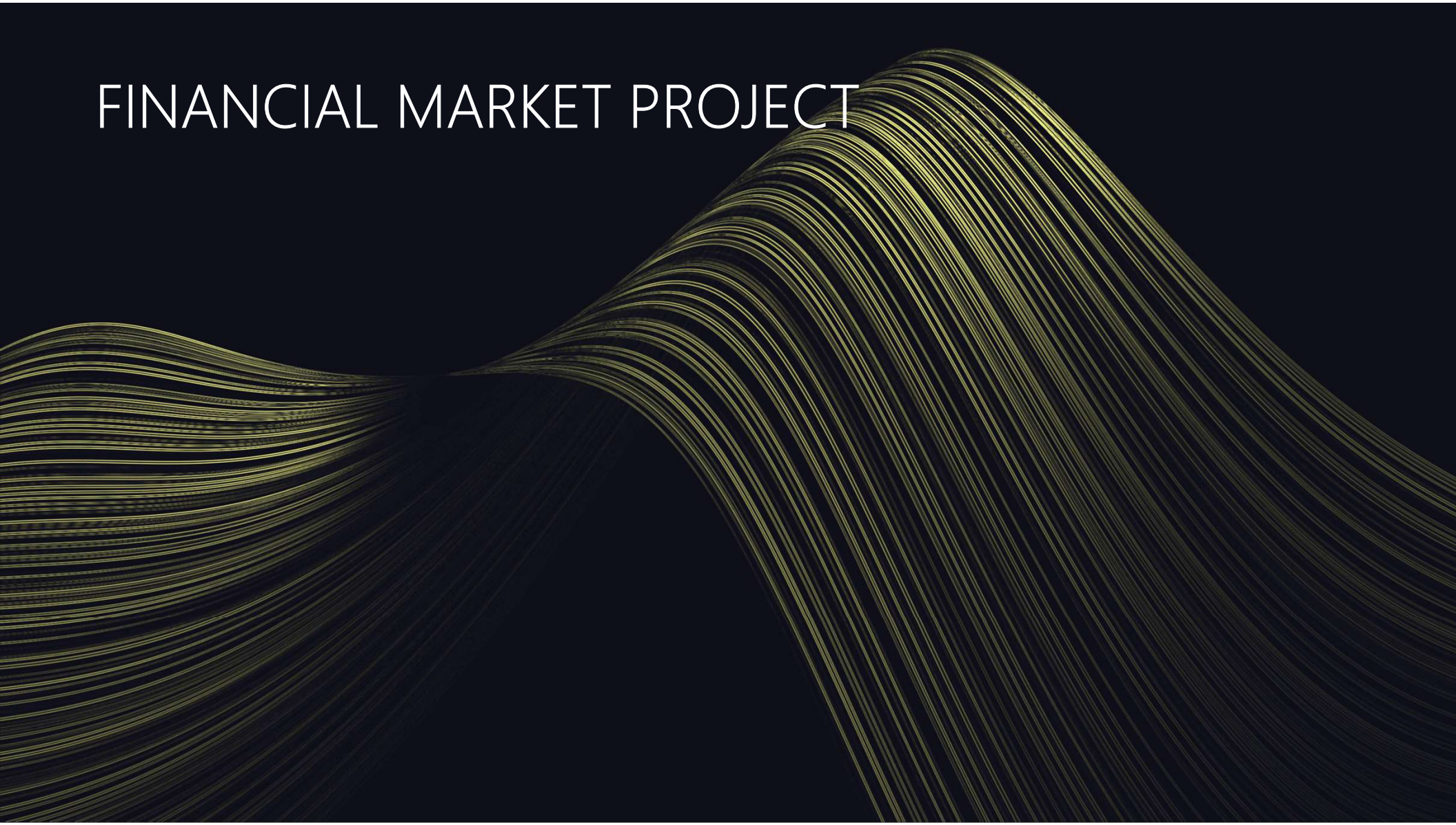
FINANCIAL MARKET PROJECT



PROJECT BACKGROUND

Global Market Interconnectedness

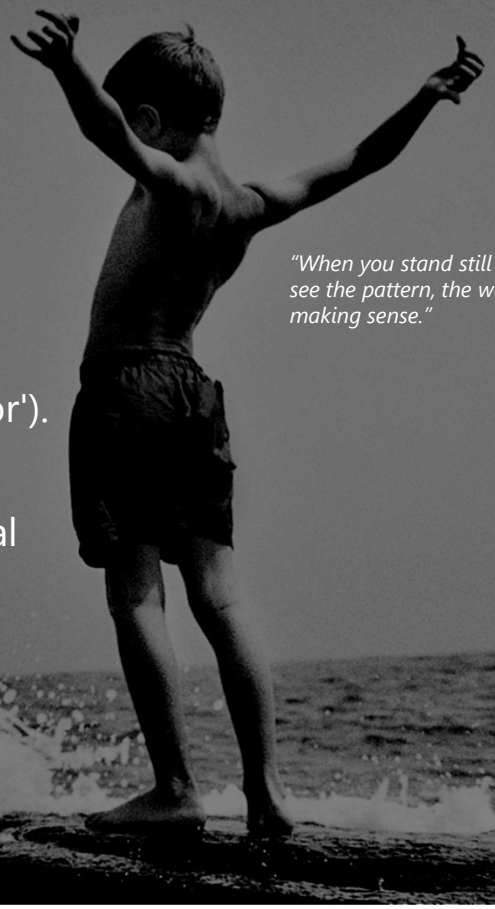
This project seeks to analyse the interconnectedness of major global equity markets following a period of significant systemic shocks.

The study focuses on six major global market indices—**Nifty 50** (India), **Dow Jones Index**, **Nasdaq** (United States), **Hang Seng** (Hong Kong), **Nikkei 225** (Japan), and **DAX** (Germany)—alongside the **VIX** (Volatility Index or 'Fear Factor').

The primary period of analysis covers daily returns from **2019 to 2024**.

We observed a **clear volatility cycle**: a sharp expansion during the 2020 global shock (COVID-19) followed by contraction as markets stabilized.

The analysis confirms that shocks in major economies (e.g., the U.S.) transmit globally but with varying intensity across regions.



"When you stand still long enough to see the pattern, the world starts making sense."

PROJECT BACKGROUND

Predictive Goal

The core objective is to determine whether movements in global indices and volatility act as significant indicators for the daily opening direction of the Nifty 50.

This insight is foundational for subsequent predictive modelling, where **global co-movements serve as explanatory features** for Nifty's directional dynamics.

We aim to integrate metrics capturing risk appetite and stress dynamics (like the VIX) at the start of the day to enhance predictive frameworks.

The analysis includes an examination of short-term sentiment spillover, such as how often global indices record a positive daily return when the Nifty opens strong.

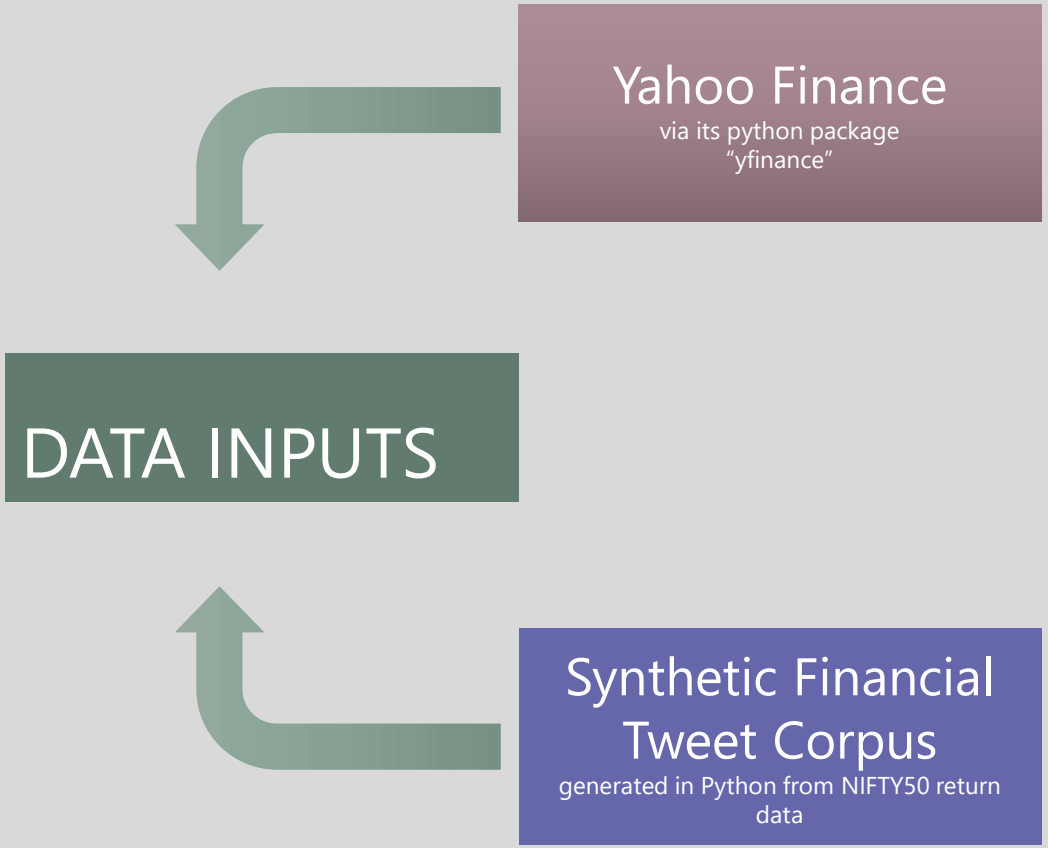
"Every insight begins as a moment of openness."

DATA SOURCES

This project draws on two primary data sources that together form the foundation of the analysis.

The first source is Yahoo Finance, accessed programmatically through Python's yfinance package, which provided daily price and return data for major global market indices.
This dataset enabled the creation of a comprehensive master table used throughout the modelling and exploratory analysis phases.
The second input is a Synthetic Financial Tweet Corpus, generated in Python using real NIFTY50 return data to simulate market-aware tweets.
This allowed the project to incorporate sentiment analysis without relying on paid Twitter API access.

Together, these two inputs support both the quantitative market modelling and the text-based sentiment components of the project.



DATA SNAPSHOT

	Nifty_Close	DowJones_Close	Nasdaq_Close	HangSeng_Close	Nikkei_Close	DAX_Close	VIX_Close	Nifty_Return	DowJones_Return	Nasdaq_Return	HangSeng_Return	Nikkei_Return	DAX_Return	Year	Quarter	Month
2020/01/02	12 282	28 869	9 092	28 544	23 488	13 386	12.470	0.935	1.158	1.333	1.255	-0.478	0.703	2020	1	1
2020/01/03	12 227	28 635	9 021	28 452	23 375	13 219	14.020	-0.452	-0.810	-0.786	-0.322	-0.481	-1.246	2020	1	1
2020/01/06	11 993	28 703	9 071	28 226	23 205	13 127	13.850	-1.911	0.239	0.562	-0.792	-0.729	-0.697	2020	1	1
2020/01/07	12 053	28 584	9 069	28 322	23 576	13 227	13.790	0.499	-0.417	-0.032	0.340	1.598	0.761	2020	1	1
2020/01/08	12 025	28 745	9 129	28 088	23 205	13 320	13.450	-0.229	0.565	0.669	-0.827	-1.573	0.706	2020	1	1
2020/01/09	12 216	28 957	9 203	28 561	23 740	13 495	12.540	1.585	0.737	0.813	1.684	2.306	1.313	2020	1	1
2020/01/10	12 257	28 824	9 179	28 638	23 851	13 483	12.560	0.335	-0.460	-0.267	0.270	0.466	-0.087	2020	1	1
2020/01/13	12 330	28 907	9 274	28 955	23 915	13 452	12.320	0.594	0.289	1.036	1.106	0.268	-0.236	2020	1	1
2020/01/14	12 362	28 940	9 251	28 885	24 025	13 456	12.390	0.266	0.113	-0.244	-0.241	0.463	0.037	2020	1	1
2020/01/15	12 343	29 030	9 259	28 774	23 917	13 432	12.420	-0.154	0.313	0.080	-0.386	-0.452	-0.180	2020	1	1
2020/01/16	12 356	29 298	9 357	28 883	23 933	13 429	12.320	0.099	0.921	1.063	0.380	0.069	-0.021	2020	1	1
2020/01/17	12 352	29 348	9 389	29 056	24 041	13 526	12.100	-0.025	0.172	0.340	0.600	0.452	0.720	2020	1	1
2020/01/20	12 225	29 282	9 378	28 796	24 084	13 549	12.598	-1.035	-0.227	-0.113	-0.897	0.176	0.169	2020	1	1
2020/01/21	12 170	29 196	9 371	27 985	23 865	13 556	12.850	-0.447	-0.292	-0.080	-2.815	-0.909	0.051	2020	1	1
2020/01/22	12 107	29 186	9 384	28 341	24 031	13 516	12.910	-0.517	-0.033	0.138	1.271	0.699	-0.296	2020	1	1
2020/01/23	12 180	29 160	9 402	27 909	23 795	13 388	12.980	0.607	-0.090	0.199	-1.524	-0.982	-0.942	2020	1	1
2020/01/24	12 248	28 990	9 315	27 950	23 827	13 577	14.560	0.557	-0.584	-0.931	0.145	0.133	1.406	2020	1	1
2020/01/27	12 119	28 536	9 139	27 672	23 344	13 205	18.230	-1.055	-1.566	-1.885	-0.993	-2.030	-2.739	2020	1	1
2020/01/28	12 056	28 723	9 270	27 402	23 216	13 324	16.280	-0.521	0.655	1.426	-0.977	-0.547	0.901	2020	1	1
2020/01/29	12 130	28 734	9 275	27 161	23 379	13 345	16.390	0.611	0.040	0.059	-0.879	0.705	0.160	2020	1	1
2020/01/30	12 036	28 859	9 299	26 449	22 978	13 157	15.490	-0.772	0.435	0.256	-2.620	-1.718	-1.408	2020	1	1
2020/01/31	11 962	28 256	9 151	26 313	23 205	12 982	18.840	-0.612	-2.091	-1.591	-0.516	0.990	-1.331	2020	1	1
2020/02/03	11 708	28 400	9 273	26 357	22 972	13 045	17.970	-2.125	0.509	1.338	0.169	-1.005	0.487	2020	1	2
2020/02/04	11 980	28 808	9 468	26 676	23 085	13 282	16.050	2.321	1.436	2.098	1.210	0.490	1.813	2020	1	2
2020/02/05	12 089	29 291	9 509	26 787	23 320	13 478	15.150	0.914	1.677	0.430	0.415	1.018	1.480	2020	1	2

The **Master Data** was finalized after several cleaning and transformation steps:
Daily Returns Calculation: Calculated based on close prices: (Y_{t}-Y_{t-1})/Y_{t-1}\times 100.
Data Merging: All index files were merged using an outer join, noting that global holidays differ.
Missing Data Imputation: Missing data (due to holidays or non-trading days) were imputed using the **Linear interpolation** method.
Indicator Variables: Indicator variables for "Year," "Quarter," and "Month" were created to support time-series analysis.
Target Variable: The dependent variable, **Nifty_Open_Dir**, was created, defined as 1 if the Nifty 50 Open at time t is greater than the Close at t-1, and 0 otherwise.

OBJECTIVES

1. Build a unified financial market dataset

Combine global index data (Nifty, Dow Jones, Nasdaq, Hang Seng, Nikkei, DAX, VIX) into a structured master table.

2. Analyse market behaviour over time

Examine volatility, return patterns, and correlations across indices.

Identify key periods such as COVID-19 volatility spikes and post-pandemic stabilisation.

3. Predict short-term index direction

Develop classical statistical and machine-learning models to forecast Nifty50 opening direction based on global market movements.

4. Compare modelling approaches

Evaluate logistic regression, decision trees, random forests, and other classifiers.

Assess accuracy, confusion matrices, ROC curves, and AUC to determine the best-performing model.

5. Integrate text sentiment with market analysis

Generate a large corpus of synthetic, market-aware tweets for NLP and sentiment classification.

Analyse how sentiment shifts correspond to market movements.

6. Demonstrate a full end-to-end data-science pipeline

Data collection → cleaning → feature engineering → EDA → modelling → evaluation → text mining.

Present insights and actionable findings supported by visualisation and statistical evidence.

ANALYSIS PLAN

Data Preparation
Downloaded global index data (Nifty, Dow Jones, Nasdaq, Hang Seng, Nikkei, DAX, VIX) using yfinance.
Cleaned and merged all indices into a single master dataset (2018–2024).
Calculated daily percentage returns and created date-based features (year, month, quarter).

Exploratory Data Analysis
Analysed return distributions and volatility trends.
Produced year-by-year boxplots to compare market behaviour (including COVID period).
Computed correlation matrices to examine inter-market relationships.

Project Aim
Understand global market movements and their relationships.
Explore whether global returns help explain Nifty direction.
Extend the analysis using tweet sentiment generated from market behaviour.

ANALYSIS PLAN

(CONTINUED)

Modelling
Built classical models (e.g., logistic regression) to assess relationships between global returns and Nifty direction.
Trained machine-learning models including **Decision Trees** and **Random Forests**.
Compared model performance using accuracy, confusion matrices, and ROC–AUC scores.

Sentiment Analysis
Generated a synthetic corpus of NIFTY50-styled tweets using a custom Python script.
Extracted sentiment scores (compound, positive, negative, neutral) using **VADER**.
Analysed sentiment distribution and linked it back to market tone.

Outputs Produced
- Summary statistics and EDA visualisations.
- Model performance tables and ROC curves.
- Sentiment bar charts and tweet-based insights.

CHALLENGES

Challenge: Access to Twitter Market Data

Authentic historical financial tweets were locked behind paid APIs (Twitter/X Developer access), making direct sentiment extraction impossible.

Solution Implemented:

Built a **custom Python tweet generator** that produced realistic, market-aware tweets based on actual NIFTY50 return behaviour.

The script mapped daily returns to tone categories (e.g., strong_up, down, flat) and generated linguistically diverse tweets using predefined vocabulary, emojis, macro terms, and timestamps.

This allowed creation of a **large, sentiment-rich corpus** suitable for NLP and VADER analysis without requiring paid API access.

Challenge: Missing Market Data on Certain Dates

Financial markets do not trade every day. Weekends, holidays, or index-specific closures created gaps in the dataset. This caused breaks in the time series and made return calculations inconsistent.

Solution Implemented:

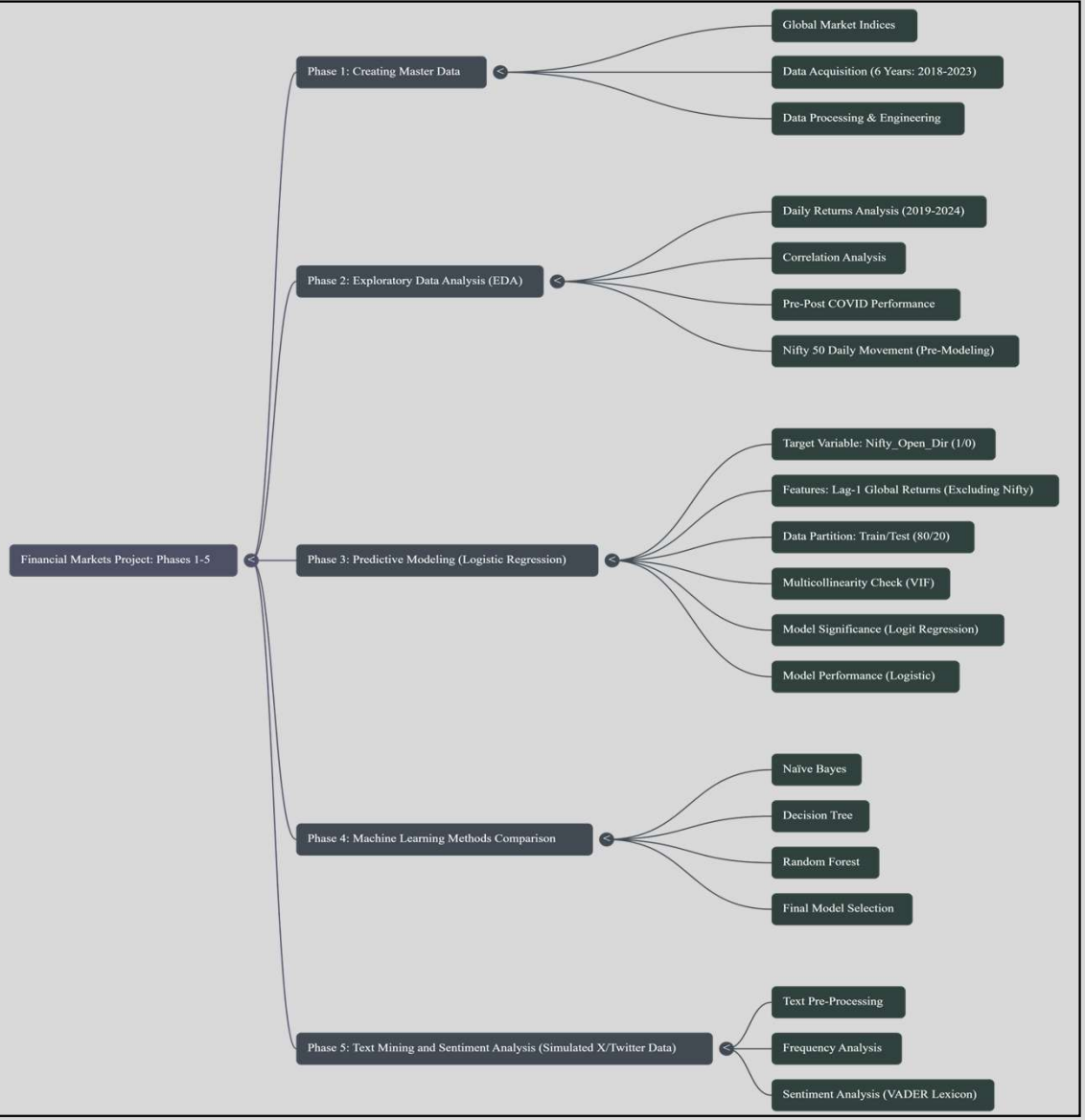
First, **removed all weekends** and **excluded 1 January** each year (a known global holiday) to avoid artificially creating missing values.

For any remaining gaps, applied **linear interpolation** separately to the Open and Close prices.

This method filled only the missing values while preserving all real data points and maintaining smooth, realistic market transitions.

PROJECT PLAN

Phase	Description	Key Deliverables	Deliverables Month
Phase 1: Data Acquisition & Cleaning	Collect OHLC data (6 years), calculate returns, handle missing data, and create indicator variables (Year/Quarter).	Finalized Master Data Table.	end of May 2025
Phase 2: Exploratory Data Analysis (EDA)	Analyze 5-year performance (Box Plots, Summary Tables, Heat Maps), conduct correlation analysis, and analyze pre-post COVID recovery periods.	Correlation Matrices (Full Period & 2024), Recovery Timeline Analysis.	end of May 2025
Phase 3: Binary Logistic Regression	Split data (80/20), run linear model to predict Nifty direction using lagged global returns/VIX, check multicollinearity (VIF), and optimize threshold.	Logit Model with Significant Variables, Train/Test AUC, and Optimal Threshold (0.489) Metrics.	July 2025
Phase 4: Machine Learning Comparison	Apply Naïve Bayes, Decision Tree, and Random Forest methods to the directional prediction task.	Model Comparison Table, Final Model Selection (Random Forest, Test AUC 0.826).	September 2025
Phase 5: Text Mining & Sentiment Analysis	Clean simulated Nifty Twitter/X data, perform tokenization/stopword removal, generate WordCloud, and apply VADER Sentiment Analysis.	Sentiment Distribution Visualizations (Bar Chart, Pie Chart, Histogram).	October 2025



KEY INSIGHTS

The modelling phase showed that global market movements do contain useful signals for anticipating how the NIFTY50 is likely to open on the following day.

Among all approaches tested, the **Random Forest model** consistently delivered the most reliable predictions, handling the complexity and subtle patterns in the data better than simpler statistical methods.

While no model can perfectly predict short-term market direction, the Random Forest demonstrated a strong ability to separate meaningful trends from noise and remained stable when tested on unseen data.

Overall, the project confirms that combining multiple global indices into a single predictive framework provides a solid foundation for short-term market insight, with the Random Forest emerging as the most dependable choice for this task.