COMPSCI 711

Machine Learning Assignment Report.

Description:

The "soil.csv" collection offers details about the properties of the soil. It has 5 features and 8641 instances. The "track" feature is the target feature. The remaining four features are:

Feature Name	Type	Distinct/Missing Values
track (target)	numeric	40 distinct values 0 missing attributes
northing	numeric	7011 distinct values 0 missing attributes
easting	numeric	6069 distinct values 0 missing attributes
resistivity	numeric	5726 distinct values 0 missing attributes
isns	nominal	2 distinct values 0 missing attributes

Regression analysis is used in this dataset's task with the aim of predicting the value of the "track" feature based on the other attributes. Since there are no classes and no missing values in the dataset, the issue is not one of categorization.

Overall, this dataset comprises data on soil parameters and can be used to estimate the value of the "track" feature using regression analysis.

Task 1:

We will compare the 4 regression rmse's with their Bagged versions. Below we can find the table.

Method	Decision Tree	K Nearest Neighbor	Linear Regression	SVR
Base	2.1794	4.1657	**4.4901**	11.4876
Bagged	**1.9257**	**3.7844***	4.4905	**11.4661**
P_value	0.0522	0.0092	0.3707	0.7509

Table Description:

The provided table presents the results of four regression methods, namely Decision Tree, K Nearest Neighbor, Linear Regression, and Support Vector Regression (SVR), and compares their base errors and bagged errors. The base errors represent the errors of the unmodified regression methods, while the bagged errors represent the errors of the regression methods that were modified through bagging.

The table also contains the p-values, which show the importance of the difference between base errors and bagged errors for each approach in terms of statistics. The Decision Tree and K Nearest Neighbor algorithms' p-values are less than 0.05, indicating that there is statistical significance in the difference between their base and bagged errors. The p-values, however, are higher than 0.05 for the SVR and linear regression approaches, indicating that the difference between their base and bagged errors is not statistically significant.

Thoughts & Conclusions:

The study examines Decision Tree, K Nearest Neighbor, Linear Regression, and Support Vector Regression's base error and bagged error. (SVR).

The findings reveal that bagging can increase the accuracy of the Decision Tree and K Nearest Neighbor approaches since the bagged versions had lower errors than the corresponding base versions. The base and bagged errors for the Linear Regression approach, however, are the same, which raises the possibility that bagging may not be useful in this situation.

The bagged error of the SVR algorithm is somewhat smaller than its base error, indicating a marginally better performance. In general, bagging can be a good way for increasing the accuracy of regression methods, however the usefulness of this method can vary depending on the regression method utilized.

Task 2:

We will compare the base values of the four rmse's with their boosted versions:

Method	Decision Tree	K Nearest Neighbor	Linear Regression	SVR
Base	2.1794	4.1657	**4.4901**	11.4876
Boosted	**1.5964**	**3.3208**	5.5948	**10.228***
P-values	0.0572	0.0149	0.1609	0.0019

Table Description:

The Decision Tree, K Nearest Neighbor, Linear Regression, and Support Vector Regression results are shown in the table. (SVR). While the Boosted errors represent the errors of the regression techniques that were improved through boosting, the Base errors represent the errors of the unmodified regression methods. In accordance with the findings, all methods—aside from linear regression—have boosted errors that are lower than their base errors, indicating that boosting can raise the precision of these techniques.

The p-values for the difference between the base errors and the Boosted errors for each method are also included in the table. These values show the statistical significance of the difference.

Thoughts & Conclusions:

The table compares the base errors and Boosted errors of the four regression techniques: Support Vector Regression, Decision Tree, and K Nearest Neighbor. (SVR). The errors of the updated regression methods by boosting are represented by the boosted errors.

The findings reveal that, except for linear regression, all methods' boosted mistakes are lower than their base errors, demonstrating that boosting can greatly improve these approaches' accuracy. The Support Vector Regression method demonstrated the greatest increase in accuracy through boosting of all the techniques.

Additionally, each method's base errors and Boosted errors are shown in the table along with their respective p-values, which indicate how statistically significant they are different. The Decision Tree, K Nearest Neighbor, and Linear Regression methods all have p-values greater than 0.05, indicating that there is no statistically significant difference between their base and Boosted errors. The SVR approach, in contrast, has a p-value of less than 0.05, indicating that the difference between its base and Boosted mistakes is statistically significant.

Overall, the table demonstrates that boosting can be a useful way for improving the precision of regression techniques, particularly for the SVR method. However, the regression technique employed may have an impact on how effective boosting is.

Task 3:

Method	Voted Regression	Decision Tree	K Nearest Neighbor	Linear Regression	SVR
RMSE's	4.2630	***2.1794***	4.1657	4.4901	11.4876
P-value	0.0108		0.00518	0.01708	9.667269230788565e-05 (This is less than 0.05)

Table Description:

Lower values in the table indicate greater performance when comparing the output of five machine learning algorithms on a dataset. Voted Regression, Decision Tree, KNN, Linear Regression, and SVR are the five algorithms. The statistical significance of the performance differences between Voted Regression and the other methods is shown in the P-value column. Voted Regression performs noticeably worse than Decision Tree and KNN because their p-values are both less than 0.05.

Voted Regression works noticeably better than these two techniques, as evidenced by the fact that the p-values for Linear Regression and SVR are substantially lower than 0.05. Accordingly, the findings imply that Decision Tree and KNN are superior options to Voted Regression for this dataset, while Linear Regression and SVR perform badly when compared to Voted Regression.

Thoughts & Conclusions:

The lower RMSE values imply better performance of the algorithms. Among the five algorithms, Decision Tree performed the best with an RMSE of 2.1794, while Voted Regression had an RMSE of 4.2630, which is higher than the RMSE of Decision Tree and KNN but lower than the RMSE of Linear Regression and SVR.

The P-value column signifies the statistical significance of the differences in performance between Voted Regression and the other algorithms. The p-values for Decision Tree and KNN are less than 0.05, indicating that Voted Regression performs significantly worse than these two methods. The p-values for Linear Regression and SVR are much lower than 0.05, indicating that Voted Regression performs significantly better than these two methods. Consequently, the findings propose that Decision Tree and KNN are better options than Voted Regression for this dataset, while Linear Regression and SVR are inferior to Voted Regression.

Filename:	Report 2.docx
Directory:	C:\Users\jajia\Documents
Template:	C:\Users\jajia\AppData\Roaming\Microsoft\Templates\Normal.dotm
Title:	
Subject:	
Author:	Sri Adi Narayana Repudi
Keywords:	
Comments:	
Creation Date:	09-04-2023 19:37:00
Change Number:	23
Last Saved On:	10-04-2023 09:54:00
Last Saved By:	Sri Adi Narayana Repudi
Total Editing Time:	496 Minutes
Last Printed On:	10-04-2023 09:54:00

As of Last Complete Printing

Number of Pages:	3
Number of Words:	1,119 (approx.)
Number of Characters:	6,380 (approx.)