Slide 1: Title
Hello, everyone. Today, we're diving into the world of applying data science to predict stock prices. Let's explore how we can leverage machine learning techniques to unlock insights into the financial markets.

Slide 2: Agenda
In today's presentation, we will go through five parts which are Data Science Flow, Problem Definition, What you can learn, Implementation and Conclusion respectively.

Slide 3: Data Science Flow
Before we delve into the details, let's briefly outline the flow of our data science project. It starts with data collection, followed by modeling, and then evaluation.

Slide 4: Problem Definition
Let's begin by outlining the main challenge we're addressing in our project.

First, the problem statement: Accurately predicting stock prices is essential for investors and traders to make informed decisions. However, the ever-changing nature of financial markets presents significant hurdles to achieving this.

Second, our objectives: We aim to develop a robust model that can predict stock prices based on historical data. This will provide valuable insights, empowering investors to make more informed and strategic trading decisions.

Slide 5: What you can learn
Now, let's discuss what you can learn from our project:

As we delve into the challenge of predicting stock prices, you'll gain insight into the complexities of financial markets and the importance of accurate predictions for investors and traders.

Throughout this presentation, you'll see how data science techniques, especially machine learning, are applied to analyze historical data and forecast future stock prices.

We'll walk you through the process of building predictive models, from data collection and preprocessing to model evaluation. This will give you a clear understanding of the steps involved in developing robust models for financial forecasting.

Furthermore, we'll explore various evaluation techniques such as Mean Squared Error, T-test, Wilcoxon-test, and Feature Importance. Understanding these methods will enable you to assess model performance and interpret results effectively.

Finally, we'll demonstrate how these techniques can be applied in real-world scenarios, providing practical insights into financial market trends and empowering decision-making.

By the end of this presentation, you'll have a comprehensive understanding of how data science can be leveraged to analyze and predict stock prices, equipping you with valuable skills for navigating the complexities of financial markets.

Slide 6: Implementation
With a clear understanding of our objectives and learning outcomes, let's discuss how we will implement these concepts in our project.

Slide 7: Data Collection
Step 1, Data Collection. We'll start by collecting historical stock price data using the Yahoo Finance API. This will serve as the foundation for our analysis and predictive modeling efforts. In the project, we take TSMC32.SA for example.

Slide 8: Data Analysis
Step 2, Data Analysis. We'll analyze the data to gain insights into stock price movements and volume trends. Visualization techniques will help us identify patterns and correlations that will inform our modeling process.

Slide 9: Model Building
Step 3, Model Building. We'll split the data into training and testing sets and construct a Random Forest Regressor model. It works by constructing multiple decision trees during training and outputting the average prediction of the individual trees. This ensemble approach helps to reduce overfitting and improve the model's robustness.
By leveraging the collective intelligence of multiple decision trees, we aim to capture complex patterns in the data and make accurate predictions of stock prices.

To optimize the model, GridSearchCV finds optimal hyperparameters by evaluating a model across various hyperparameter combinations. It searches over a defined grid of hyperparameters, aiming to enhance the model's predictive performance.

Slide 10: Model Building
This visualization helps to see how well the model predictions align with the actual values. If the points are closely clustered around the diagonal line, it indicates good predictions. If they are scattered far from the line, it suggests the model's predictions are not as accurate. As you can see, most of the points are closely clustered around the diagonal line. Therefore, it indicates that our model is quite accurate.

Slide 11: Model Evaluation (1)
After training our model, we'll evaluate its performance using metrics such as Mean Squared Error also called as MSE, which calculates the average squared difference between the actual and predicted values of the target variable. A lower MSE suggests that our model's predictions are closer to the actual values, indicating higher accuracy and effectiveness. If the MSE is higher than expected, it could indicate issues such as underfitting or overfitting. Underfitting occurs when the model is too simple to capture the underlying structure of the data, leading to high bias. Overfitting, on the other hand, happens when the model is too complex and captures noise in the training data, leading to high variance. With an MSE of 0.54, our model's predictions are quite accurate. This step is crucial for validating the effectiveness of our model and identifying areas for improvement.

Slide 12: Model Evaluation (2)
The second model evaluation techniques are T-test and Wilcoxon-test. Both tests assess whether there's a significant difference between actual and predicted prices.

In the T-test and the Wilcoxon-test, since the p-value is greater than the significance level, we fail to reject the null hypothesis. Therefore, there is no significant difference between actual and predicted prices. Both tests suggest that the model's predictions are statistically similar to the actual prices, with no significant difference detected.

Given that the data is typically not normally distributed in finance because prices often follow a skewed distribution, and the sample size might not be large enough, the Wilcoxon test, being non-parametric and making fewer assumptions, is often more suitable for hypothesis testing in this context. Therefore, in this scenario, the Wilcoxon test would be more appropriate.

Slide 13: Model Evaluation (3)

The third model evaluation method is Feature Importance which based on mean decrease in impurity measures the ability of each feature to differentiate between classes in a Random Forest model. The blue bars represent feature importances, with error bars showing inter-tree variability. Higher bars indicate more important features for predicting the target variable. Upon examining the graph, we observe that the top three important features are opening price, high price, and low price. These features play a significant role in the model's decision-making process, suggesting that variations in these prices are strongly associated with changes in the target variable. Understanding feature importance not only helps us comprehend which variables are crucial for predicting stock prices but also provides insights into the underlying relationships between these variables and the target.

Slide 14: Prediction

Finally, we'll demonstrate how our trained model can predict stock prices for the next day based on the latest data. This real-world application showcases the practical value of our data science techniques in financial markets.

Slide 15: Conclusion

In conclusion, our journey through applying data science to predict stock prices has been illuminating. We've learned how machine learning techniques can provide valuable insights into financial markets. From collecting and analyzing data to building and evaluating models, we've demonstrated the power of data science in empowering investors to make informed decisions. Through techniques like MSE, T-test, Wilcoxon-test, and Feature Importance, we've validated our models and identified key features driving predictions. Finally, with our trained model predicting stock prices, we've showcased the practical value of data science in financial markets. This is the end of our presentation. Thank you!