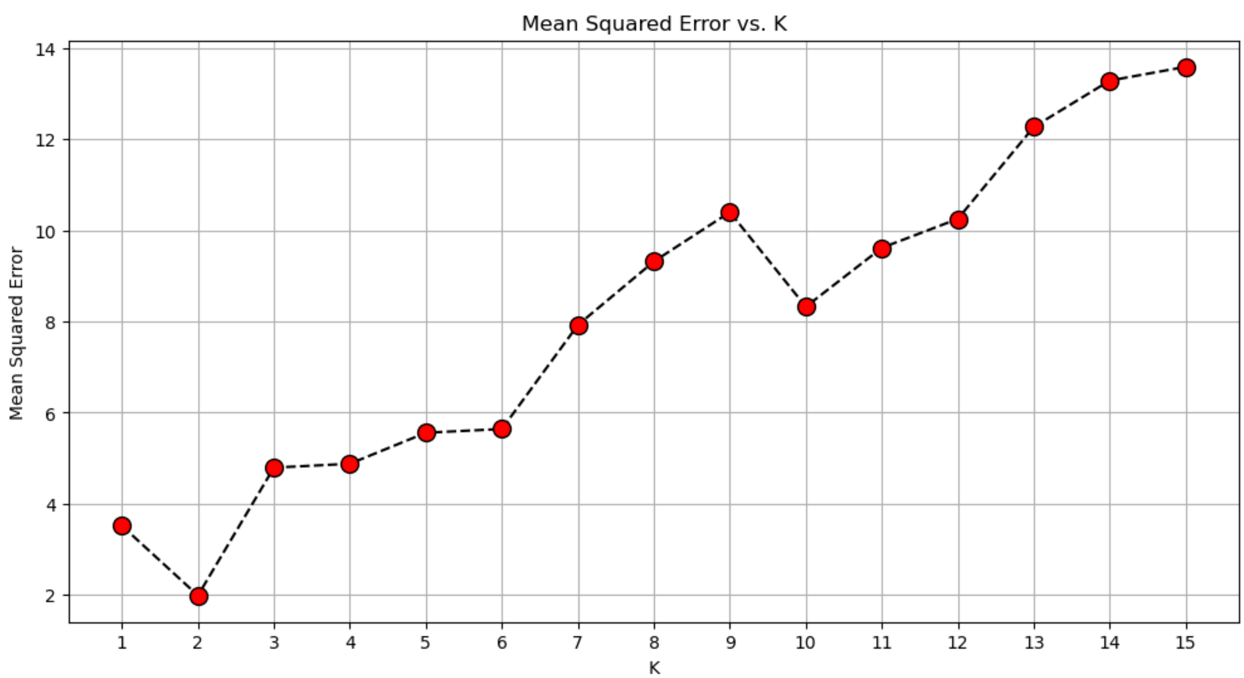
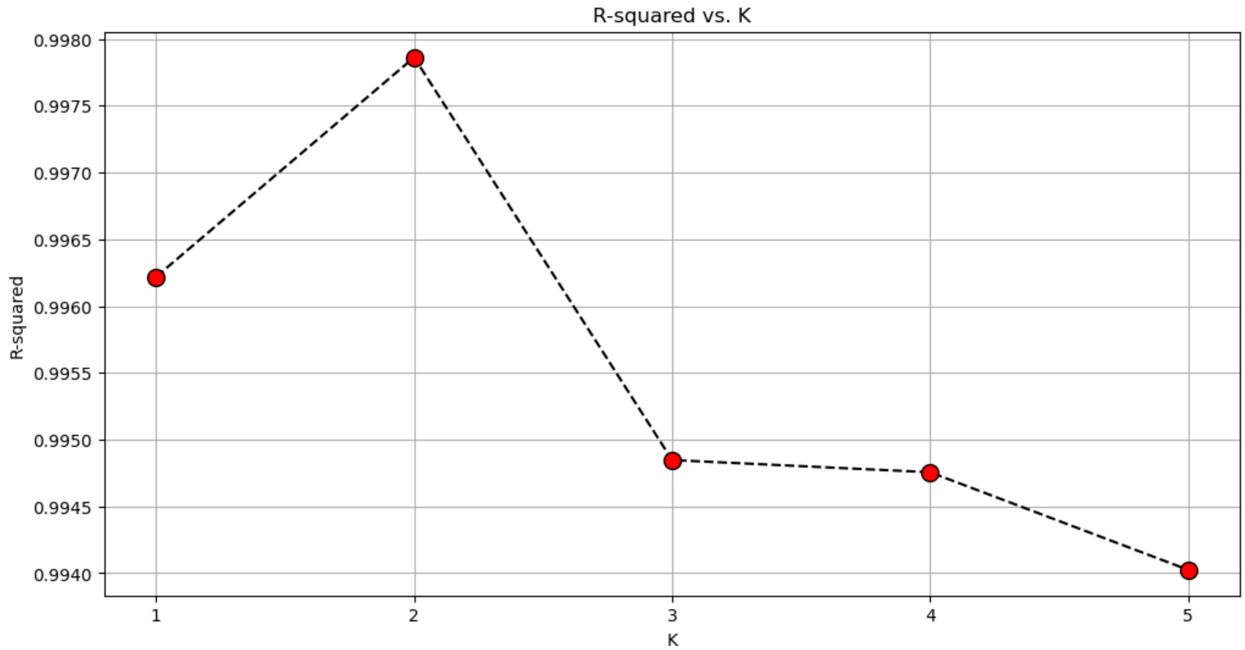
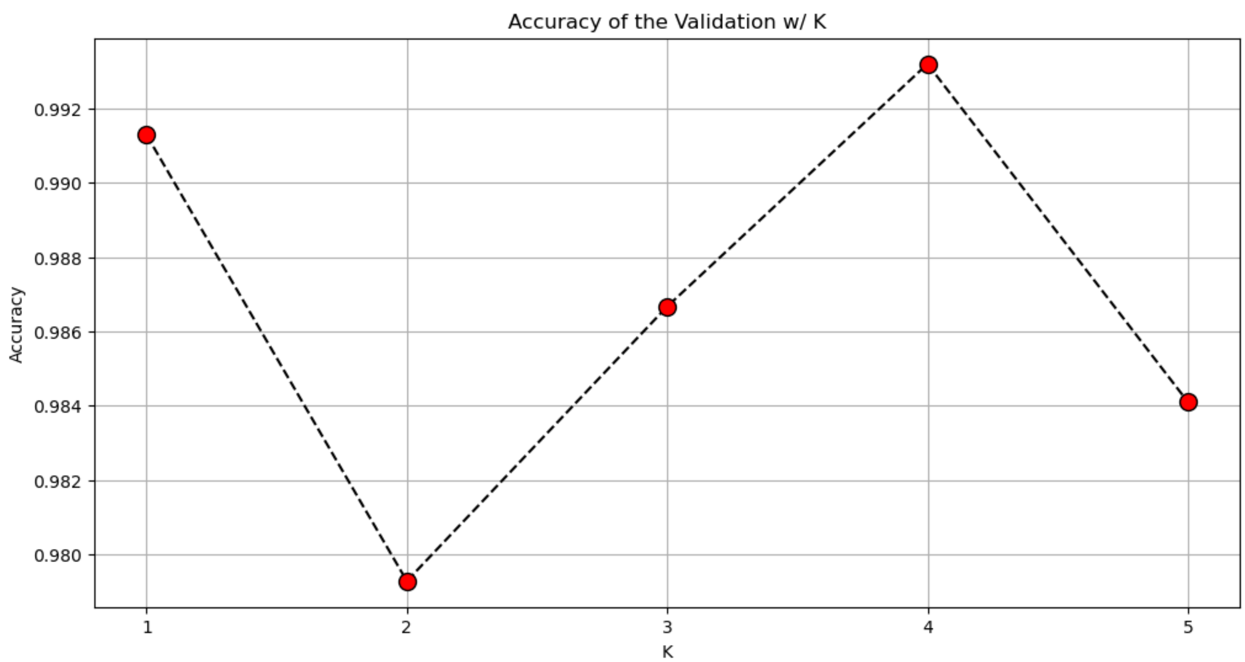
Main Objective:

For this part of the project, I will investigate the relationship between stock market performance and the progress of COVID-19 vaccinations by focusing on the predictive capabilities of machine learning algorithms. For that, I will utilize the S&P_500_Price and Nasdaq_100_Price columns as features and PopulationVaccinated column as the label to explore how the percentage of vaccinated individuals influences market indices. I plan to first use K-Nearest Neighbors (KNN) model, then Decision Tree model, and at the end I will compare the given the results and decide which one is the most compatible method for my dataset. I will verify the models by dividing the merged dataset into training (80%), validation (10%), and testing (10%) sets to ensure robust evaluation and minimize overfitting.

KNN Graphs:



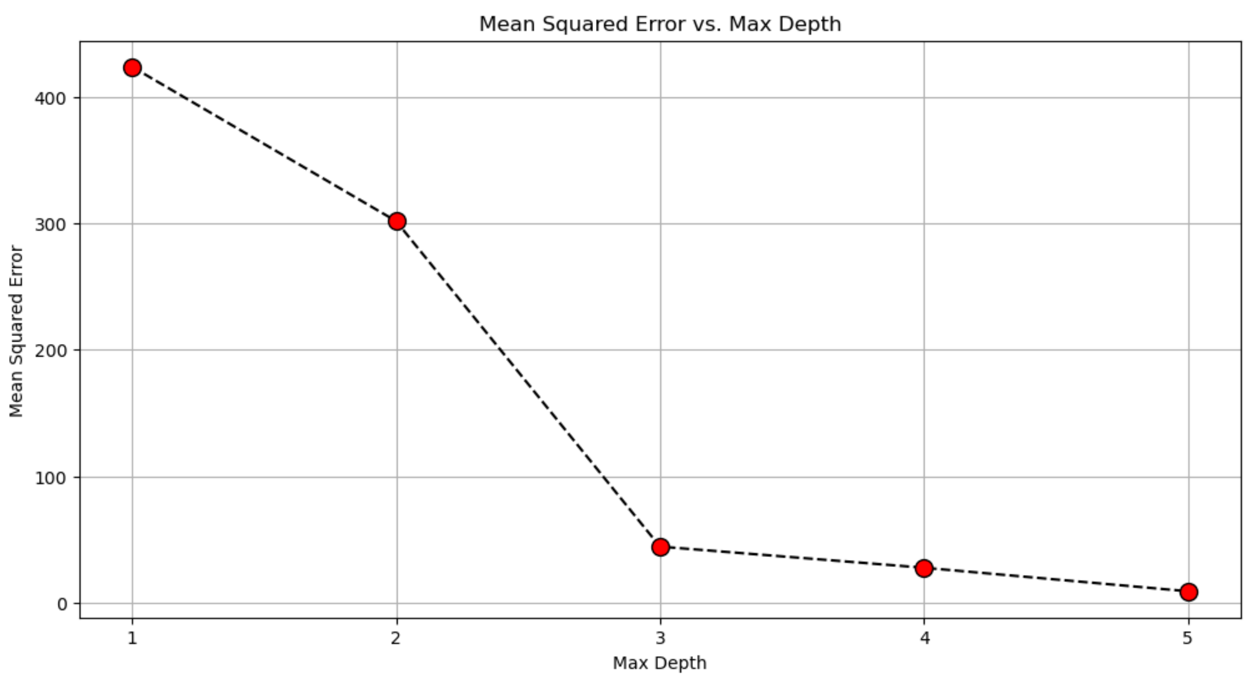

R-squared vs. K
0.9980
0.9975
0.9970
0.9965
0.9960
0.9955
0.9950
0.9945
0.9940
R-squared
1
2
3
4
5
K

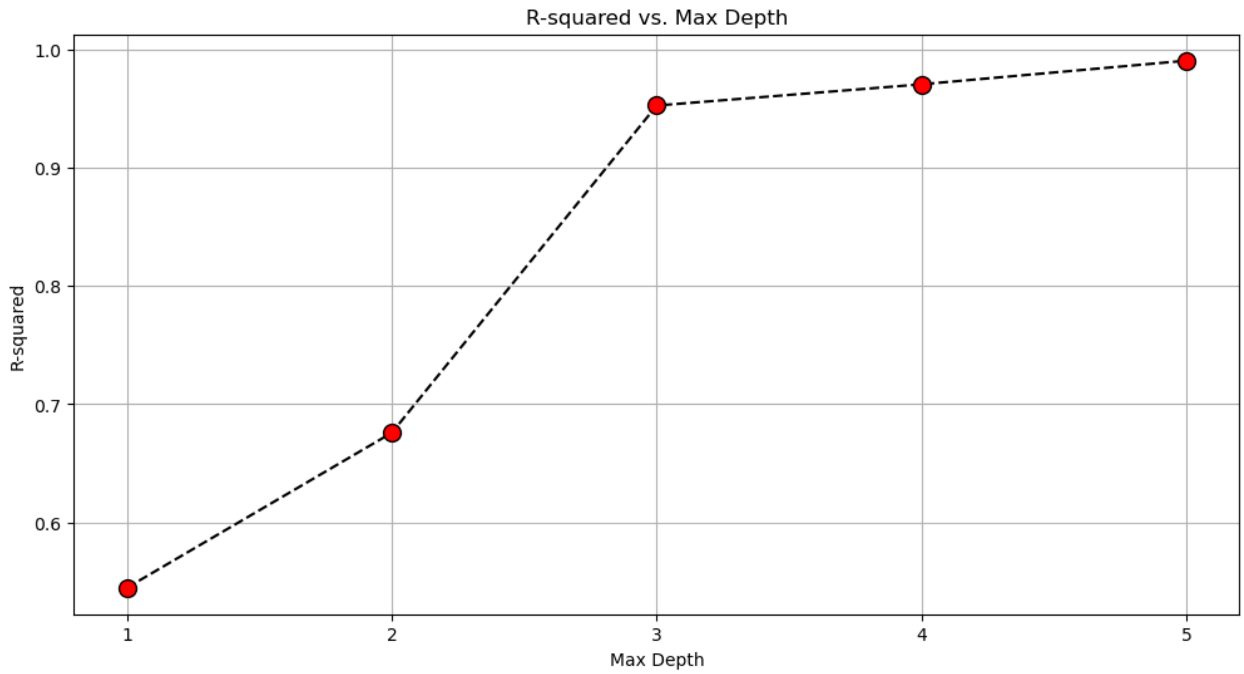

Accuracy of the Validation w/ K
0.992
0.990
0.988
0.986
0.984
0.982
0.980
Accuracy
1
2
3
4
5
K

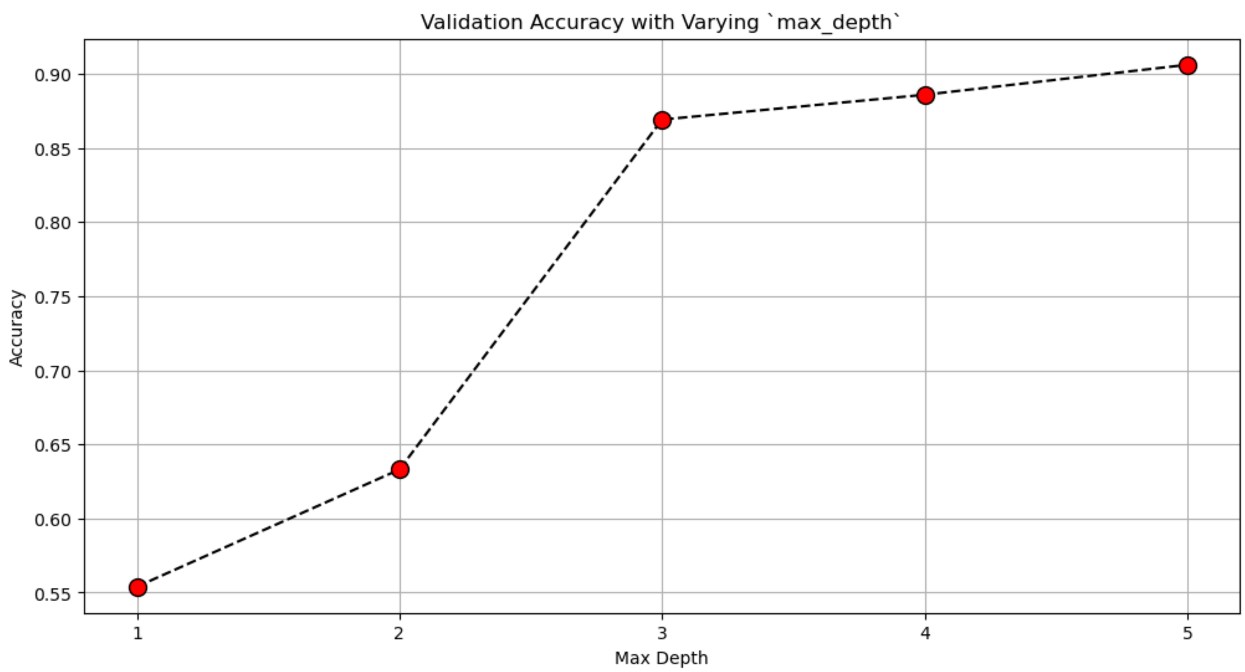
KNN Report:

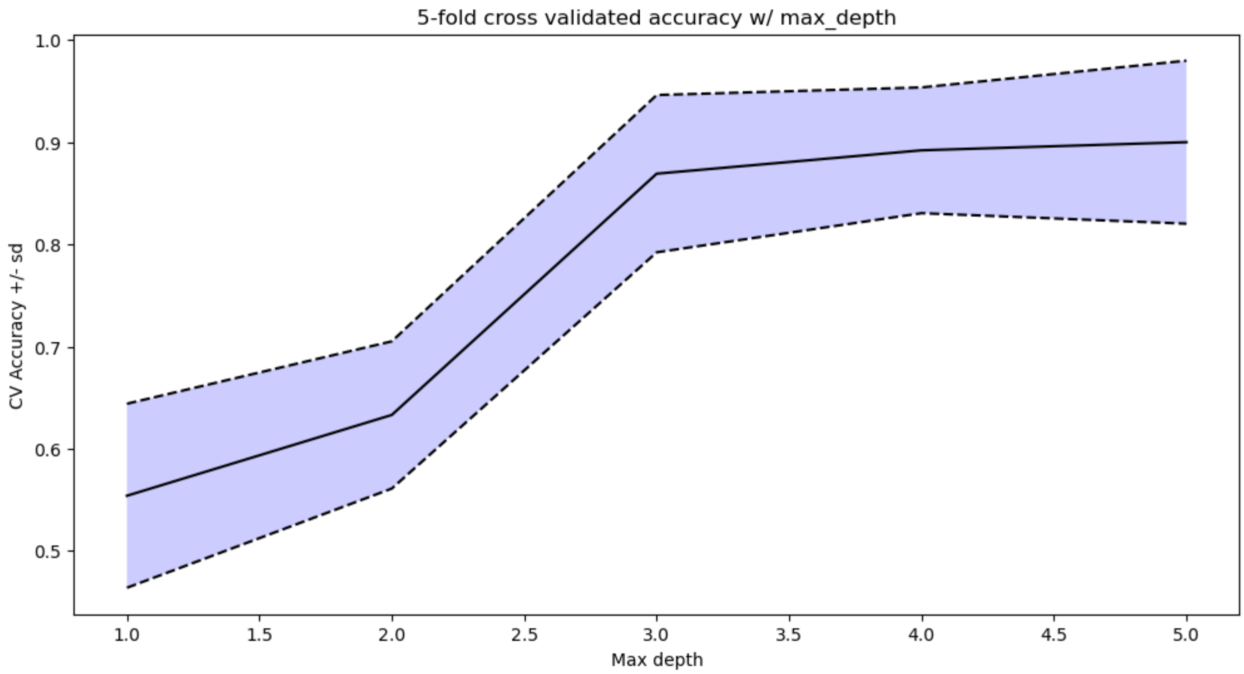
The K-Nearest Neighbors (KNN) model demonstrates significant compatibility with the dataset, as evidenced by its performance across multiple metrics. From the analysis, KNN's pros include its simplicity and the intuitive nature of its tunability through the number of neighbors (K), allowing for flexible adjustment to optimize for different performance measures such as Mean Squared Error, R-squared, and accuracy. It shows particularly strong performance at lower K values (between K=2 and K=4), where it achieves low MSE, high R-squared, and high accuracy, indicating effective balance and generalization capabilities (Best Features). However, the cons include its sensitivity to the scale of the data and irrelevant features, which can degrade its performance if not properly managed. The provided graphs collectively suggest that while KNN is generally effective, its performance can vary significantly with changes in K, emphasizing the need for careful parameter tuning and consideration of metric priorities in model evaluation and selection.

Decision Tree Graphs:




R-squared vs. Max Depth
R-squared
1.0
0.9
0.8
0.7
0.6
1
2
3
4
5
Max Depth


Validation Accuracy with Varying `max_depth`
Accuracy
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
1
2
3
4
5
Max Depth



Decision Tree Report:

The Decision Tree model exhibits considerable promise for the dataset as demonstrated by the analysis of the provided graphs, which show changes in Mean Squared Error (MSE), R-squared, validation accuracy, and 5-fold cross-validated accuracy across varying maximum depths. The MSE graph illustrates a sharp decline in error as the maximum depth increases from 1 to 3, suggesting that the model becomes significantly better at capturing the underlying patterns with a moderate depth, before plateauing from depth 3 to 5. This plateauing indicates a balance point where increasing depth does not introduce overfitting, maintaining model simplicity while retaining predictive power. The R-squared graph supports this, showing a steep increase in explanatory power, which levels off at a maximum depth of 3, maintaining high values through depth 5. The validation accuracy and cross-validated accuracy graphs both show an upward trend as the depth increases, reaching and sustaining high levels of accuracy, indicating robustness and generalizability of the model on unseen data. The 5-fold cross-validation graph, in particular, shows a band of variance that narrows and stabilizes at higher depths, suggesting the model's consistency across different subsets of the data. Collectively, these metrics point to a Decision Tree model that is well-tuned to this dataset with a maximum depth between 3 and 5 (Best Features), providing an excellent balance between complexity and performance, making it highly compatible and effective for the data at hand.

Comparison:

Mean Squared Error (MSE): KNN consistently shows a lower MSE than the Decision Tree model. On the test set, KNN has an MSE of approximately 52.66 compared to DT's MSE of about 97.48. Additionally, the average MSE over multiple evaluations is 32.17 for KNN and 76.93 for DT. This indicates that KNN is better at minimizing prediction errors on both the specific test dataset and on average, suggesting higher precision and less error in predictions.

R-squared (R2): KNN also exhibits superior performance in terms of the coefficient of determination, R2, which measures the proportion of variance in the dependent variable that is predictable from the independent variables. KNN's test R2 score is 0.9377, significantly higher than DT's 0.8847. Similarly, the average R2 for KNN is 0.9645 compared to DT's 0.9147. Higher R2 values for KNN suggest that it more effectively captures the variance in the dataset and can explain more of the variability in response data.

Accuracy: KNN's peak accuracy is notably higher than that of the DT model, suggesting that KNN may be better suited for this dataset when configured optimally.

Conclusion:

KNN consistently outperforms the Decision Tree model across all of the indicators, so it’s a more compatible machine learning method for my dataset.