Advanced Programming 2025

Predicting Bitcoin Price Direction Using News Sentiment

Final Project Report

Habib Elina
elina.habib@unil.ch
Student ID: 22427009

January 6, 2026

Abstract

This project studies whether news sentiment can help predict the daily direction of Bitcoin prices. Due to the high volatility of cryptocurrency markets and their strong sensitivity to public information, understanding the role of news sentiment is of particular interest. The research question is whether sentiment extracted from Bitcoin-related news articles contains useful information for short-term price direction forecasting. The problem is formulated as a binary classification task, where the objective is to predict whether the Bitcoin closing price increases or decreases from one day to the next. Daily Bitcoin price data is combined with sentiment scores aggregated from news articles. To ensure a realistic prediction setting and avoid look-ahead bias, only information available at time $t-1$, namely lagged returns and lagged sentiment, is used as input features. Several machine learning models are implemented and compared, including Logistic Regression, Random Forest, and K-Nearest Neighbors. Their performance is evaluated using standard classification metrics and compared to a naive baseline strategy that always predicts an upward price movement. The results show that machine learning models slightly outperform the baseline, indicating that news sentiment contains limited predictive information. However, overall performance remains modest, highlighting both the difficulty of predicting short-term Bitcoin price movements and the noisy nature of sentiment signals.

Keywords: data science, Python, machine learning, Bitcoin, cryptocurrency, sentiment analysis, financial prediction, time series

Contents

1 Introduction

Bitcoin is one of the most prominent cryptocurrencies and is widely known for its extreme price volatility. Since its creation, Bitcoin has attracted significant attention from investors, researchers, and policymakers, both as a speculative asset and as a potential alternative to traditional financial systems. Accurately predicting Bitcoin price movements remains a challenging task due to the decentralized nature of the market and the absence of clear fundamental valuation benchmarks.

Unlike traditional financial assets, Bitcoin prices are strongly influenced by public attention, media coverage, and investor sentiment. News articles, online media, and public discourse often play a central role in shaping expectations and triggering market reactions. As a result, sentiment analysis has become an increasingly popular tool for studying cryptocurrency markets.

The objective of this project is to assess whether daily news sentiment can help predict the direction of Bitcoin price movements using machine learning techniques. Rather than aiming to design a profitable trading strategy, the primary goal is methodological: to build a transparent, reproducible, and well-structured data science pipeline that combines financial time-series data with sentiment indicators and evaluates their predictive value relative to simple benchmark strategies.

This project is conducted in the context of the Advanced Programming course and emphasizes clean code structure, modular design, reproducibility, and careful handling of time-series data. The remainder of this report is structured as follows. Section 2 reviews related literature. Section 3 presents the data and methodological framework. Section 5 reports the empirical results, which are analyzed in Section 6. Section 7 concludes and outlines potential extensions.

2 Literature Review / Related Work

A substantial body of literature investigates the relationship between news sentiment and financial markets. Early work by Tetlock (2007) demonstrates that pessimistic media tone is associated with lower stock market returns and increased volatility. This line of research established that textual information extracted from news sources can contain economically meaningful signals.

With the development of natural language processing techniques, sentiment analysis has been widely applied to large textual datasets, including financial news, earnings reports, and social media content. Subsequent studies have extended sentiment-based approaches to high-frequency data and alternative asset classes.

In the context of cryptocurrencies, several studies examine whether sentiment derived from news articles, Twitter posts, or online forums can explain or predict Bitcoin price movements. Due to the speculative nature of cryptocurrency markets, sentiment is often hypothesized to play a more significant role than in traditional financial markets. Some empirical findings suggest short-term predictive effects, particularly around major news events or periods of heightened market attention. However, results remain mixed and highly sensitive to the choice of sentiment source, aggregation frequency, time horizon, and modeling strategy.

From a methodological perspective, prior studies differ substantially in how sentiment information is constructed and aligned with financial data. Some works rely on intraday or hourly sentiment measures, while others aggregate sentiment at the daily or weekly level. Similarly, prediction targets range from price levels to returns or directional movements. These design choices strongly influence empirical results and complicate direct comparisons across studies.

Overall, existing research highlights two key challenges. First, sentiment signals are often noisy, unstable over time, and difficult to exploit consistently in out-of-sample settings. Second, cryptocurrency prices are influenced by a wide range of factors beyond textual sentiment, including macroeconomic conditions, regulatory announcements, technological developments, and

market microstructure effects. Consequently, sentiment-based approaches are rarely sufficient on their own and are typically used in combination with other indicators. This project follows this literature by adopting a cautious and transparent approach, focusing on standard machine learning models, careful temporal data alignment, and explicit baseline comparisons.

3 Methodology

3.1 Data Description

This project relies on two distinct datasets that are combined to study the relationship between news sentiment and Bitcoin price movements.

The first dataset contains daily Bitcoin price data, including the closing price for each trading day. This dataset spans several years and provides a continuous time series of Bitcoin prices. From this data, daily returns are computed as percentage changes in the closing price. These returns are later used both to construct the target variable and as explanatory information.

The second dataset consists of Bitcoin-related news articles, each associated with a numerical sentiment score ranging from -1 to 1. These sentiment scores are obtained through a pre-existing sentiment analysis procedure and reflect the overall tone of each article. Since multiple news articles can be published on the same day, sentiment scores are aggregated at the daily level by computing the average sentiment across all articles published on a given date.

Since sentiment data is only available from 5 November 2021 to 12 September 2024, while Bitcoin price data extends before and beyond this period, the analysis is restricted to the overlapping time window between the two datasets. All observations outside this common period are excluded to ensure temporal consistency and to avoid introducing artificial information into the predictive framework. After alignment and preprocessing, the final dataset contains slightly more than one thousand daily observations.

Days without available news articles are assigned a neutral sentiment value. This choice allows the preservation of all trading days while avoiding the introduction of artificial bias through data deletion or interpolation.

3.2 Approach

The prediction task is formulated as a binary classification problem, where the objective is to predict the direction of the Bitcoin price from one day to the next. The target variable takes the value one if the Bitcoin closing price increases relative to the previous day, and zero otherwise. Focusing on price direction rather than on the magnitude of returns simplifies interpretation and is common practice in short-term financial prediction tasks, especially in highly volatile markets.

To ensure a realistic forecasting framework, the model is restricted to information that would have been available at the time of prediction. Two explanatory variables are therefore constructed: the Bitcoin return observed on the previous day and the average news sentiment aggregated over the previous day. Both variables are lagged by one time step. This design choice explicitly avoids look-ahead bias and prevents the use of future information during training and evaluation.

Prior to model training, several preprocessing steps are applied to ensure data consistency and reproducibility. Date fields from both datasets are converted to a common datetime format to enable accurate temporal alignment. Aggregated sentiment measures are merged with Bitcoin price data using calendar dates, and all features are constructed using only past information. Feature scaling is applied where appropriate to improve model stability and comparability across algorithms.

Three supervised learning algorithms are considered. Logistic Regression is used as a simple and interpretable linear benchmark. Random Forest is included to capture potential non-linear relationships between sentiment and price movements, while K-Nearest Neighbors provides a

non-parametric, distance-based alternative that relies on local similarities in the feature space. Using models of varying complexity allows for a more comprehensive assessment of whether sentiment-based information contributes to predictive performance.

Model selection is deliberately restricted to relatively simple and well-established machine learning methods. While more complex approaches such as deep learning or advanced time-series models could potentially be considered, they typically require substantially larger datasets and introduce additional challenges in terms of interpretability and reproducibility. In the context of financial prediction, simpler models are often preferred as strong baselines due to their robustness and transparency.

Model performance is evaluated using accuracy, precision, and recall. In addition to these machine learning models, a naive baseline strategy is introduced. This baseline always predicts an upward price movement, reflecting the overall bullish bias observed in Bitcoin prices over the considered period. Comparing model performance against this baseline provides a clear benchmark and helps determine whether the proposed approaches offer meaningful predictive improvements beyond a trivial strategy.

Finally, the use of a deliberately small set of explanatory variables is a conscious methodological choice. Given the limited size of the dataset and the noisy nature of sentiment signals, a more complex feature space could lead to overfitting and unstable results. Prioritizing simplicity and interpretability aligns with the objective of the project, which is not to maximize predictive performance at all costs, but to assess whether news sentiment contains any robust incremental predictive signal.

3.3 Implementation

The project is implemented entirely in Python and follows a modular and reproducible design. The codebase is structured into separate modules, each responsible for a specific stage of the machine learning pipeline. This organization improves clarity, facilitates debugging, and allows individual components to be extended or replaced if needed.

Data loading and preprocessing are handled in the `data_loader.py` module. This module reads raw Bitcoin price data and news sentiment data from CSV files, converts date columns into a consistent format, aggregates sentiment scores at the daily level, and merges both datasets. Lagged features are constructed to ensure a proper temporal structure, and feature scaling is applied using standardization.

Model training is implemented in the `models.py` module. Three supervised learning algorithms are considered: Logistic Regression, Random Forest, and K-Nearest Neighbors. Each model is defined in a separate function, which takes the training data as input and returns a fitted model object. The models rely primarily on default or commonly used hyperparameters to ensure transparency and reproducibility. Only minimal adjustments are introduced when necessary for technical reasons, such as ensuring convergence or limiting overfitting, rather than for performance optimization. This choice reflects the objective of the project, which is to assess the predictive contribution of sentiment features rather than to fine-tune model performance.

Model evaluation is handled by the `evaluation.py` module. This module computes classification metrics and produces detailed performance reports. Finally, the `main.py` script orchestrates the entire pipeline, including data preparation, model training, evaluation, baseline comparison, visualization, and result saving.

All figures and tables are generated automatically and saved to disk, ensuring full reproducibility of the empirical results.

Example code snippet: Construction of lagged features

```
# Use yesterday's return and sentiment to predict today's price movement
df["return_lag1"] = df["return"].shift(1)
df["mean_sentiment_lag1"] = df["mean_sentiment"].shift(1)
```

```
4
5 # Remove the first observation where lagged values are unavailable
6 df = df.dropna()
```

4 Code Maintenance and Reproducibility

Code maintenance and reproducibility are central aspects of this project. Version control is managed using Git, and the full codebase is hosted on GitHub. The repository follows a clear and logical structure, separating raw data, source code, generated results, and documentation.

All experiments can be reproduced by running the main script after installing the required dependencies listed in the repository. The modular design of the codebase facilitates maintenance and future extensions, such as adding new predictive models, alternative sentiment indicators, or additional features. While no formal unit tests were implemented due to the exploratory nature of the project, the clear separation of responsibilities across modules helps reduce implementation errors.

This emphasis on code organization, reproducibility, and version control reflects best practices in data science and software development, and aligns closely with the objectives of the Advanced Programming course.

5 Results

5.1 Experimental Setup

All experiments were conducted on a standard personal computer using Python 3. The main libraries used include pandas and `numpy` for data manipulation, `scikit-learn` for machine learning algorithms, and `matplotlib` and `seaborn` for visualization.

The dataset spans from November 2021 to 2024 and is split chronologically into training and testing sets. The first 80% of observations are used for training, while the remaining 20% are reserved for testing. This time-based split is essential in time-series contexts, as random splitting could introduce information leakage from future observations.

To maintain methodological clarity, only limited hyperparameter adjustments are applied. Logistic Regression is trained with an increased maximum number of iterations to ensure convergence. The Random Forest model uses a restricted depth to mitigate overfitting, while the K-Nearest Neighbors model relies on a fixed number of neighbors.

5.2 Performance Evaluation

Model performance is evaluated using accuracy, precision, and recall. Accuracy measures the overall proportion of correct predictions, while precision and recall provide additional insight into how well upward price movements are identified.

In addition to the machine learning models, a naive baseline strategy is introduced. This baseline always predicts that the Bitcoin price will increase. Although simplistic, this strategy provides a meaningful benchmark, as Bitcoin exhibits a general upward trend over the sample period.

Table 1 summarizes the performance metrics for all models. Logistic Regression achieves the highest accuracy, marginally outperforming Random Forest and KNN. All machine learning models slightly exceed the baseline accuracy, indicating that sentiment and lagged returns provide some predictive information. However, these improvements remain modest, highlighting the difficulty of short-term Bitcoin price prediction.

Regarding class-specific metrics, the baseline strategy achieves a perfect recall by construction, as it always predicts an upward movement. Its precision, however, remains limited, reflecting a high number of false positives. In contrast, machine learning models exhibit a more

balanced trade-off between precision and recall. While their ability to selectively identify upward movements remains limited, this behavior reflects the noisy and weakly informative nature of short-term cryptocurrency returns rather than model misspecification.

Table 1: Model Performance Metrics

Model	**Accuracy**	**Precision**	**Recall**
Baseline (Always Up)	0.51	0.51	1.00
Logistic Regression	0.56	0.57	0.59
Random Forest	0.55	0.56	0.55
KNN	0.53	0.55	0.52

5.3 Visualizations

Several visualizations are produced to explore the relationship between news sentiment and Bitcoin returns.

Figure 1 shows the distribution of lagged daily sentiment scores. The distribution is highly concentrated around neutral values, with relatively few extreme observations. This lack of strong sentiment variation may partly explain the limited predictive power of sentiment-based models.

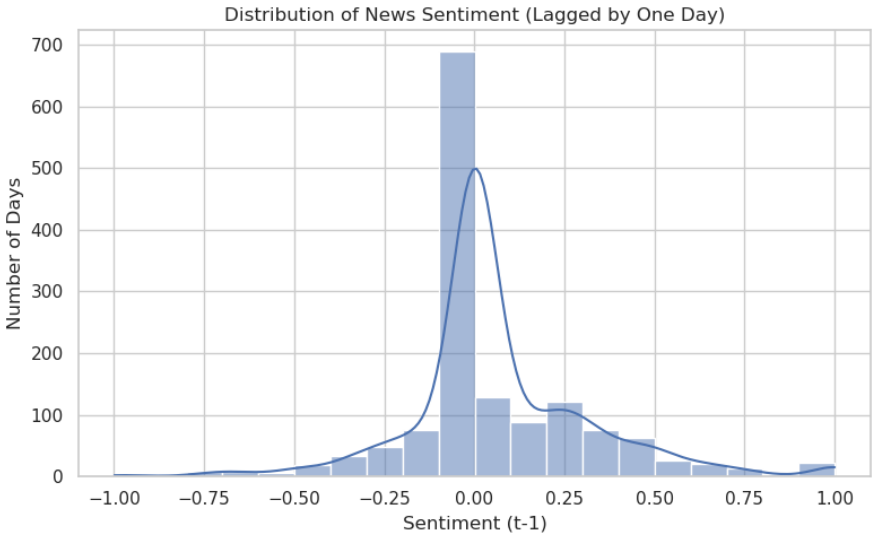

Figure 1: Distribution of daily aggregated news sentiment

Figure 2 presents a scatter plot of yesterday's sentiment versus today's Bitcoin return. The points are widely dispersed, and no clear linear relationship emerges. While some positive sentiment days coincide with positive returns, the overall relationship appears weak and noisy.

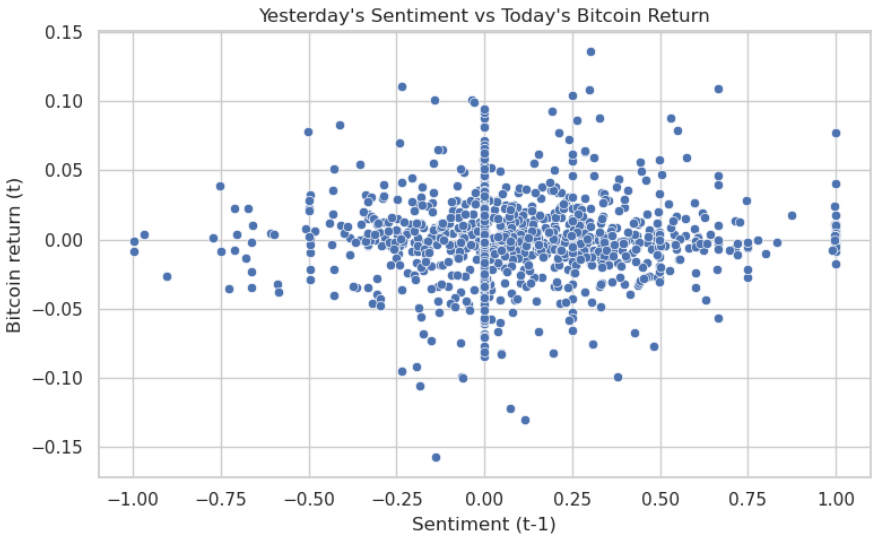

Figure 2: Relationship between lagged sentiment and daily Bitcoin returns

Figure 3 displays seven-day rolling averages of sentiment and returns. Smoothing the series reveals occasional co-movements between sentiment and returns, but these patterns are not persistent over time.

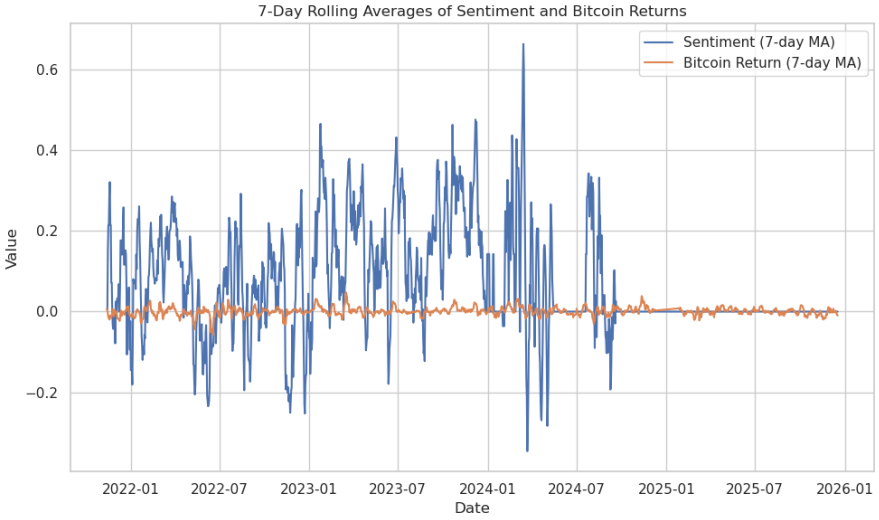

Figure 3: Seven-day rolling averages of sentiment and Bitcoin returns

An additional correlation analysis between lagged news sentiment and daily Bitcoin returns confirms this observation, yielding a correlation close to zero (0.025). This result is consistent with the visual evidence and suggests that daily sentiment alone is insufficient to explain short-term price movements.

6 Discussion

The results indicate that incorporating news sentiment into short-term Bitcoin price direction prediction leads to only limited improvements. Although machine learning models outperform the naive baseline, their accuracy remains close to random guessing. This finding is consistent

with the highly volatile and speculative nature of cryptocurrency markets, where price movements are often driven by rapidly changing expectations and exogenous shocks.

Several factors may explain the modest performance. First, the daily aggregation of sentiment scores may be too coarse to capture rapid market reactions to news, especially in a market that operates continuously and reacts within minutes or hours rather than days. Second, Bitcoin prices are influenced by a wide range of factors beyond textual sentiment, including macroeconomic conditions, regulatory developments, liquidity dynamics, and broader financial market trends. Finally, sentiment scores themselves may contain measurement noise or bias, which can further reduce their explanatory power when used as predictive features.

Despite these limitations, the inclusion of a naive baseline model plays a crucial role in contextualizing the results. The fact that all trained models slightly outperform the baseline suggests that news sentiment does contain some predictive information, even if the signal is weak. From a methodological perspective, this highlights the importance of benchmarking machine learning models against simple strategies to avoid overstating performance gains.

From a practical standpoint, these results suggest that sentiment-based indicators should be interpreted as complementary signals rather than standalone decision tools. Their value may lie in combination with richer market data, higher-frequency sentiment measures, or more advanced modeling approaches. This observation also motivates future work aimed at extending the feature set and exploring alternative representations of sentiment dynamics.

7 Conclusion and Future Work

7.1 Summary

This project examined whether daily news sentiment can help predict the direction of Bitcoin price movements. By combining financial time-series data with aggregated sentiment scores and applying standard machine learning models, the study provides a transparent and reproducible evaluation framework.

The results show that Logistic Regression achieves the best performance, but overall predictive accuracy remains modest. While sentiment appears to contribute marginally, it is insufficient on its own to support strong predictive claims.

7.2 Future Directions

Several extensions could be explored in future work. First, higher-frequency data, such as hourly sentiment and returns, may better capture short-term market reactions. Second, incorporating additional explanatory variables, including trading volume or volatility indicators, could improve performance. Finally, more advanced natural language processing techniques, such as transformer-based sentiment models, may provide richer representations of news content.

References

1. Gupta, Rakshit. (2025). *A Comprehensive Review of Stock Market Price Prediction through News Sentiment Analysis and Machine Learning Approaches*. International Journal of Research & Technology, 13(3), 268–280. Available at: `https://ijrt.org/j/article/view/413`

2. Tetlock, Paul C. (2007). *Giving Content to Investor Sentiment: The Role of Media in the Stock Market*. SSRN Electronic Journal. Available at: `https://business.columbia.edu/sites/default/files-efs/pubfiles/3097/Tetlock_Media_Sentiment_JF.pdf`

3. Du, Kelvin, et al. (2024). *Financial Sentiment Analysis: Techniques and Applications*. ACM Computing Surveys. Available at: `https://sentic.net/financial-sentiment-analysis-survey.pdf`

4. Naeem, Muhammad Abubakr, et al. (2021). *Predictive Role of Online Investor Sentiment for Cryptocurrency Market: Evidence from Happiness and Fears*. International Review of Economics & Finance. Available at: `https://www.sciencedirect.com/science/article/abs/pii/S1059056021000083`

5. Koltun, Vladlen, and Ivan P Yamshchikov. (2023). *Pump It: Twitter Sentiment Analysis for Cryptocurrency Price Prediction*. Risks (MDPI), 11(9), 159. Available at: `https://www.mdpi.com/2227-9091/11/9/159`

Datasets Used

1. Kaggle. (2024). *Crypto Currencies Daily Prices (BTC.csv)*. Available at: `https://www.kaggle.com/datasets/svaningelgem/crypto-currencies-daily-prices?select=BTC.csv`

2. Kaggle. (2024). *Sentiment Analysis of Bitcoin News 2021-2024*. Available at: `https://www.kaggle.com/datasets/imadallal/sentiment-analysis-of-bitcoin-news-2021-2024`

Libraries Used

1. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python.

2. Harris, C. R. et al. (2020). NumPy: Array programming in Python.

3. McKinney, W. (2010). Data structures for statistical computing in Python.

4. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering.

5. Waskom, M. L. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software.

Tools Used

ChatGPT (Open AI) and Claude AI were used as support tools for debugging Python code, improving code readability and assisting with report structure.

A Additional Figures

Figure 4 illustrates the distribution of upward and downward Bitcoin price movements in the test set. While the two classes are nearly balanced, upward movements occur slightly more frequently. This mild imbalance helps explain why the naive baseline strategy, which always predicts an upward movement, achieves a perfect recall despite its limited overall predictive power.

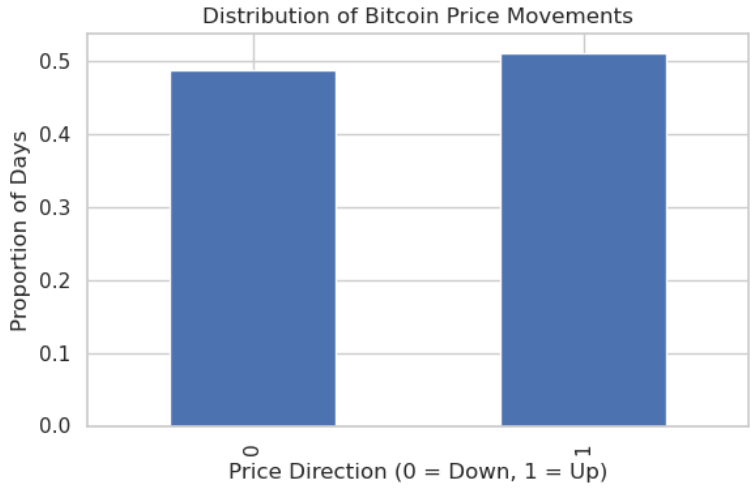

Figure 4: Distribution of upward and downward Bitcoin price movements in the test set.

B Code Repository

GitHub Repository: `https://github.com/Elinahbb/elina_projet.git`

Repository structure:

- README.md: Project description and instructions
- PROPOSAL.md: Project proposal
- project_report.pdf : Final project report (PDF)
- environment.yml: Python dependencies
- data/: CSV datasets used in the project
- main.py: Main script to run the entire pipeline
- src/: Python modules (data_loader.py, models.py, evaluation.py)
- results/: Generated figures and results
- notebooks/: notebooks for data exploration and visualization

Installation instructions & How to reproduce results:

To create the environment and execute the pipeline, the following commands can be used:

```
cd elina_projet
conda env create -f environment.yml
conda activate elina-projet
python main.py
```

The script loads and preprocesses the data, trains the machine learning models, evaluates their performance, and generates the figures, which are saved in the `results/` folder. All results and figures should match those presented in the report.