64
15  encoder__clarity_SI2    53940 non-null  float64
16  encoder__clarity_VS1    53940 non-null  float64
17  encoder__clarity_VS2    53940 non-null  float64
18  encoder__clarity_VVS1   53940 non-null  float64
19  encoder__clarity_VVS2   53940 non-null  float64
20  remainder__carat        53940 non-null  float64
21  remainder__depth        53940 non-null  float64
22  remainder__table        53940 non-null  float64
23  remainder__x            53940 non-null  float64
24  remainder__y            53940 non-null  float64
25  remainder__z            53940 non-null  float64
```

Task 1:

	Linear regression	Decision trees	K-nearest neighbor	Support vector machine
Base	**1123.33**	877.38	**1180.79**	**3414.02**
Bagged	1126.53	**784.52**	1190.6134	3421.86
p-value	0.27	0.006*	0.55	0.76

- Results (RMSE) shown in bold were better when compared in the same column. Lower RMSE values indicate better performance, as they indicate that the model's predictions are closer to the actual values.

  The RMSE value for the bagged decision tree model is lower than that of the base model (single decision tree), it suggests that the bagging method has improved the model's performance. This is because the bagging method reduces the variance in the predictions by averaging over multiple trees, which can help to reduce overfitting.

  However, if the RMSE values are very similar, it may not be clear whether the bagging method has significantly improved the performance. In such cases, further analysis may be required to determine which model is better.

- The result marked with* were found to be statistically significant at $p < 0.05$ level using two-tailed paired t-test when compared with the result in the same column.

  For instance, the bagged decision tree outperformed the single decision tree in a statistically significant manner. However, if the p-value is greater than the significance level, it suggests that the difference in performance between the models is not statistically significant (for Linear regression and Bagged Linear regression, K-nearest neighbor and Bagged K-nearest neighbor, Support vector machine and Bagged Support vector machine).

Task 2:

	Linear regression	Decision trees	K-nearest neighbor	Support vector machine
Base	**1123.33**	877.38	**1180.79**	**3414.02**
Boosted	1362.67	**784.31**	1313.49	3480.22
p-value	0.08	0.58	0.03*	0.42

- Results (RMSE) shown in bold were better when compared in the same column. Lower RMSE values indicate better performance, as they indicate that the model's predictions are closer to the actual values.

  The boosted decision tree has a lower average RMSE than the single decision tree, it suggests that the boosting method has improved the model's performance. Boosting can help to reduce bias and variance in the predictions and improve the model's generalization ability.

  However, if the difference in the average RMSE values is not significant, it may be difficult to conclude that one model is better than the other. In such cases, you may need to perform additional experiments, such as comparing the variance in the predictions or evaluating the models on different datasets, to determine which model is more robust and reliable.

- The result marked with* were found to be statistically significant at $p < 0.05$ level using two-tailed paired t-test when compared with the result in the same column.

  The p-value associated with the k-nearest neighbor model is 0.03, which is below the commonly used threshold of 0.05 for statistical significance. From these findings, we can conclude that the k-nearest neighbor model may be a better choice than the boosted model for this particular dataset and problem. The statistical significance of the p-value associated with the k-nearest neighbor model suggests that the observed difference in performance between the two models is unlikely to be due to chance.

  However, it's important to note that the choice of model should not be based solely on statistical significance. Other factors, such as the interpretability of the model, computational resources required, and practical considerations, should also be taken into account when selecting a model. Additionally, it may be worth considering other evaluation metrics beyond just the RMSE to get a more complete picture of model performance.

Task 3:

Part 1.

	Linear regression	Decision trees	K-nearest neighbor	Support vector machine
Base	**1123.33**	**877.38**	1180.79	3414.02
Voting	1172.56	1172.56	**1172.56**	**1172.56**
p-value	0.78	0.11	0.93	0.003*

- Results (RMSE) shown in bold were better when compared in the same column. Lower RMSE values indicate better performance, as they indicate that the model's predictions are closer to the actual values.

- The result marked with* were found to be statistically significant at $p < 0.05$ level using two-tailed paired t-test when compared with the result in the same column.

Part 2.

	Voting	Linear regression	Decision trees	K-nearest neighbor	Support vector machine
RMSE	1172.56	1123.33	***877.38***	1180.79	3414.02

	Voting	Linear regression	K-nearest neighbor	Support vector machine
	Decision trees	**Decision trees**	**Decision trees**	**Decision trees**
p-value	0.11	0.02*	0.008*	0.006*

- The best result is shown in bold and italics. Decision trees model has better performance than others.

- The results marked with* were found to be statistically significant at $p < 0.05$ level using two-tailed paired t-test when compared with the result in the same column.

  The statistical significance of the p-value associated with the k-nearest neighbor and Decision trees model, Linear regression and Decision trees model, Support vector machine and Decision trees model suggests that the observed difference in performance between the two models is unlikely to be due to chance.