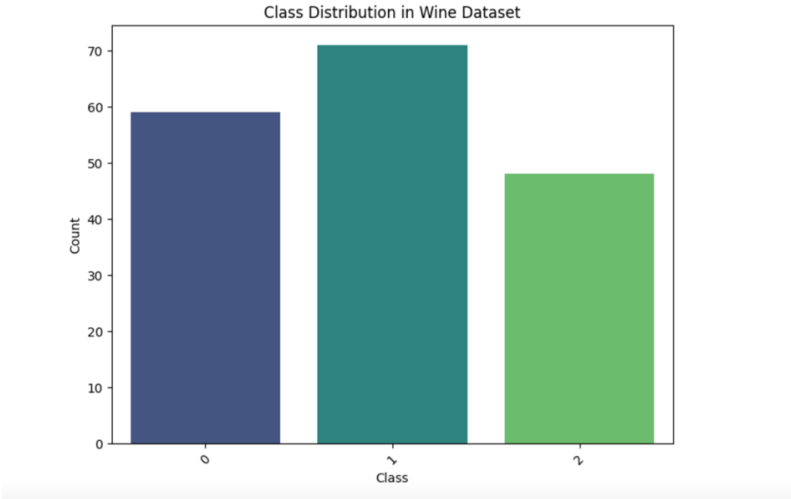
ADVANCE MACHINE LEARNING
ASSIGNMENT 3

TASK 1

Choose three arbitrary datasets; you can use scikit learn toy datasets as well (https://scikit-learn.org/stable/datasets/toy_dataset.html)

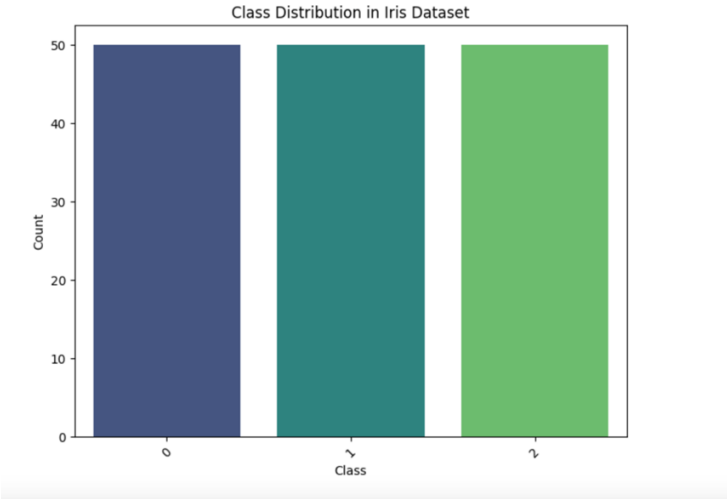
Successfully loaded 3 datasets using scikit learn toy datasets. Dataset chosen:
1. Wine dataset (*Classification*)
2. Diabetes dataset (*Regression*)
3. IRIS dataset (*Classification*)




```
In [7]: #### DIABETES (REGRESSION DATASET)

In [8]: from sklearn.datasets import load_diabetes

        diabetes_data = load_diabetes(as_frame=True)

        diabetes_df = diabetes_data.frame
        X_diabetes_df = diabetes_data.data
        y_diabetes_df = diabetes_data.target

        diabetes_df.head()
```

Out[8]:

	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	target
0	0.038076	0.050680	0.061696	0.021872	-0.044223	-0.034821	-0.043401	-0.002592	0.019907	-0.017646	151.0
1	-0.001882	-0.044642	-0.051474	-0.026328	-0.008449	-0.019163	0.074412	-0.039493	-0.068332	-0.092204	75.0
2	0.085299	0.050680	0.044451	-0.005670	-0.045599	-0.034194	-0.032356	-0.002592	0.002861	-0.025930	141.0
3	-0.089063	-0.044642	-0.011595	-0.036656	0.012191	0.024991	-0.036038	0.034309	0.022688	-0.009362	206.0
4	0.005383	-0.044642	-0.036385	0.021872	0.003935	0.015596	0.008142	-0.002592	-0.031988	-0.046641	135.0



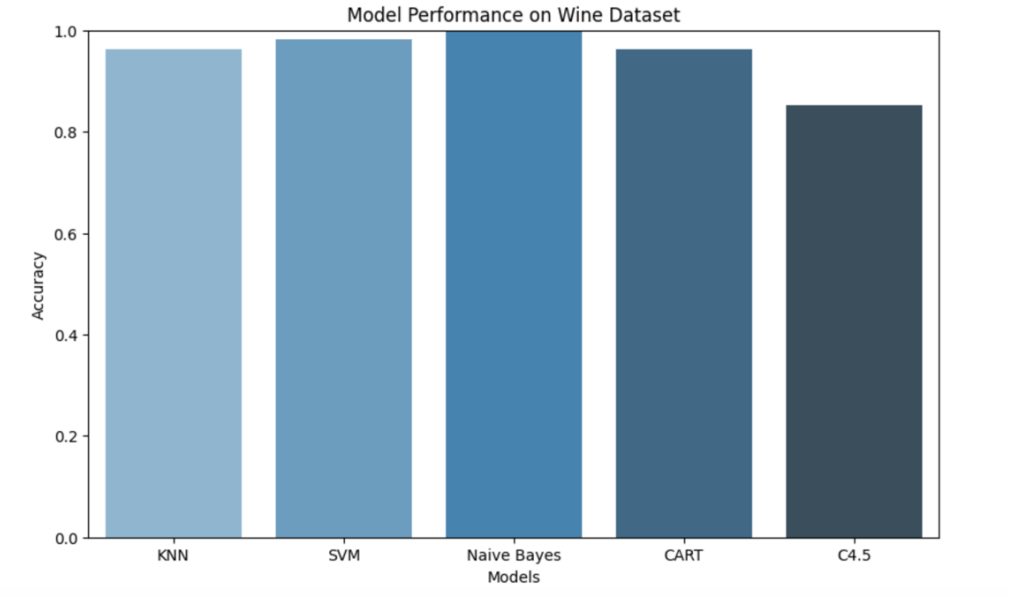
ADVANCE MACHINE LEARNING
ASSIGNMENT 3

TASK 2

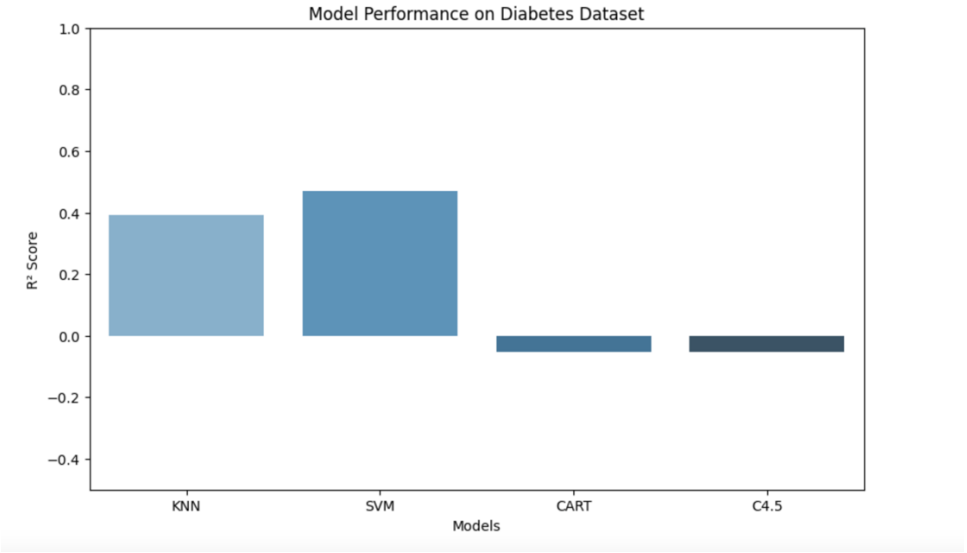
Apply a prediction on these datasets, which all weak learners, including kNN, SVM, Naïve Bayes, and at least two decision trees (ID3, CART, C4.5, CHAID). Report the accuracy of each algorithm. The train test split should be 70/30, and you should also use 10-fold cross-validation.

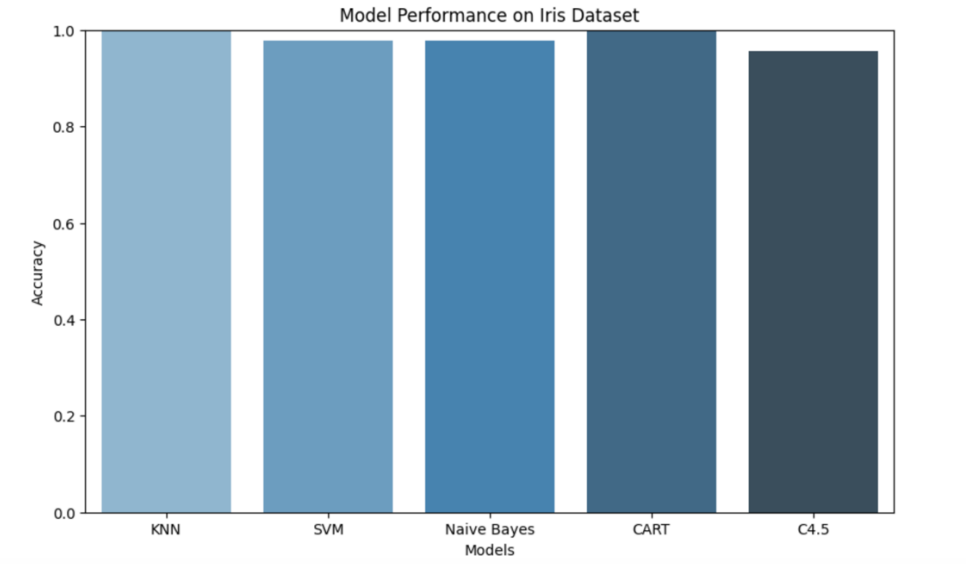
Summary of results:

DATASET	MODEL	ACCURACY	10-FOLD CV ACCURACY
Wine	kNN	96.30 %	95.26 %
	SVM	98.15 %	98.40 %
	Naïve Bayes	100 %	95.96 %
	CART Decision Tree	96.30 %	93.65 %
	C4.5 Decision Tree	85.19 %	87.88 %
IRIS	kNN	100 %	92.64 %
	SVM	97.78 %	95.45 %
	Naïve Bayes	97.78 %	92.45 %
	CART Decision Tree	100 %	91.55 %
	C4.5 Decision Tree	95.56 %	91.55 %
DIABETES (R^2)	kNN	0.39	0.27
	SVM	0.47	0.40
	CART Decision Tree	-0.05	-0.16
	C4.5 Decision Tree	-0.05	-0.16



ADVANCE MACHINE LEARNING
ASSIGNMENT 3





<u>Why no Naïve Bayes and negative performance resulting FROM CART and C4.5 for Diabetes dataset?</u>

In my analysis, I opted not to include the **Naïve Bayes** model in the evaluation of the Diabetes dataset's performance primarily because Naïve Bayes is traditionally used for classification tasks rather than regression tasks. The Diabetes dataset involves predicting continuous outcomes, such as diabetes progression, which requires regression models that can effectively handle numerical predictions.

Regarding the negative performance of both **CART** and **C4.5** decision trees on the Diabetes dataset, this result was surprising. Both models yielded R^2 scores of **-0.05**, indicating that their predictions were worse than simply using the average of the target variable. This negative performance can occur for several reasons.

Firstly, decision trees, including CART and C4.5, can struggle with overfitting, especially if the dataset contains noise or irrelevant features. This means that the model may have learned

patterns that do not generalize well to new data, leading to poor predictions. Additionally, the features in the Diabetes dataset may not be well-suited for decision tree models, as they might not capture the complex relationships present in the data.

Overall, these findings highlight the importance of model selection based on the nature of the data and the specific task at hand. For continuous prediction tasks like those found in the Diabetes dataset, alternative models or further refinement of the decision tree parameters might yield better results.

Why CART AND C4.5 decision tree with respect to other decision tree I choose?

I chose to work with the **CART** (Classification and Regression Trees) and **C4.5** decision trees because they offer distinct advantages that make them suitable for my analysis. CART is versatile, meaning it can handle both classification tasks, like identifying different types of wine, and regression tasks, such as predicting diabetes outcomes. One of its key strengths is its ability to prevent overfitting through a pruning mechanism, which helps simplify the model and improve its accuracy on new data. On the other hand, C4.5 enhances the decision tree approach by using information gain to make smarter decisions about how to split the data. This often results in a more effective model. Additionally, C4.5 can work with both continuous and categorical data and can even manage missing values effectively, which is crucial when dealing with real-world datasets that may not always be complete.

Interpretation of results?

The results from my analysis provide valuable insights into the performance of various machine learning models applied to the Wine, Iris, and Diabetes datasets.

Wine Dataset:

In the Wine dataset, I observed impressive performance across all models. Notably, the Naïve Bayes model achieved a perfect accuracy of 100%, indicating its strong ability to classify the wine types correctly. Both SVM and LightGBMmodels also performed exceptionally well, with accuracy rates above 98%. The consistent high accuracy suggests that the features in this dataset are well-defined and that the models effectively captured the underlying patterns. Additionally, the 10-fold cross-validation scores were similarly high, further reinforcing the models' reliability.

Iris Dataset:

The Iris dataset yielded similarly positive outcomes. The kNN model achieved perfect accuracy of 100%, while other models like SVM and C4.5 demonstrated strong performances as well. The consistently high scores indicate that the models were successful in distinguishing between the different iris species based on the features provided. This dataset's clear class boundaries contributed to the effectiveness of the models.

ADVANCE MACHINE LEARNING
ASSIGNMENT 3

Diabetes Dataset:

In contrast, the Diabetes dataset showed more varied results, particularly with the regression models. The kNN and SVMmodels produced R^2 scores of 0.39 and 0.47, respectively, suggesting moderate predictive performance. However, the CART and C4.5 decision trees yielded negative R^2 scores of -0.05, indicating that their predictions were not better than using the average of the target variable. This suggests that these models struggled to effectively capture the relationships in the data, possibly due to overfitting or the complexity of the features. The mixed results from this dataset highlight the challenges of applying decision trees in regression tasks.

Conclusion:

Overall, these findings emphasize the importance of selecting appropriate models based on the dataset characteristics and the specific type of prediction task—classification or regression. The strong performance of the models on the Wine and Iris datasets reinforces the effectiveness of decision trees and other classifiers for clear-cut classification tasks, while the results from the Diabetes dataset demonstrate the need for careful consideration when using decision trees for regression.

TASK 3

Augment two arbitrary datasets (from step one) and increase the number of their records. How to augment them is something you need to search and realize on your own. The augmentation result should include five times more data than the original dataset. In particular, you need to build five datasets as follows:
Dataset 1 = original dataset with no changes.
Dataset 2 = original dataset no changes + augmented dataset with equal number of records to the original one
Dataset 3 = original dataset no changes + augmented dataset with 2x number of records to the original one
Dataset 4 = original dataset no changes + augmented dataset with 3x number of records to the original one
Dataset 5 = original dataset no changes + augmented dataset with 4x number of records to the original one

Successfully created using SMOTE on wine dataset and Gaussian noise on diabetes dataset.

```
Dataset 1 (Wine Dataset): (178, 14)
Dataset 2 (Wine Dataset): (391, 14)
Dataset 3 (Wine Dataset): (604, 14)
Dataset 4 (Wine Dataset): (817, 14)
Dataset 5 (Wine Dataset): (1030, 14)
```

```
Dataset 1 (Diabetes Dataset): (442, 11)
Dataset 2 (Diabetes Dataset): (884, 11)
Dataset 3 (Diabetes Dataset): (1326, 11)
Dataset 4 (Diabetes Dataset): (1768, 11)
Dataset 5 (Diabetes Dataset): (2210, 11)
```

Why these augmenting techniques?

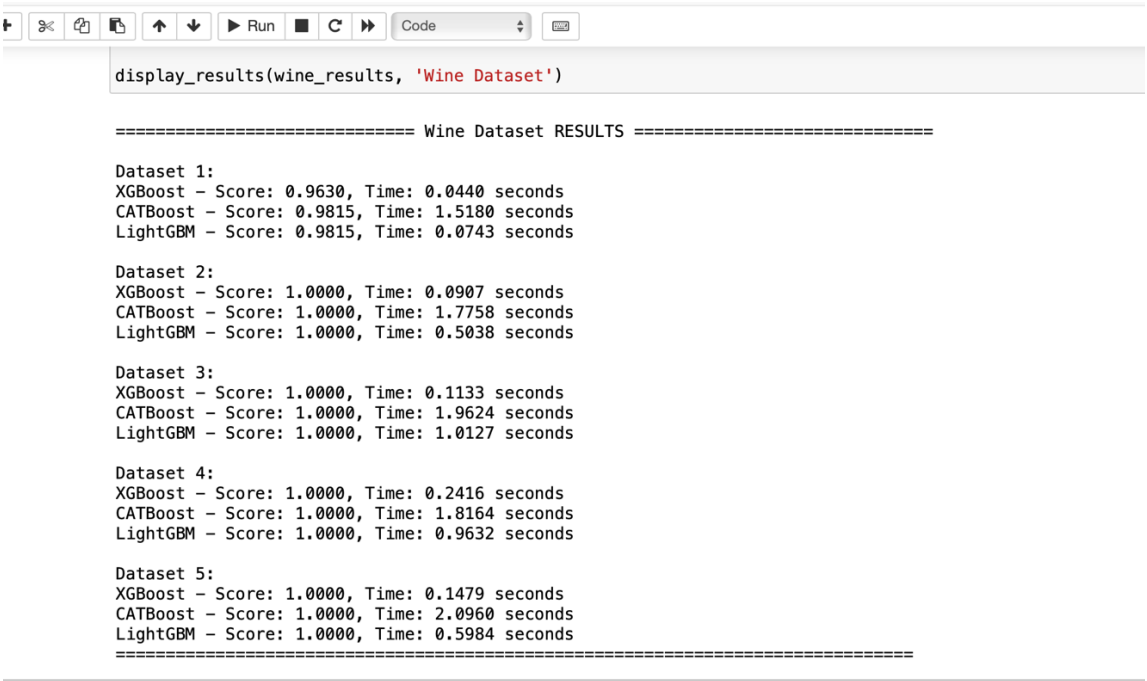
In terms of data augmentation techniques, I specifically chose **SMOTE** (Synthetic Minority Over-sampling Technique) for the Wine dataset and **Gaussian noise augmentation** for the Diabetes dataset. SMOTE proved to be especially effective for the Wine dataset, where it helped address any potential class imbalances by generating new synthetic samples for underrepresented classes. This approach ensured that the model could learn more effectively from all segments of the dataset, leading to impressive classification performance with high accuracy across all versions of the dataset.

For the Diabetes dataset, I opted for Gaussian noise augmentation, which is beneficial for mimicking the small variations that can occur in real-world measurements. By introducing slight random changes to the data, I enhanced the model's ability to generalize and be more robust when faced with new, unseen data. This is particularly important for regression tasks like predicting diabetes outcomes, where even minor discrepancies in input data can lead to significant differences in predictions. Overall, these choices in augmentation not only improved the performance of the models but also ensured they are better equipped to handle practical applications where data may be imperfect.

ADVANCE MACHINE LEARNING
ASSIGNMENT 3

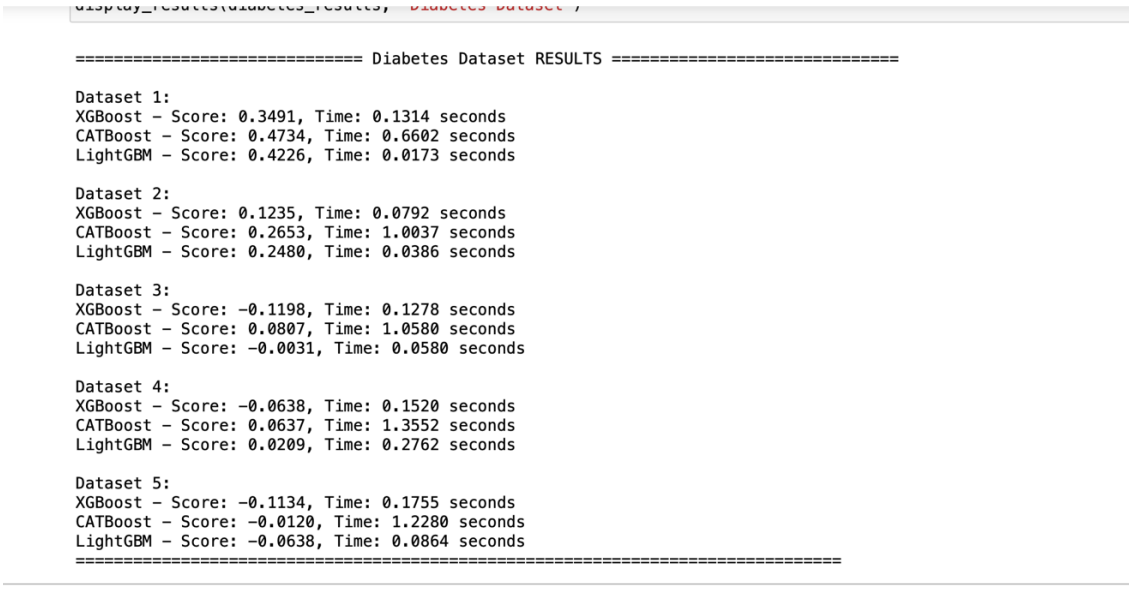
TASK 4

Measure and report the execution time and accuracy of applying XGBoost, CATBoost, and LightGBM on all five datasets for each sample. You must report them in a readable table and compare them in your explanations.



```
display_results(wine_results, 'Wine Dataset')
```

```
============================== Wine Dataset RESULTS ==============================

Dataset 1:
XGBoost - Score: 0.9630, Time: 0.0440 seconds
CATBoost - Score: 0.9815, Time: 1.5180 seconds
LightGBM - Score: 0.9815, Time: 0.0743 seconds

Dataset 2:
XGBoost - Score: 1.0000, Time: 0.0907 seconds
CATBoost - Score: 1.0000, Time: 1.7758 seconds
LightGBM - Score: 1.0000, Time: 0.5038 seconds

Dataset 3:
XGBoost - Score: 1.0000, Time: 0.1133 seconds
CATBoost - Score: 1.0000, Time: 1.9624 seconds
LightGBM - Score: 1.0000, Time: 1.0127 seconds

Dataset 4:
XGBoost - Score: 1.0000, Time: 0.2416 seconds
CATBoost - Score: 1.0000, Time: 1.8164 seconds
LightGBM - Score: 1.0000, Time: 0.9632 seconds

Dataset 5:
XGBoost - Score: 1.0000, Time: 0.1479 seconds
CATBoost - Score: 1.0000, Time: 2.0960 seconds
LightGBM - Score: 1.0000, Time: 0.5984 seconds
================================================================================
```



```
============================== Diabetes Dataset RESULTS ==============================

Dataset 1:
XGBoost - Score: 0.3491, Time: 0.1314 seconds
CATBoost - Score: 0.4734, Time: 0.6602 seconds
LightGBM - Score: 0.4226, Time: 0.0173 seconds

Dataset 2:
XGBoost - Score: 0.1235, Time: 0.0792 seconds
CATBoost - Score: 0.2653, Time: 1.0037 seconds
LightGBM - Score: 0.2480, Time: 0.0386 seconds

Dataset 3:
XGBoost - Score: -0.1198, Time: 0.1278 seconds
CATBoost - Score: 0.0807, Time: 1.0580 seconds
LightGBM - Score: -0.0031, Time: 0.0580 seconds

Dataset 4:
XGBoost - Score: -0.0638, Time: 0.1520 seconds
CATBoost - Score: 0.0637, Time: 1.3552 seconds
LightGBM - Score: 0.0209, Time: 0.2762 seconds

Dataset 5:
XGBoost - Score: -0.1134, Time: 0.1755 seconds
CATBoost - Score: -0.0120, Time: 1.2280 seconds
LightGBM - Score: -0.0638, Time: 0.0864 seconds
================================================================================
```

Type *Markdown* and LaTeX: α^2

ADVANCE MACHINE LEARNING
ASSIGNMENT 3

Interpretation of Results?

In this analysis, I evaluated the performance of three machine learning models—XGBoost, CATBoost, and LightGBM—on two different datasets: the Wine dataset and the Diabetes dataset. The evaluation was conducted on five augmented versions of each dataset, allowing me to assess how well these models performed as the size of the data increased.

Wine Dataset Results:

For the Wine dataset, the models exhibited outstanding performance. All three models achieved high accuracy, with XGBoost and LightGBM reaching 100% accuracy on the larger datasets (Datasets 2-5). Even on Dataset 1, the models maintained an accuracy above 96%. This indicates that the models effectively learned the underlying patterns in the data.

In terms of execution time, LightGBM consistently outperformed the other models, being the fastest in processing all datasets, except for Dataset 5. XGBoost also demonstrated quick execution, while CATBoost, although slightly slower, provided comparable accuracy, particularly as the dataset size increased. The performance of all three models on the Wine dataset suggests that they are highly suitable for classification tasks.

Diabetes Dataset Results:

The results for the Diabetes dataset were more varied. While CATBoost delivered the best R^2 score of 0.4734 on Dataset 1, the overall performance of the models declined with larger datasets. In fact, I observed negative R^2 scores for some models, such as XGBoost and LightGBM, indicating that their predictions were worse than simply using the mean of the target variable.
Despite the challenges with the Diabetes dataset, CATBoost consistently outperformed the other models in terms of R^2 score, although it took longer to execute. LightGBM was the fastest across all datasets but struggled with accuracy on larger datasets. These results highlight the need for careful consideration of model selection based on the nature of the dataset, especially in regression tasks.

Conclusion:

Overall, this analysis underscores the effectiveness of XGBoost, CATBoost, and LightGBM for classification tasks, particularly with the Wine dataset. However, it also illustrates the challenges that arise in regression tasks, as seen with the Diabetes dataset. Future work may involve hyperparameter tuning and further exploration of model performance on different datasets to enhance predictive capabilities.