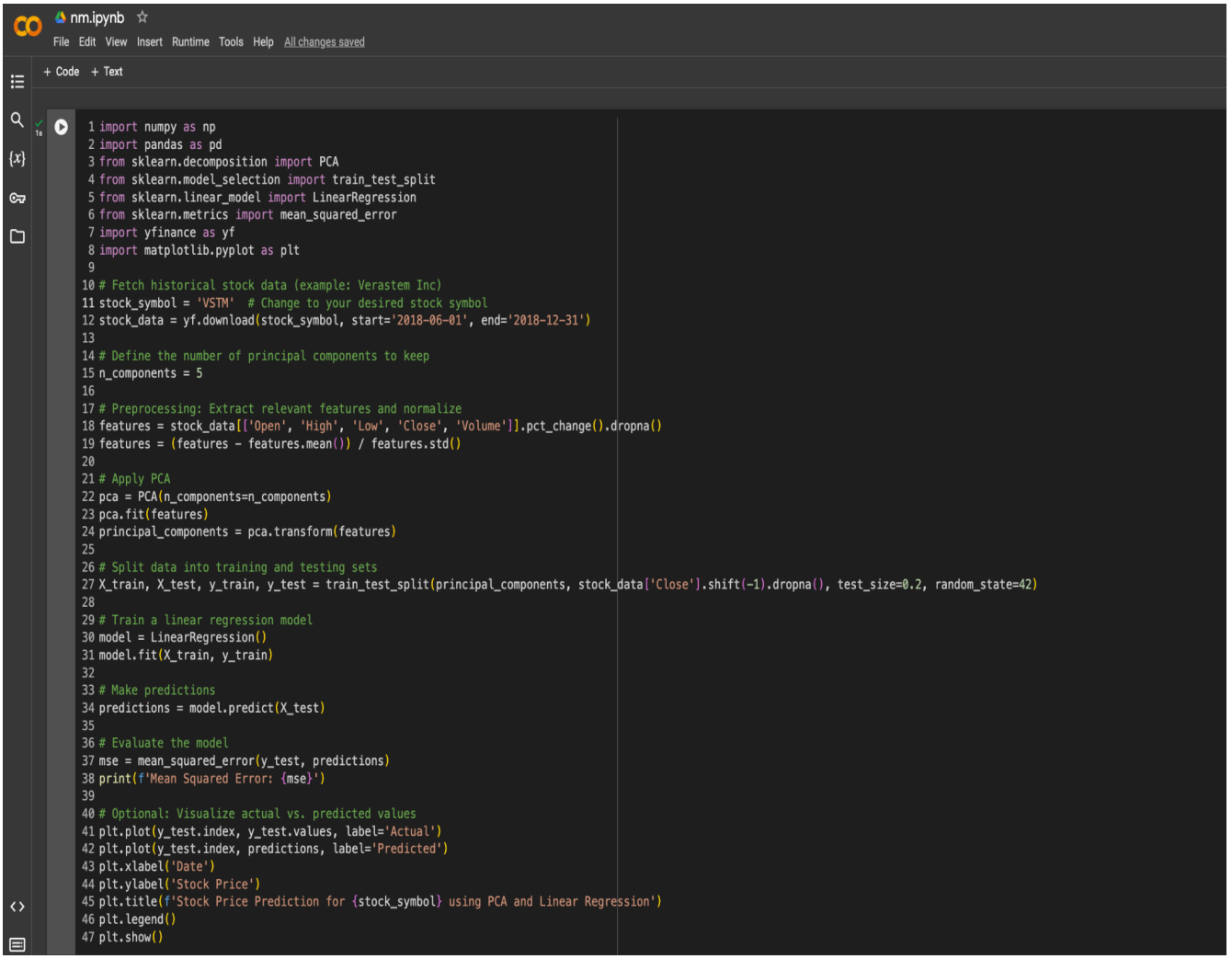
Codes Used in the Project

Team Members

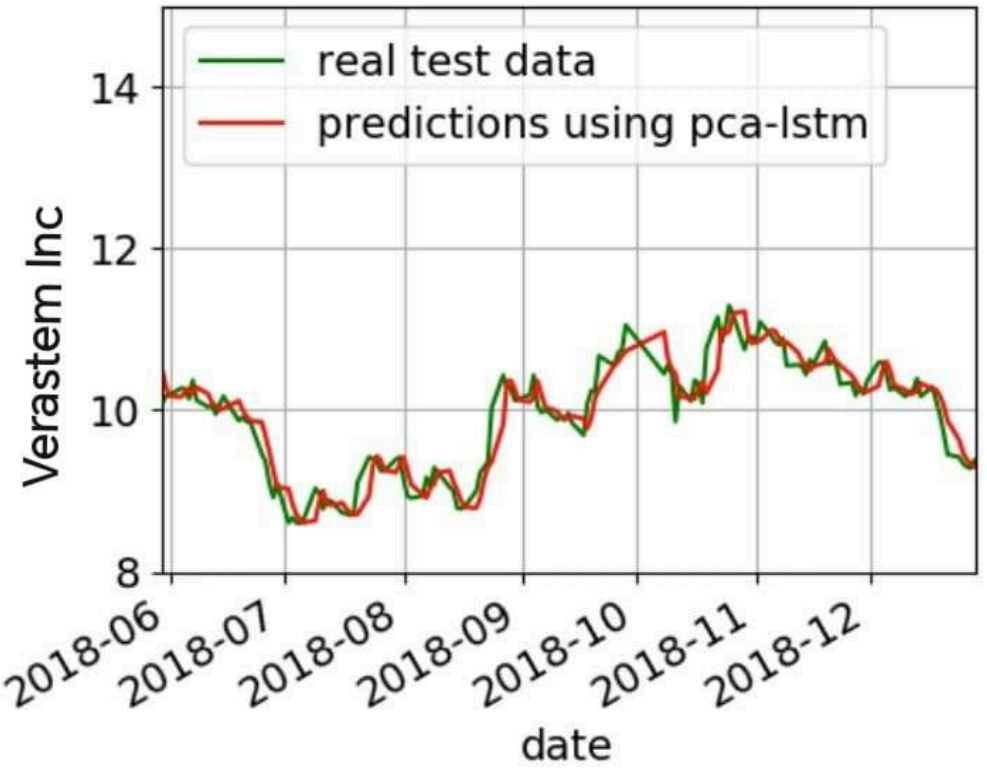
Ujjwal Raj (ME22B1072)
Aman Gupta (ME22B1071)
Prashant Tyagi (ME22B1069)
Vinay (ME22B1070)

Python Code of the stock price prediction




```python
import numpy as np
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import yfinance as yf
import matplotlib.pyplot as plt

# Fetch historical stock data (example: Verastem Inc)
stock_symbol = 'VSTM'  # Change to your desired stock symbol
stock_data = yf.download(stock_symbol, start='2018-06-01', end='2018-12-31')

# Define the number of principal components to keep
n_components = 5

# Preprocessing: Extract relevant features and normalize
features = stock_data[['Open', 'High', 'Low', 'Close', 'Volume']].pct_change().dropna()
features = (features - features.mean()) / features.std()

# Apply PCA
pca = PCA(n_components=n_components)
pca.fit(features)
principal_components = pca.transform(features)

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(principal_components, stock_data['Close'].shift(-1).dropna(), test_size=0.2, random_state=42)

# Train a linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
predictions = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, predictions)
print(f'Mean Squared Error: {mse}')

# Optional: Visualize actual vs. predicted values
plt.plot(y_test.index, y_test.values, label='Actual')
plt.plot(y_test.index, predictions, label='Predicted')
plt.xlabel('Date')
plt.ylabel('Stock Price')
plt.title(f'Stock Price Prediction for {stock_symbol} using PCA and Linear Regression')
plt.legend()
plt.show()
```

Output of the code for Verastem Inc Dataset



Latex Code of the Report

```
\documentclass{article}
\usepackage{graphicx}
\usepackage{lipsum} % For generating placeholder text

\title{\textbf{\centering{ \Huge Report on \\Stock Price Prediction using Principal Component Analysis}}}

\date{}

\begin{document}

\maketitle

\vfill
```

\begin{center}
\textbf{\huge Team Members}\\
{\large Ujjwal Raj(ME22B1072)\\
Aman Gupta(ME22B1071)\\ Prashant Tyagi(ME22B1069)\\
Vinay(ME22B1070)}
\end{center}

\newpage
\section{Abstract}
\Large {The categorization of high-dimensional data presents a fascinating challenge to machine learning models, as the frequent presence of highly correlated dimensions or attributes can affect the accuracy of classification models. In this paper, we investigate the problem of high dimensionality in stock exchange data to predict market trends by applying Principal Component Analysis (PCA) with linear regression. PCA can help improve the predictive performance of machine learning methods while reducing redundancy in the data. We conduct experiments on high-dimensional spectral data from three stock exchanges: New York Stock Exchange, London Stock Exchange, and Karachi Stock Exchange. We compare the accuracy of linear regression classification models before and after applying PCA. The experiments demonstrate that PCA can enhance the performance of machine learning models, provided that the relative correlation among input features is carefully investigated, and principal components are selected judiciously. We use Root Mean Square Error (RMSE) as an evaluation metric to assess the classification model.}
\section{Introduction}
Prediction is a process to make assumption of future based on existing data. The more precise the prediction, the more it could be easier to make decision for future.
Prediction of stock exchange trends has been an interesting topic in the field of pattern recognition and machine learning because of its possible monetary profit. A stock market is an organized set-up with a regulatory body and has registered members who can buy or sell shares. It’s a public market, where different companies invest high capital and do trading of their shares. Stock market prediction provides information about stock market, which can help the shareholders to make decision about trading. It may serve as warning system for long-term shareholders while for short-term investors may serve as recommender system.

One of the problems with Stock market data is that the available data is highly volatile with very high dimensions. Many of the attributes are highly correlated and makes prediction of stock market a highly challenging, complicated and daunting task. The best solution for above problem is to reduce dimensions of data. Measuring correlation between data dimensions can do reduction and discarding those attributes, or dimensions which have least impact on overall prediction model. Principal component analysis is one of effective technique, which can be applied to reduce the dimensionality of data.
In this paper, we have predicted the trend of Verastem Inc exchanges by using linear regression as a classification model. The past values are used to build a classification function and then uses this function to predict about future values. Principal component analysis is used with linear regression model to check that whether PCA has improved the accuracy of model or not.

\section{Problem Definition}
Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction. In this project, our objective is to harness the power of PCA to predict stock prices. By reducing the dimensionality of the data, we aim to extract meaningful patterns and improve the accuracy of our predictions.

\section{Motivation}
Predicting even slight fluctuations in stock movements can translate into substantial financial gains. Yet, grappling with voluminous datasets teeming with diverse features presents a formidable obstacle. The curse of dimensionality further complicates this scenario by inundating the analysis with redundancy and noise, impeding accurate predictions. However, Principal Component Analysis (PCA) emerges as a robust antidote to this predicament. By distilling complex datasets into more manageable dimensions, PCA effectively mitigates the curse of dimensionality while preserving essential information critical for accurate forecasting. Through dimensionality reduction, PCA streamlines the analysis process, facilitating clearer insights into stock market trends. Thus, PCA stands as a powerful tool in the arsenal of predictive analytics, enabling investors to make informed decisions amid the intricacies of financial markets.

\section{Approach}
To combat the curse of dimensionality, we employ two primary strategies: feature selection and feature extraction (dimensionality reduction). Feature selection involves identifying the most influential features, thereby enhancing model interpretability and performance. Dimensionality reduction techniques, such as PCA, mitigate the curse of dimensionality by transforming the data into a lower-dimensional space. This not only improves computational efficiency but also facilitates data visualization and understanding.

\section{Example: Housing Dataset}
Consider a scenario where a homebuyer must evaluate numerous features before making a purchase decision, such as proximity to amenities, neighborhood safety, and property size. Despite the abundance of factors, the ultimate decision often revolves around a few critical features. Similarly, in stock prediction, PCA helps identify key factors amidst a multitude of variables, simplifying the prediction process and improving decision-making.

\section{Mathematical Method: Eigen Decomposition}
Eigen decomposition is a fundamental concept in PCA. It involves decomposing a covariance matrix into its constituent eigenvalues and eigenvectors. These eigenvectors represent the principal components, with each capturing a certain amount of variance in the data. By retaining the principal components with the highest variance, we effectively reduce the dimensionality of the dataset while preserving the most significant information. Moreover, the concepts of linear regression seamlessly integrate with PCA, enabling us to model stock price trends accurately.

\section{Mathematical Steps In PCA}

\subsection{Standardize the Data}
First, we standardize the data to have a mean of 0 and a standard deviation of 1. This is given by:
\[
X_{\text{std}} = \frac{X - \mu}{\sigma}
\]
where \(X\) is the original feature vector, \(\mu\) is the mean of the feature vector, and \(\sigma\) is the standard deviation of the feature vector.

\subsection{Compute the Covariance Matrix}

Next, we compute the covariance matrix of the standardized data. The covariance matrix is given by:
$$
\Sigma = \frac{1}{n-1}(X_{\text{std}} - \mu)(X_{\text{std}} - \mu)^T
$$
where n is the number of observations.

Compute the Eigenvalues and Eigenvectors

We then compute the eigenvalues and eigenvectors of the covariance matrix. The eigenvector v and eigenvalue λ of a matrix A satisfy the equation:
$$
Av = \lambda v
$$
The eigenvectors represent the directions or components for the reduced subspace of X_{std}, while the eigenvalues represent the magnitudes for the directions.

Sort Eigenvalues and their Corresponding Eigenvectors

Next, we sort the eigenvalues in decreasing order and choose the first k eigenvectors, which correspond to the k largest eigenvalues, where k is the number of dimensions of the new feature subspace ($k \leq d$).

Transform the Original Dataset

Finally, we transform the original standardized dataset via the selected eigenvectors to obtain a k-dimensional feature subspace.
$$
Y = X_{\text{std}} \cdot W
$$
where Y is the transformed k-dimensional feature subspace and W is the matrix of the top k eigenvectors.

Finding Principal Components with Max Variance

To identify the principal components that capture the maximum variance in the data, we compute the eigenvalues associated with each eigenvector. The eigenvector corresponding to the highest eigenvalue represents the principal component with the most significant variance. By projecting all components onto this principal component, we effectively reduce the dimensionality of the dataset while retaining the most critical information. This process transforms

the data into a lower-dimensional space, facilitating more efficient analysis and prediction.

```
\section{Stock Price Prediction with PCA}
\begin{center}
  \includegraphics[width=1.5 \textwidth ]{code1.png}
  \includegraphics[width=1.5 \textwidth ]{graph1.jpeg}
\end{center}

\section{Dataset Explanation}
```

The code starts by fetching historical stock data for Verastem Inc(an American pharmaceutical company that develops medicines to treat certain cancers) using the yfinance library. It then preprocesses the data by calculating the percentage change for the ‘Open’, ‘High’, ‘Low’, ‘Close’, and ‘Volume’ features and standardizing these features to have a mean of 0 and a standard deviation of 1. The code then applies Principal Component Analysis (PCA) to reduce the dimensionality of the data, keeping only the top 5 principal components. The data is then split into training and testing sets, and a linear regression model is trained on the training set.

```
\section{Stock Price Prediction and Evaluation}
```

The trained model predicts stock prices. Its performance is evaluated using Mean Squared Error and visualized by plotting actual vs predicted prices.

```
\section{Summarizing the Project}
```

We have successfully explored and demonstrated the efficacy of PCA in predicting stock prices, highlighting its role in mitigating dimensionality issues. We believe that our work can serve as a foundation for future research in this area.

```
\section*{Thank You}
```

We extend our sincere gratitude for the opportunity to undertake this project and for your support and guidance throughout the process.

```
\end{document}
```

Latex code for the presentation

```
\documentclass{beamer}
```

```
\usetheme{Madrid}
\usepackage{graphicx}
\usepackage{amsmath}

\title{Stock Price Prediction using Principal Component Analysis}
\author{Ujjwal Raj (ME22B1072)\\ Aman Gupta (ME22B1071)\\ Prashant Tyagi (ME22B1069)\\ Vinay Kumar (ME22B1070)}
\date{\today}

\begin{document}

\frame{\titlepage}

\begin{frame}{Problem Definition}
\begin{block}{Understanding Principal Component Analysis (PCA)}
Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of data while retaining most of its variance. It does this by transforming the data to a new coordinate system in which the greatest variance lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.
\end{block}

\begin{block}{Our Objective}
Our objective in this project is to leverage PCA for predicting stock prices. Predicting stock prices is a crucial task in financial markets and can lead to significant financial gains if done accurately.
\end{block}
\end{frame}

\begin{frame}{Motivation}
\begin{block}{Significance of Predicting Stock Trends}
Even slight insights into stock movement can lead to significant financial gains, making accurate prediction highly desirable. However, predicting stock prices is a complex task due to the high volatility and uncertainty in the stock market.
\end{block}

\begin{block}{Why PCA?}
PCA is chosen due to its effectiveness in mitigating the curse of dimensionality often encountered in stock data analysis. The curse of dimensionality refers to
```

the exponential increase in volume associated with adding extra dimensions to Euclidean space, which can lead to problems in machine learning and data analysis.
\end{block}
\end{frame}

\begin{frame}{Dealing with Curse of Dimensionality}
\begin{block}{There are two ways to remove the curse of dimensionality}
\end{block}

\begin{block}{Feature Selection}
It is the process which help us understand the most important feature.
\end{block}

\begin{block}{Dimensionality Reduction}
Dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. Techniques like PCA are used for this purpose. By reducing the dimensionality, we can improve model performance and can easily understand and visulize the data.
\end{block}
\end{frame}

\begin{frame}{Feature Reduction Example: Housing Dataset}
\begin{block}{Analogy with Housing Features}
Consider buying a house with numerous attributes such as proximity to amenities, number of rooms, size of the house, etc. Despite the abundance of features, the decision often boils down to a few critical factors such as location, price, and size. This is similar to feature reduction.
\end{block}

\begin{block}{Application in Stock Prediction}
In the context of stock prediction, PCA helps identify key factors amidst a multitude of variables, simplifying the prediction process. For example, instead of considering all the features of a stock such as its previous prices, volume, etc., we can use PCA to find a smaller set of features that capture the most variance in the data.Like PE Ratio, Inflation, Consumer Sentiment etc
\begin{figure}
    \centering
    \includegraphics[width=0.5\linewidth]{graph1.png}

```
    \caption{Enter Caption}
    \label{fig:enter-label}
\end{figure}
\end{block}
\end{frame}

\begin{frame}{Mathematical Method: Eigen Decomposition}
\begin{block}{Understanding Eigen Decomposition}
Eigen Decomposition is a type of matrix decomposition that plays a pivotal role in PCA. It involves decomposing a square matrix into a set of eigenvectors and eigenvalues. The eigenvectors represent the directions of the new space, and the eigenvalues represent the magnitudes or lengths for the corresponding eigenvectors.
\end{block}

\begin{block}{Linear Regression Insights}
Concepts from linear regression, such as least squares fitting, can seamlessly integrate with PCA, contributing to accurate modeling of stock price trends. In particular, the residuals from a linear regression model can be used to calculate the variance that is captured by each principal component.
\end{block}
\end{frame}

\begin{frame}{Mathematical Steps in PCA}
\begin{block}{Step 1: Standardize the Data}
First, we standardize the data to have a mean of 0 and a standard deviation of 1. This is given by:

$$ X_{std} = \frac{X - \mu}{\sigma} $$

where $X$ is the original feature vector, $\mu$ is the mean of the feature vector, and $\sigma$ is the standard deviation of the feature vector.
\end{block}

\begin{block}{Step 2: Compute the Covariance Matrix}
Next, we compute the covariance matrix of the standardized data. The covariance matrix is given by:

$$ \Sigma = \frac{1}{n-1} (X_{std} - \mu)(X_{std} - \mu)^T $$
```

where n is the number of observations.
\end{block}
\end{frame}
\begin{frame}{Mathematical Steps in PCA}
\begin{block}{Step 3: Compute the Eigenvalues and Eigenvectors}
We then compute the eigenvalues and eigenvectors of the covariance matrix. The eigenvector v and eigenvalue λ of a matrix A satisfy the equation:

$$ A v = \lambda v $$

The eigenvectors represent the directions or components for the reduced subspace of X_{std}, while the eigenvalues represent the magnitudes for the directions.
\end{block}

\begin{block}{Step 4: Sort Eigenvalues and their Corresponding Eigenvectors}
Next, we sort the eigenvalues in decreasing order and choose the first k eigenvectors, which correspond to the k largest eigenvalues, where k is the number of dimensions of the new feature subspace ($k \leq d$).
\end{block}
\end{frame}

\begin{frame}{Mathematical Steps in PCA}
\begin{block}{Step 5: Transform the Original Dataset}
Finally, we transform the original standardized dataset via the selected eigenvectors to obtain a k-dimensional feature subspace.

$$ Y = X_{std} \cdot W $$

where Y is the transformed k-dimensional feature subspace and W is the matrix of the top k eigenvectors.
\end{block}
\end{frame}

\begin{frame}{Finding Principal Components with Max Variance}
\begin{block}{Utilizing Eigenvalues}
Eigenvalues aid in identifying principal components with maximum variance, a crucial step in dimensionality reduction. The principal components are the eigenvectors of the covariance matrix of the data, and the magnitude of the

corresponding eigenvalue indicates the amount of variance captured by that principal component.
\end{block}

\begin{block}{Projection for Dimensionality Reduction}
By projecting all components onto those with maximum variance, we effectively reduce dimensions while preserving the essential information. This is done by calculating the dot product of the data points with the eigenvectors.
\end{block}
\end{frame}

\begin{frame}{Linear Regression Diagram}
\begin{center}
  \includegraphics[width=1\textwidth]{lr.png}
\end{center}
\end{frame}

\begin{frame}{Stock Price Prediction with PCA Code}
\begin{center}
  \includegraphics[width=1\textwidth]{code1.png}
\end{center}
\end{frame}

\begin{frame}{Stock Price Prediction with PCA Graph}
\begin{center}
  \includegraphics[width=1\textwidth]{graph1.jpeg}
\end{center}
\end{frame}

\begin{frame}{Dataset Explanation}
\begin{block}{Data Preprocessing and Model Training}
The code starts by fetching historical stock data for Verastem Inc(an American pharmaceutical company that develops medicines to treat certain cancers) using the yfinance library. It then preprocesses the data by calculating the percentage change for the ‘Open’, ‘High’, ‘Low’, ‘Close’, and ‘Volume’ features and standardizing these features to have a mean of 0 and a standard deviation of 1. The code then applies Principal Component Analysis (PCA) to reduce the dimensionality of the data, keeping only the top 5 principal

components. The data is then split into training and testing sets, and a linear regression model is trained on the training set.
\end{block}
\begin{block}{Stock Price Prediction and Evaluation }
The trained model predicts stock prices. Its performance is evaluated using Mean Squared Error and visualized by plotting actual vs predicted prices.
\end{block}
\end{frame}

\begin{frame}{Conclusion}
\begin{block}{Summarizing the Project}
We've successfully explored and demonstrated the efficacy of PCA in predicting stock prices, highlighting its role in mitigating dimensionality issues. We believe that our work can serve as a foundation for future research in this area.
\end{block}

\begin{block}{Thank You}
We extend our gratitude for the opportunity to undertake this project and for your attention. We look forward to any feedback or questions you may have.
\end{block}
\end{frame}

\end{document}

Thank You