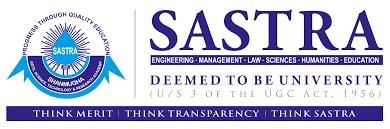
****

**Analysis of SP500 Index using HMM**

July - November 2025

*Submitted By*

Aravind Suresh - 125018007

B. Tech Computer Science and Business Systems

Submitted To

DR. Swetha Varadarajan

**TABLE OF CONTENT**

1. **ABSTARCT**
2. **INTRODUCTION**
3. **BACKGROUND**
4. **MEHTODOLOGY**
5. **DISCUSSION**
6. **LEARNING OUTCOMES**
7. **REFERENECE**

**ABSTRACT**

This project applies a Hidden Markov Model (HMM) to analyse and predict volatility regimes within the S&P 500 index, aiming to provide insights into cyclical market behaviours. Using historical stock data, the model identifies three hidden states—low, moderate, and high volatility—that correspond to different market conditions. By leveraging key features like volatility and return, the HMM captures complex patterns in price fluctuations, classifying market phases to help investors understand underlying trends and potential risk levels. The model’s performance is evaluated with metrics like log-likelihood and pseudo-accuracy, achieving an 82.9% alignment with volatility-based pseudo-labels, which reflects its predictive capability. Results reveal the HMM’s effectiveness in distinguishing regime shifts, providing a valuable framework for volatility forecasting in financial markets.

**INTRODUCTION**

Financial markets are inherently volatile, with periods of stability and turbulence driven by a range of economic, political, and psychological factors. Identifying and forecasting these market regimes can help investors and analysts make informed decisions by understanding potential risks and opportunities. The S&P 500, a prominent index representing a broad cross-section of the U.S. stock market, exhibits distinct volatility patterns that can be analysed through machine learning techniques to detect cyclical phases. Hidden Markov Models (HMMs), commonly used for time-series data, are well-suited for this type of analysis as they can uncover hidden states in observable data, making them a powerful tool for modelling market behaviour.

In this project, we implement an HMM to segment S&P 500 historical data into volatility regimes, providing insight into the market’s underlying structure. By defining three distinct hidden states that represent different levels of volatility, we aim to classify daily market conditions and predict regime changes over time. Using only two input features—volatility and return—the model captures essential characteristics of market behaviour and provides a foundation for potential applications in risk management and trading strategies. This study underscores the potential of HMMs in financial analysis, highlighting their ability to reveal meaningful patterns and assist in navigating the complexities of stock market fluctuations.

**BACKGROUND**

*Models Used:* The project employs an HMM with parameters such as the number of hidden states (ranging from 2 to 6) to capture different market phases. By leveraging log returns and 10-day SMA volatility, the model identifies states representing stable, cautious, and high-risk market phases. Statistical measures like AIC and BIC guide model optimization, ensuring the balance between model complexity and predictive accuracy.

*Preprocessing Techniques:* The dataset undergoes rigorous preprocessing, including filling missing values and calculating indicators like volatility and moving averages. For each day, 10-day SMA and single-day returns are computed, normalized using Z-scores to prevent scale discrepancies. After constructing training and testing datasets, no dimensionality reduction is applied as all features contribute to defining market phases.

**METHODOLOGY**

*Experimental Design:* This project uses an HMM with configurations from 2 to 6 hidden states to categorize S&P 500 trends. To build a robust model, daily volatility and returns are computed from historical data, with the model trained to capture hidden states. The optimal model configuration is chosen using criteria like AIC, BIC, CAIC, and HQIC, which help prevent overfitting.

*Environment and Tools:* Implemented in Google Colab, the project uses Python, with pandas, numpy, and matplotlib for data manipulation and visualization. The hmmlearn library is used for HMM implementation, and seaborn is utilized for visualizations, providing an interactive environment for model training and evaluation.

*Code Location:* The codebase, including scripts for data loading, preprocessing, model fitting, and regime prediction, is available on GitHub, making it accessible for review. The repository also contains project documentation for reference.

*Preprocessing Steps:* Initial preprocessing involves calculating 10-day SMA and daily returns, which are then normalized. The data is divided into training and testing sets, with thousands of records ensuring a balanced model. Outliers and feature reduction were not performed, as each feature holds significance in modelling volatility and return-based market regimes.

**DISCUSSION**

*Overall Results:* The HMM effectively identified S&P 500 regimes, classifying data into low, moderate, and high-volatility phases. Visualization of these regimes over time reveals changes in market stability, highlighting useful trends for traders. The results illustrate the model’s potential in providing a framework to anticipate market movements, facilitating data-driven investment decisions.

*Overfitting and Underfitting:* Model selection criteria like AIC and BIC balanced model complexity, avoiding overfitting with excessive hidden states while preventing underfitting. The use of 3-5 hidden states achieved an optimal balance for capturing volatility shifts.

*Hyperparameter Tuning:* The number of hidden states was varied to identify the optimal setting, supported by model evaluation metrics. Future tuning could incorporate different initialization techniques and modify training iterations to further refine predictions.

*Model Comparison and Selection:* HMM’s flexibility in time series analysis makes it a suitable choice for regime detection in financial markets. Compared to alternatives like ARIMA, HMM's state-based prediction model uniquely captures hidden trends, essential for regime classification.

**LEARNING OUTCOMES**

1. **Colab Link**

https://colab.research.google.com/drive/1NAmGOMxinUqutwW8UL0NNlbucvscp6PH?usp=sharing

1. **GitHub link**

<https://github.com/visionblack3/SP500-Stock-Prediction>  
  
 **3**. **Dataset link**

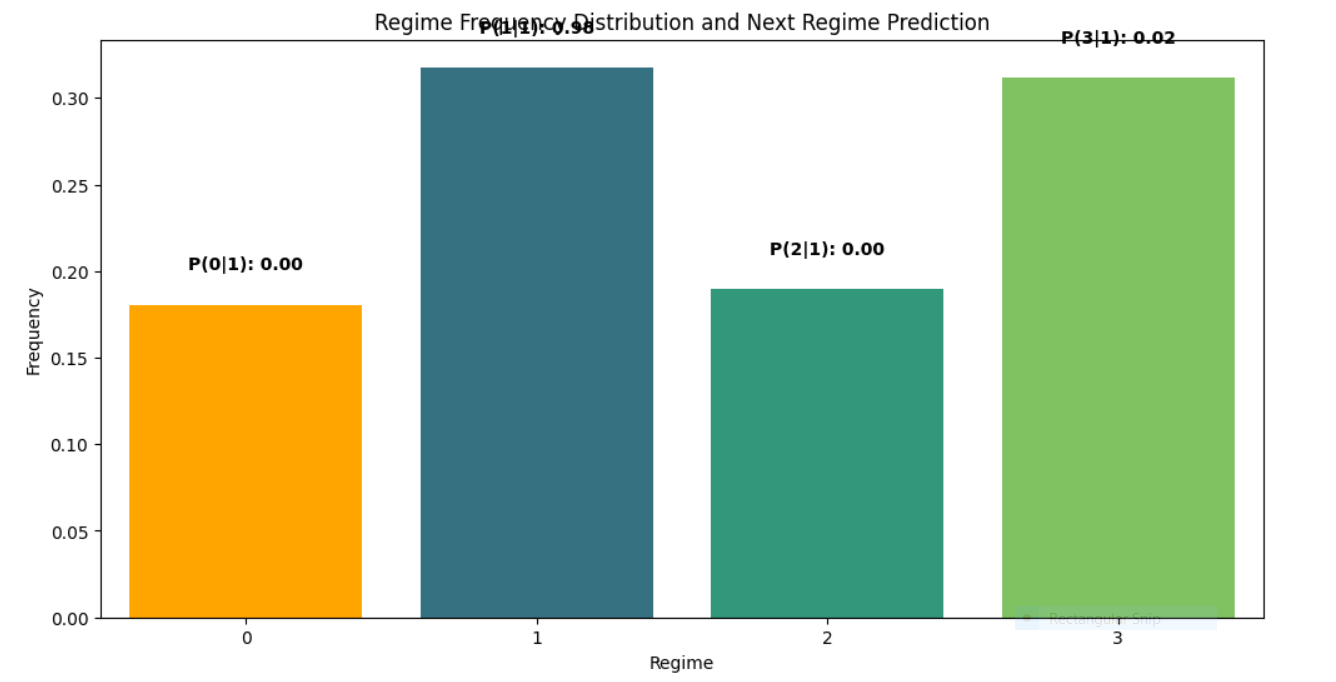
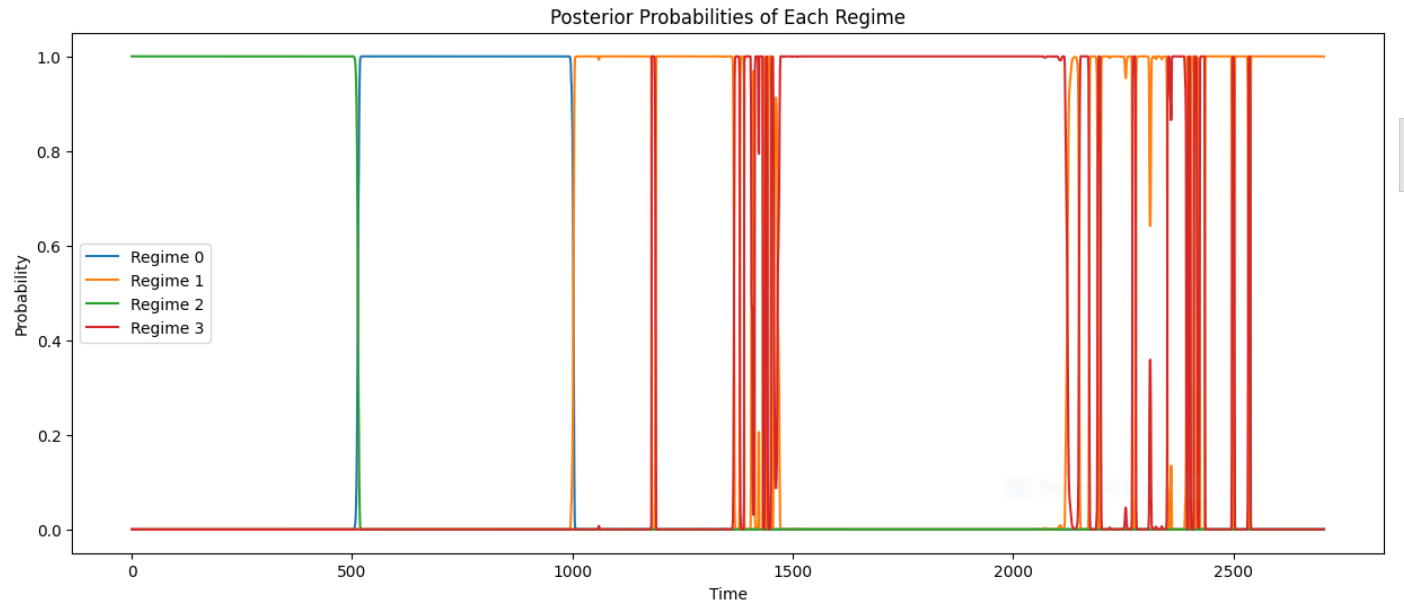
The dataset I have included in GitHub repository check that in the file was included

**Learning from this Project**

This project provided valuable insights into the complexities of financial modelling and the application of machine learning to interpret time-series data. By focusing on the S&P 500 and employing a Hidden Markov Model (HMM), we explored how to detect hidden volatility regimes within financial markets. The non-linear dynamics of market behaviour often mask underlying trends, and the HMM’s ability to probabilistically transition between hidden states—interpreted here as low, medium, and high volatility regimes—enabled us to model these shifts effectively. Working with HMMs also deepened our understanding of sequential data modelling, particularly how the transition matrix and emission probabilities contribute to capturing patterns in market behaviour over time.

Additionally, this project highlighted the importance of careful data handling and feature selection in financial analysis. Choosing relevant features, such as returns and volatility, and pre-processing financial data to reduce noise were crucial to maintaining model accuracy and interpretability. The experience underscored the need for customized evaluation metrics, as conventional measures may not fully capture model performance when using inferred or pseudo-labels for financial states. Overall, this project has not only strengthened our technical understanding of HMMs but also enhanced our ability to interpret complex financial data, skills that are essential for future work in market analysis, forecasting, and risk management applications.

**Results**

****

***The predicted next regime is: Regime 0 - Low Volatility – Stable***

**AIC and BIC result**

**States: 2, AIC: 11973.15, BIC: 12014.47**

**States: 3, AIC: 9900.19, BIC: 9982.83**

**States: 4, AIC: 9001.63, BIC: 9137.40**

**Optimal number of states based on BIC: 4**

**REFERENCE**

1.ChatGPT  
2.<https://www.mdpi.com/2227-7072/6/2/36>  
3.<https://repositories.lib.utexas.edu/server/api/core/bitstreams/5c797dfb-fec6-49ef-b164-cabe66e871b8/content>  
4.<https://www.fidelity.com/learning-center/smart-money/bear-vs-bull-market>  
5.<https://www.researchgate.net/publication/266597121_A_Six_State_HMM_for_the_SP_500_Stock_Market_Index>