

**HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY**

**SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**----🙣🕮🙡----**

****

**PROJECT**: **PREDICT STOCK PRICE**

**USING GAUSSIAN PROCESS**

INTRODUCTION TO DATA SCIENCE

Instructor: Prof. Thân Quang Khoát

Group Member:

* + - 1. Phí Hoàng Long - 20184288
      2. Lê Minh Hiếu - 20184257
      3. Nguyễn Viết Mạnh Khoa - 20184278

**HÀ NỘI 05/1/2022**

Contents

[**Overview**](#_dija56gksmw7) **3**

[Preface](#_h87zwzaiyln6) 3

[Problem statement](#_d165x7ahghtg) 3

[Project purposes](#_o5uutfq954i7) 3

[Project goals](#_jy9ccwr6t8c7) 3

[Project scope](#_96gon9hpstv2) 3

[**Data Preparation**](#_4b9yaw4a050f) **3**

[Data source: Yahoo! Finance](#_kke9sgwpipcs) 3

[Data scraping](#_jffp6js1rq6y) 4

[Data description](#_5yxg9ab3wh22) 4

[**Data Visualization and Insight Observation**](#_69uso4snhv3q) **4**

[Important terms](#_8mhrfygvmbzo) 5

[Price volatility](#_xcu4y1y5be6z) 5

[(Trading) Volume](#_uitmc69d10iq) 6

[Overall observation (25/11/2000 - 25/11/2021)](#_v1a5bp7uyxb2) 8

[NASDAQ Composite (^IXIC)](#_bhv6cz7vasgx) 8

[Palatin Technologies](#_qckaw2dn7g6x) 8

[Cassava Sciences, Inc.](#_gntt7cg8uv0e) 9

[Latest observation (3/5/2021- 25/11/2021)](#_biv8igocxmcf) 9

[NASDAQ Composite (^IXIC)](#_d0w4s7tttr5n) 10

[Palatin Technologies](#_6ug83ymw76jf) 10

[Cassava Sciences, Inc.](#_jfpqzc8lsc6d) 12

[**Algorithm and Prediction’s Results**](#_f3acgyfbg67j) **13**

[Algorithm: Gaussian Process](#_uuatwx3vaqa7) 13

[Basic Concepts](#_sywaqf4o9vu0) 13

[Covariance Function](#_chlon7b0kt0a) 15

[Hyperparameters Optimization](#_n2ngff5b5ox2) 16

[Forecasting Methodology](#_wtwend5d08u9) 17

[Result and Comparison with other methods:](#_tqrjfux5qkxq) 18

[Result on Gaussian Process:](#_h3xpxoojsngj) 18

[Comparison with other methods: Linear Regression, LSTM Network and ARIMA Model:](#_fn1uh1cmxtp8) 21

[Linear Regression](#_8t9jcvu6zc7s) 21

[Long-Short Term Memory Network](#_hv6nes707mbh) 23

[ARIMA Model](#_rz4xqurwgo7s) 25

# Overview

## *Preface*

A classic and important problem, in that we wish to gain insights about the stock market. Knowledge of stock markets themselves usually suffices, and mathematical models have been tried and tested. By using machine learning, we can automate and reduce the time of the otherwise manual task in regressing an unknown function we assume the stock market follows.

## *Problem statement*

We formulate the problem as a learning problem, as follows: The task at hand is to predict stock prices in the near future of a past time, using past knowledge (past available data, or past predictions). This means taking a point in time in the past, examining the previous stock prices, and predicting the stock prices an amount of time after. Thus, we can evaluate its accuracy by comparing our predictions to the true values from the data we gathered.

## *Project purposes*

We wish to obtain insights about stock market prices in the past, and predict future ones. We can then examine how its behavior changes over time, or identify trends that are hard to notice with human eyes alone.

## *Project goals*

By the end of this project, we wish to predict stock prices in the future. Along the way, we will try to make insights from a source of stock data.

## *Project scope*

This project utilizes some amount of business knowledge in the stock market, and areas of mathematics such as calculus, probability, and statistics.

# Data Preparation

## *Data source: Yahoo! Finance*

Yahoo! Finance is a media property that is part of the Yahoo! network. It provides financial news, data and commentary including stock quotes, press releases, financial reports, and original content. It also offers some online tools for personal finance management. Yahoo! Finance keeps track of a myriad of companies’ stock prices, but we choose three example companies of interest:

* NASDAQ Composite (^IXIC), a stock index that includes almost all stocks listed on the NASDAQ stock exchange.
* Palatin Technologies (PTN), a bio-pharmaceutical company that develops targeted receptor-specific therapeutics for the treatment of various diseases in the United States.
* Cassava Sciences, Inc. (SAVA), a clinical-stage biotechnology company that detects and treats neurodegenerative diseases such as Alzheimer's. Cassava Sciences serves clients in the United States.

## *Data scraping*

We chose Selenium, an open-source browser that can be operated using Python code, to fetch and download the HTML pages. In addition, due to the site’s usage of lazy-loading data, it was necessary to emulate “scrolling down” using Selenium to fetch complete pages.

To extract the data we need from the pages, we used BeautifulSoup, a Python library for parsing HTML/XML documents.

By correlating visible data on the browser and the raw markup we downloaded, we were able to create manipulatable data using Pandas DataFrames. For ease of future use, we extracted the data into .csv files.

## *Data description*

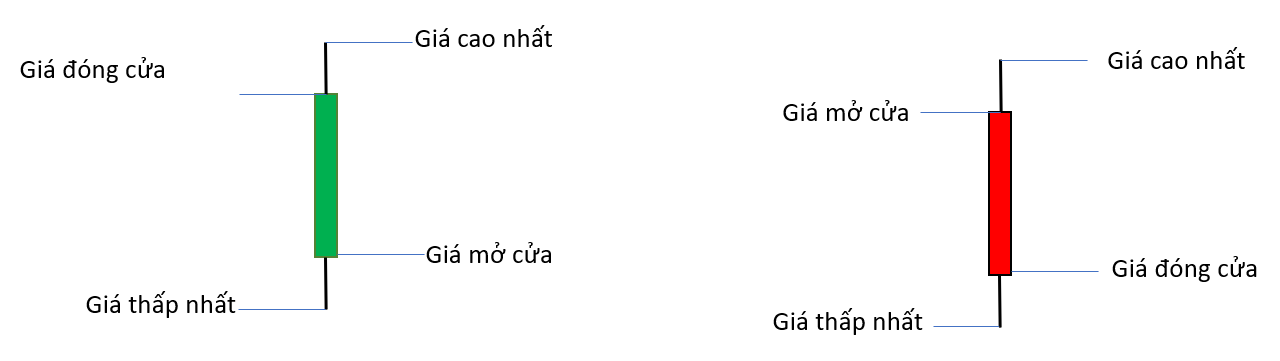
The extracted data contains the following fields:

* Date: The day that trading stocks took place.
* Open: The stock price at the beginning of the trading hours on that date.
* Close: The stock price at the end of the trading hours on that date.
* Adjusted close: The stock price at the end of the trading hours on that date, after undergoing business actions that may or may not affect the closing price.
* High: The highest stock price recorded during that date’s trading hours.
* Low: The lowest stock price recorded during that date’s trading hours.
* Volume: The number of shares traded during that date’s trading hours.

# Data Visualization and Insight Observation

## Important terms

### Price volatility



* Displayed by a candlestick chart.
* Per-day price is represented by 1 candle.
* Green candle: the price was higher than the opening price on the same day.
* Red candle: the price was lower than the opening price on the same day.
* Longer candle, greater variation between opening and closing prices.
* The line above the candle body, if it exists, represents the highest price of a session.
* The line below the candle body, if it exists, represents the lowest price of a session.

### (Trading) Volume



* The number of shares traded over a period of time (in this project: one day). Volume is the main motivation of price increases or decreases.
* Volume colors are treated similar as candle colors (defined above)



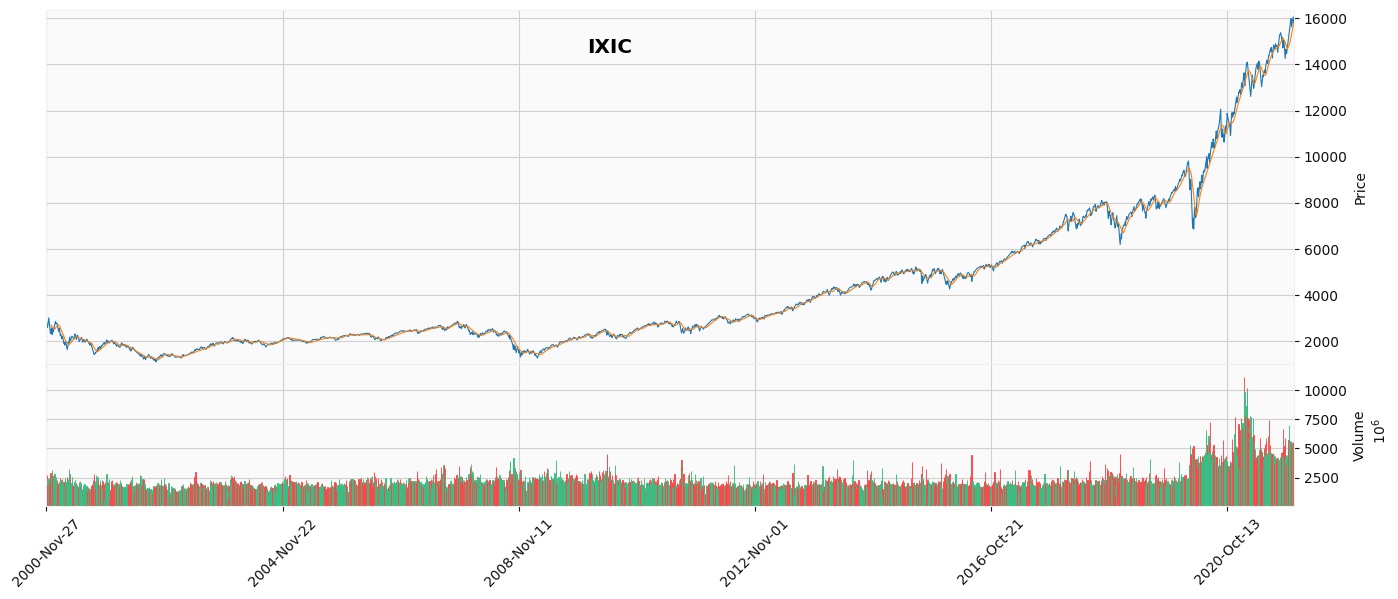
When analyzing the trading volume, it tells us how many people are buying/selling a stock, how it is buying/selling. For instance, in the image above,

* Case A: a red candle with large trading volume shows that many investors want to sell, leading to a decrease in price
* Case B: before that the trading volume is stable and the volume spike in green shows that investors want to buy so the price goes up along with the volume.

=> Trading volume partly tells us the psychology of investors.

## Overall observation (25/11/2000 - 25/11/2021)

### NASDAQ Composite (^IXIC)



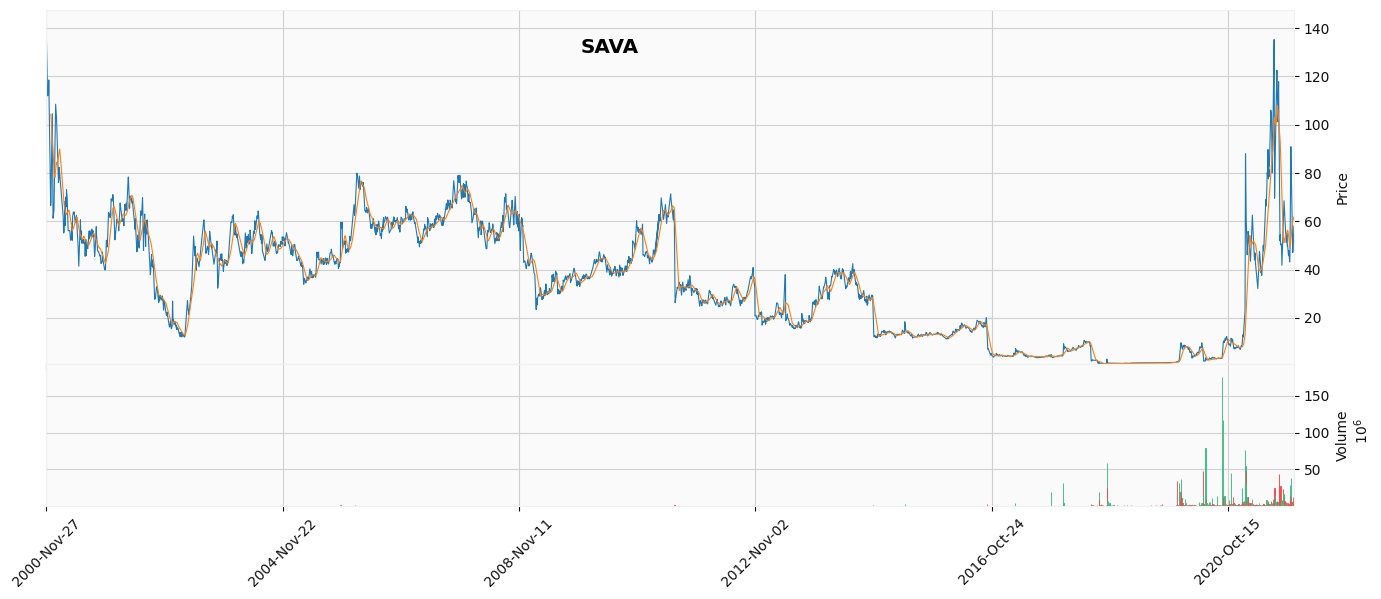
* In the early stage, investors was accumulating investments (the trading volume as well as price variation remained fairly stable until the end of 2016)
* Since 2017s, an upward trend could be observed for the stock price along with the rise of multiple volume spikes in the latest sessions, which may tell us the growth potential of the company.

### Palatin Technologies



* In the first third of the period, PTN just launched its listed shares (low trading volume), so the initial stock price fluctuated significantly, showing minor confidence from potential investors.
* In the following period, the stock suffered from a breakthrough in the zone of support in 2007s, leading to a sharp drop in the stock price with unstable trading volume. The situation did not show any considerable changes after the crisis.

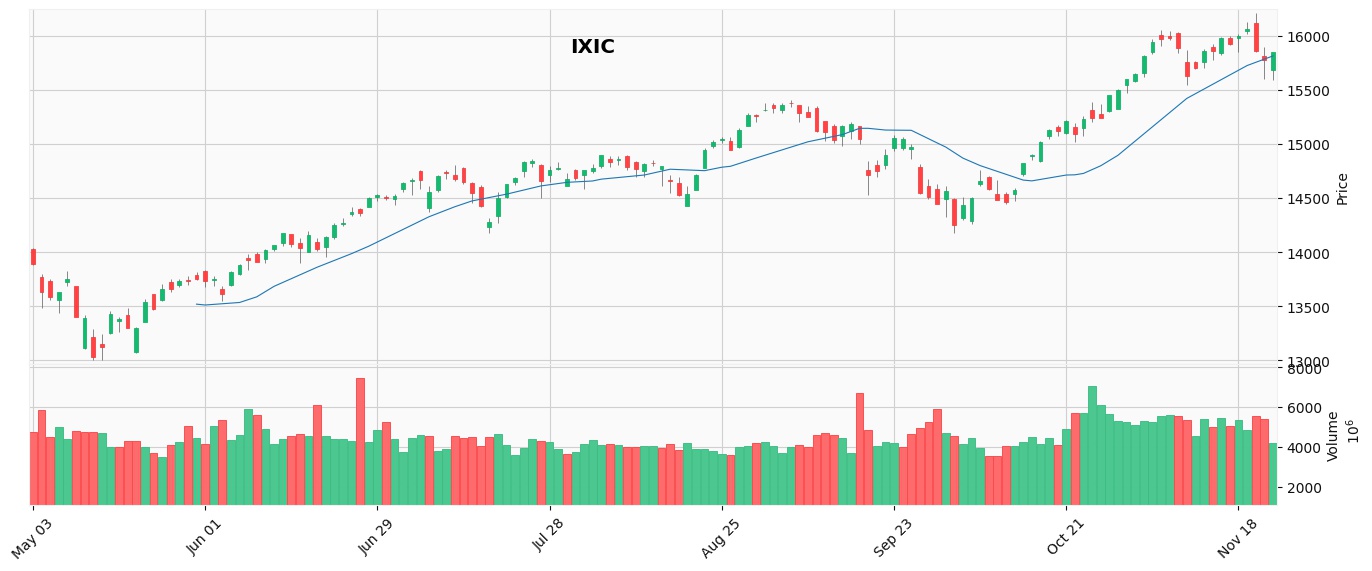
### Cassava Sciences, Inc.



* The graph shows the oscillation of the stock price with a downward trend from 11/2000 to 10/2016. The trading volume during the period was also quite scarce.
* By October 2016, the stock broke through a hard support. The accumulation line continued to stay flat for the next 4 years (until October 2020), followed by a series of volume spikes recently.

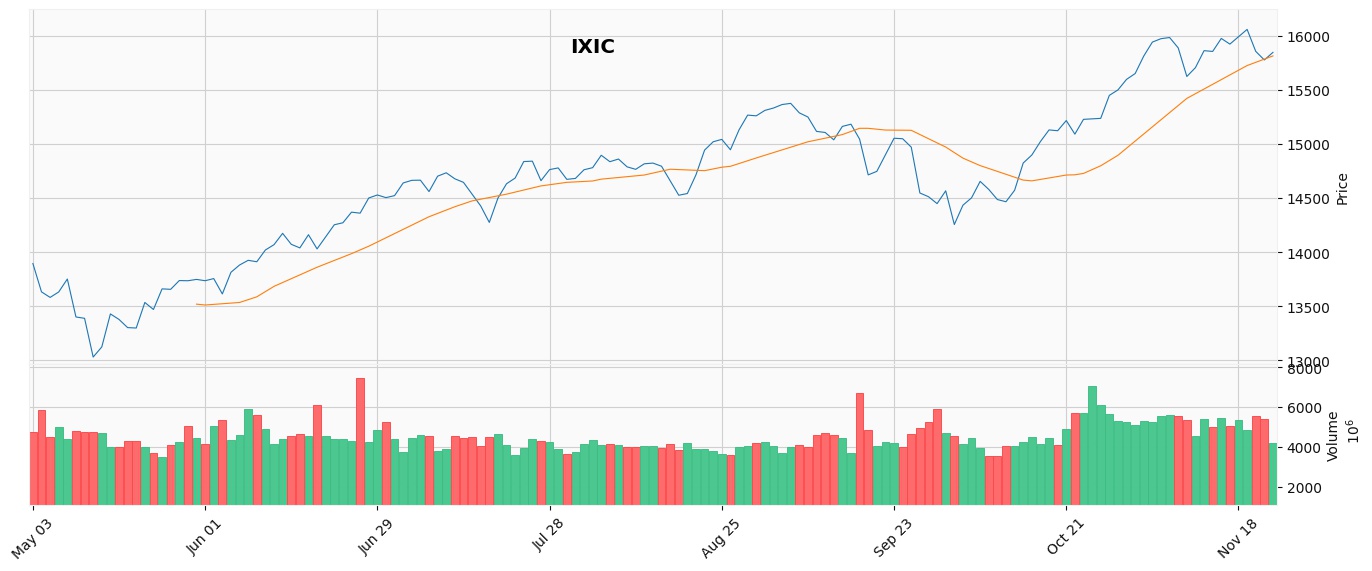
## *Latest observation (3/5/2021- 25/11/2021)*

### NASDAQ Composite (^IXIC)



* It can be observed that this company’s stocks attracted a lot of investment due to continuous demand. At the same time, the fluctuation in both price and volume among sessions are insignificant, but still followed an upward trend.

=> This is a stock worth investing in in the current period.



## 

## 

### Palatin Technologies



* Stocks did not attract investment due to the scarcity in demand (buying volume was negligible)
* MA line (yellow line) is trending down (broken zone of support)
* There was a breakthrough in the latest session where trading volume is more than 100 times higher than previous sessions

=> Stocks are depreciated and easily manipulated by the stock market and therefore should not be invested.

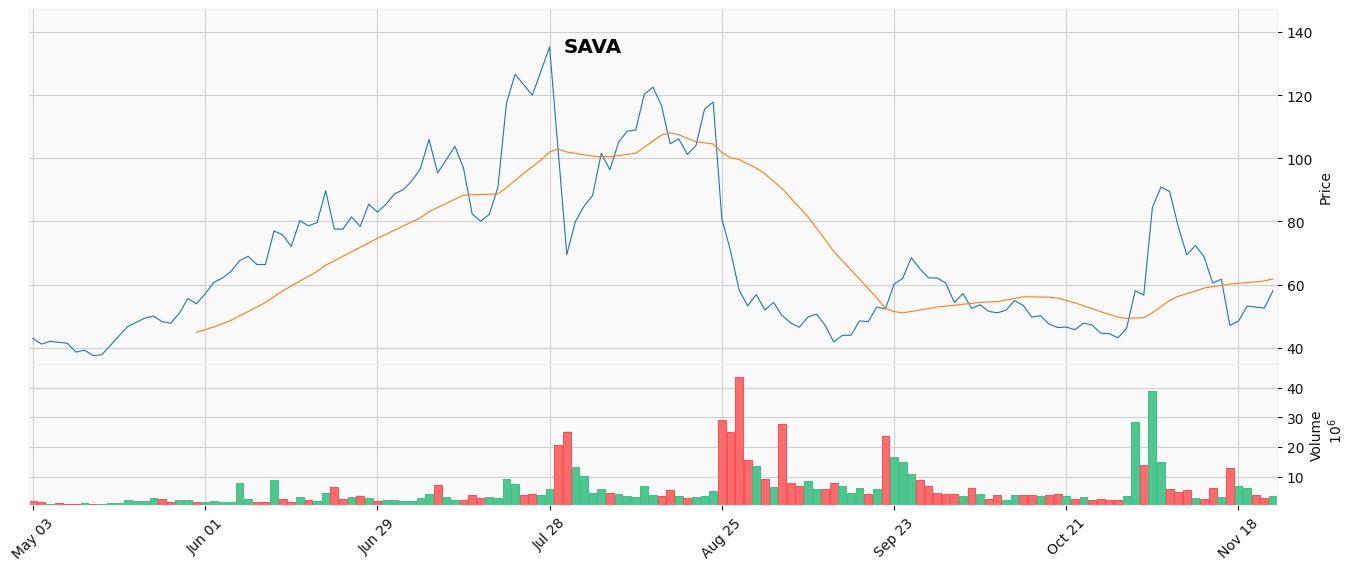


### Cassava Sciences, Inc.



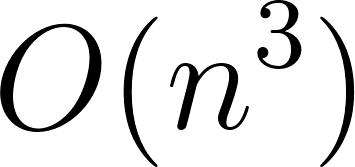
* The price volatility among several sessions was relatively high.
* The most noticeable turnaround could be observed in the period from July to September, where the stock price dropped significantly and also the selling rate was much higher than the buying rate
* The MA trend afterwards fluctuates continuously. However, the buying volume was still modest, hence the stock stayed not very attractive to investors then.

=> Observations should be more paid before investing in.



# Algorithm and Prediction’s Results

## *Algorithm: Gaussian Process*

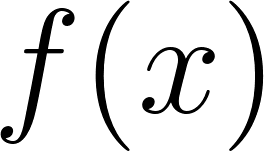
Gaussian processes are a general and flexible class of models for nonlinear regression and classification. They have received attention in the machine learning community over last years, having originally been introduced in geostatistics. They differ from neural networks in that they engage in a full Bayesian treatment, supplying a complete posterior distribution of forecasts. For regression, they are also computationally relatively simple to implement, the basic model requiring only solving a system of linear equations with computational complexity [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=O(n%5E3)#0).

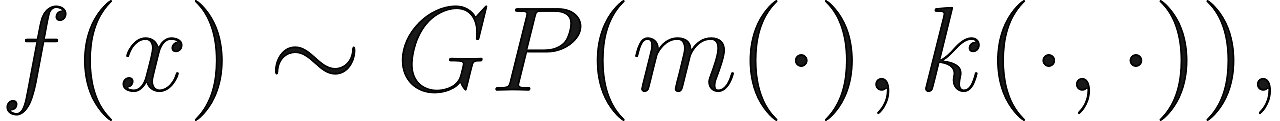
This section will briefly review Gaussian processes at a level sufficient for understanding the forecasting methodology developed in this project.

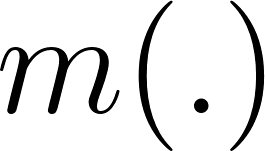
### Basic Concepts

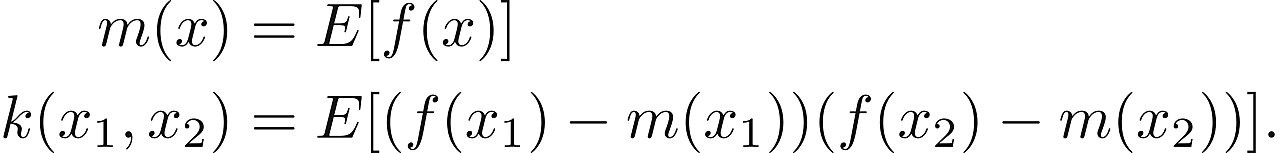
A Gaussian process is a generalization of the Gaussian distribution - it represents a probability distribution over *functions* which is entirely specified by a mean and covariance *functions*. Mathematical definition would be then as follows:

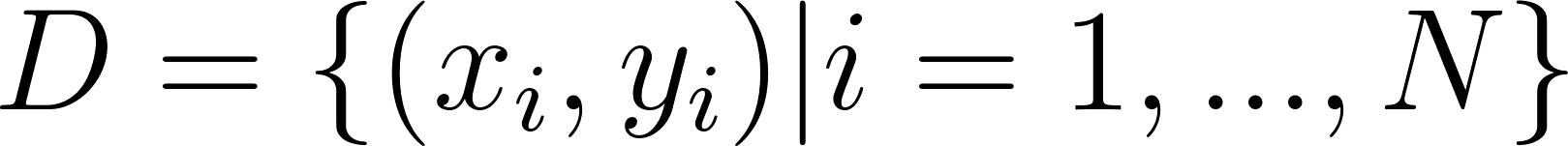
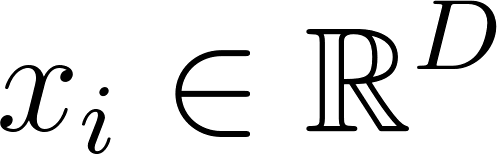
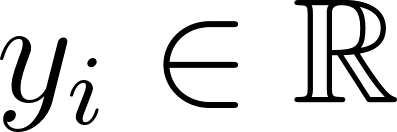
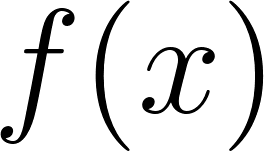
Definition: *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

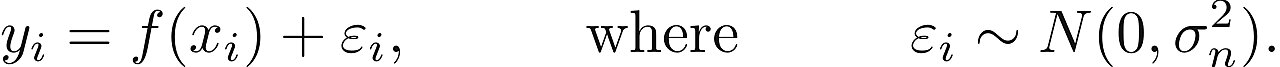
*Let* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=x#0) *be some process* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f(x)#0)*. We write:*

**

*where* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=m(.)#0) *and* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=k(.%2C%20.)#0) *are the mean and covariance functions, respectively:*

**

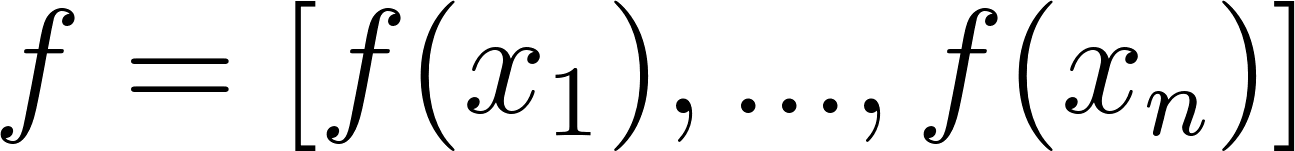
*We will assume that we have a training set* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=D%20%3D%20%5C%7B(x_i%2C%20y_i)%7Ci%20%3D%201%2C...%2CN%5C%7D#0) *where* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=x_i%20%5Cin%20%5Cmathbb%7BR%7D%5ED#0) *and* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=y_i%20%5Cin%20%5Cmathbb%7BR%7D#0)*. For sake of simplicity let* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=X#0)*be the matrix of all inputs, and* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=y#0) *the vector of targets. Also we should assume that the observations* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=y_i#0) *from the process* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f(x)#0) *are noisy:*

**

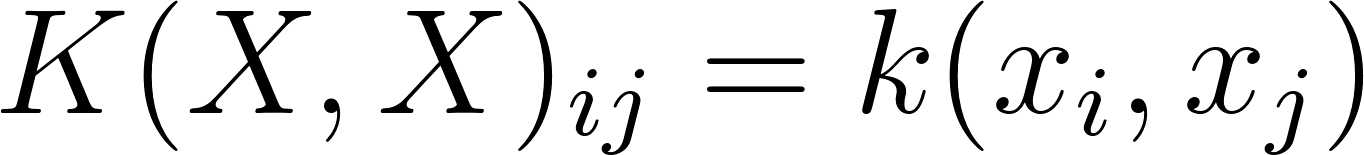
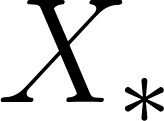
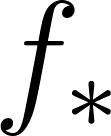
*Regression with a GP is achieved by means of Bayesian inference in order to obtain a posterior distribution over functions given a suitable prior and training data. Then, given new test inputs, we can use the posterior to arrive at a predictive distribution conditional on the test inputs and the training data.*

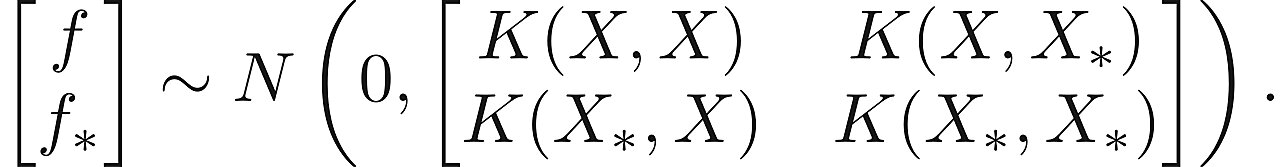
*It is often convenient to assume that the GP prior distribution has mean of zero*

**

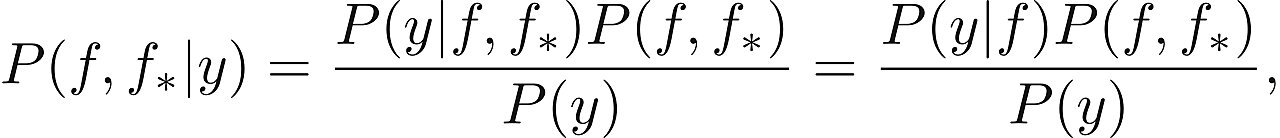
*Let* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f%3D%5C%5Bf(x_1)%2C...%2Cf(x_n)%5C%5D#0) *be a vector of function values in the training set* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=D#0)*. Their prior distribution is then:*

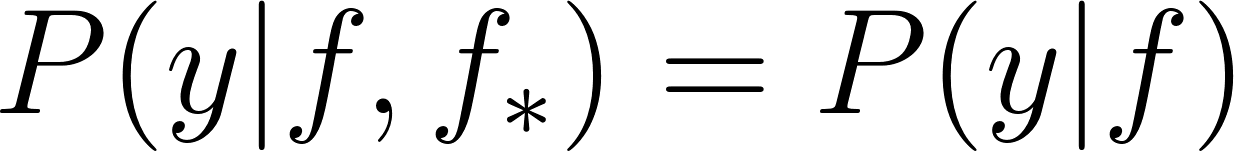
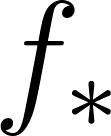
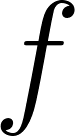
**

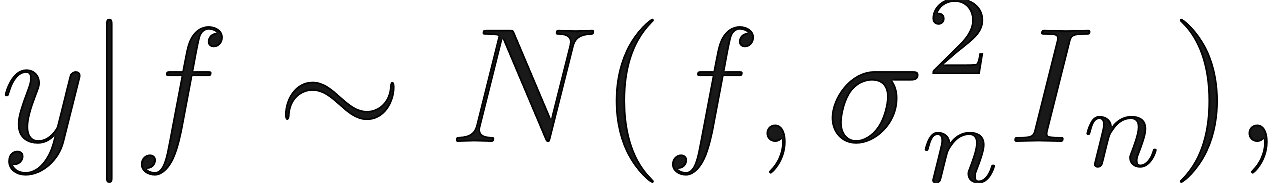
*where* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=K(X%2C%20X)_%7Bij%7D%20%3D%20k(x_i%2C%20x_j)#0) *is a covariance matrix evaluated using covariance function between given points (also known as kernel or Gram matrix). Considering the joint prior distribution between training and the test points, with locations given by matrix* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=X_*#0) *and whose function values are* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f_*#0) *we can obtain that*

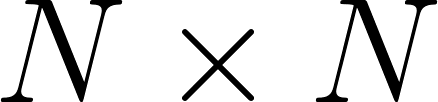
**

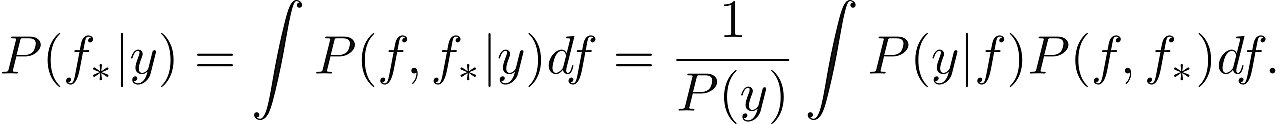
*Then, using Bayes' theorem the join posterior given training data is*

**

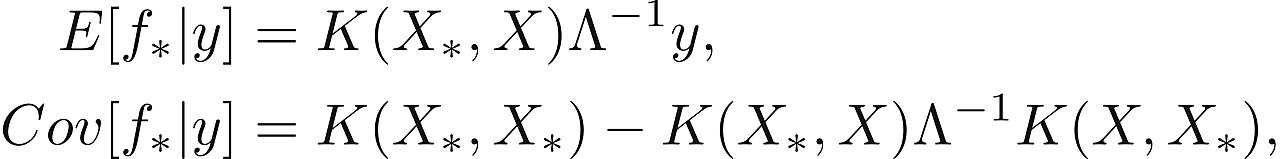
*where* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=P%5C(y%20%7C%20f%2C%20f_*%5C)%20%3D%20P%5C(y%20%7C%20f%5C)#0)*since the likelihood is conditionally independent of* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f_*#0) *given* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f#0)*, and*

**

*where* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=I_N#0) *is* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=N%20%5Ctimes%20N#0) *identity matrix. So the desired predictive distribution is*

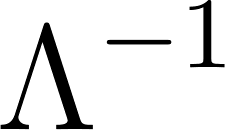
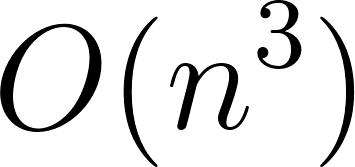
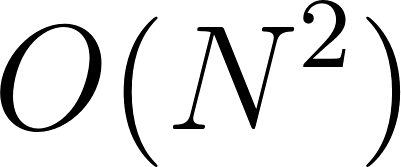
**

*Since these distributions are normal, the result of the marginal is also normal we have*

**

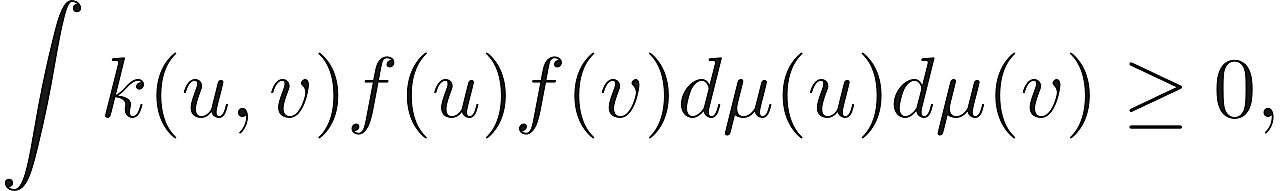
*where* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=y#0) *are test points and*

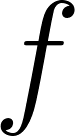
**

*The computation of* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5CLambda%5E%7B-1%7D#0) *is the most computationally expensive in GP regression, requiring as mentioned earlier* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=O%5C(n%5E3%5C)#0) *time and also* [**](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=O%5C(N%5E2%5C)#0) *space.*

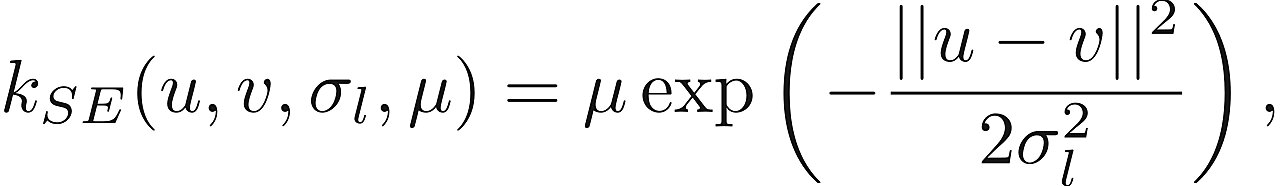
### Covariance Function

A proper choice for the covariance function is important for encoding knowledge about our problem - several examples are given in Rasmussen and Williams (2006). In order to get valid covariance matrices, the covariance function should be symmetric and positive semi-definite, which implies that the all its eigenvalues are positive,

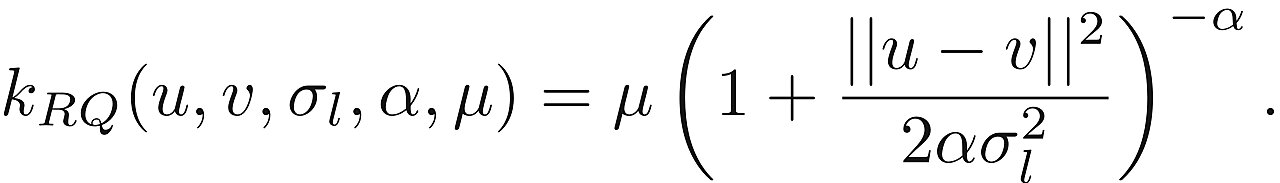


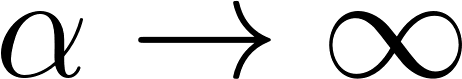
for all functions [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f#0) defined on appropriate space and measure [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Cmu#0).

The two most common choices for covariance functions are the *squared exponential* (also known as the *Gaussian* or *radial basis function* kernel):

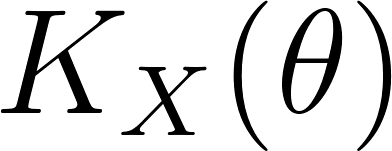


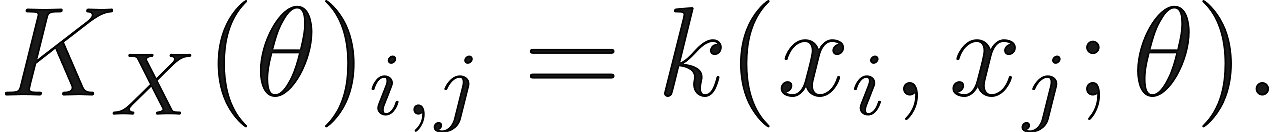
which we will use in our problem and the *rational quadratic*:



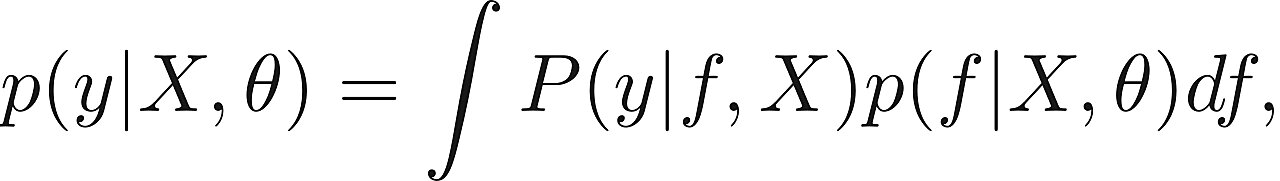
In both cases, the hyperparameter [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Csigma_%7Bl%7D#0) governs the *characteristic length scale* of covariance function, indicating the degree of smoothness of underlying random functions and [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Cmu#0) we can interpret as scaling hyperparameter. The rational quadratic can be interpreted as an infinite mixture of squared exponentials with different length-scales - it converges to a squared exponential with characteristic length-scale [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Csigma_%7Bl%7D#0) as [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Calpha%20%5Crightarrow%20%5Cinfty#0) . In this project has been used classical Gasussian kernel.

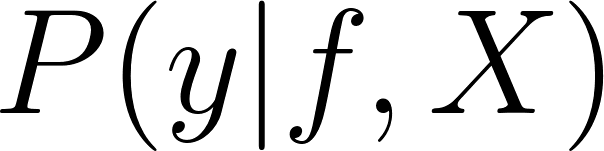
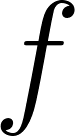
### Hyperparameters Optimization

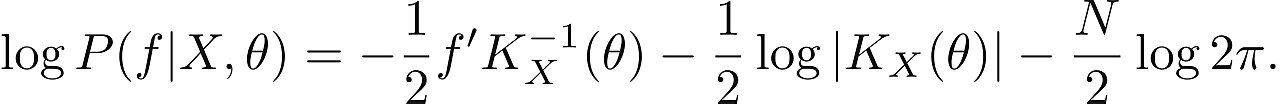
For many machine learning algorithms, this problem has often been approached by minimizing a validation error through cross-validation, but in this case we will apply alternative approach, quite efficient for GP - maximizing the *marginal likelihood* of the observerd data with respect to the hyperparameters. This function can be computed by introducing latent function values that will be integrated over. Let [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Ctheta#0) be the set of hyperparameters that have to be optimized, and [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=K_X(%5Ctheta)#0)be the covariance matrix computed by given covariance function whose hyperparameters are [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Ctheta#0),



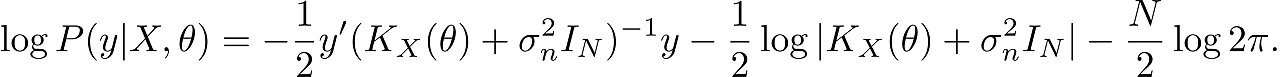
The marginal likelihood then can be written as:



where the distribution of observations [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=P(y%7Cf%2C%20X)#0)is conditionally independent of the hyperparameters given the latent function [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f#0). Under the Gaussian process prior, we have that [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=f%7CX%2C%5Ctheta%20%5Csim%20%5Cmathcal%7BN%7D(0%2C%20K_X(%5Ctheta))#0), or in terms of log-likelihood



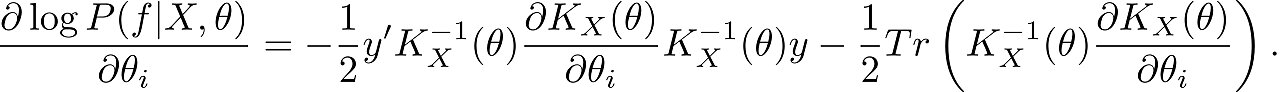
Since the distributions are normal, the marginalization can be done analytically to yield



This expression can be maximized numerically, for instance by a conjugate gradient or like in our case - the default python's sklearn optimizer fmin\_l\_bfgs\_b to yield the selected hyperparameters:

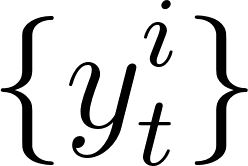
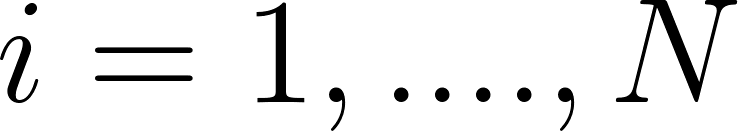
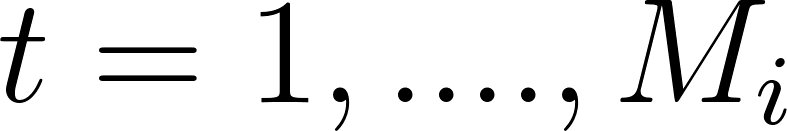


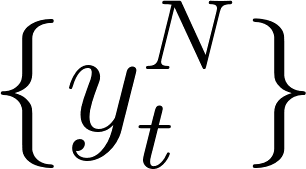
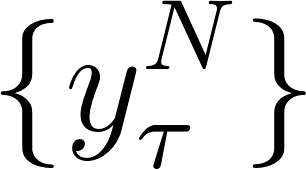
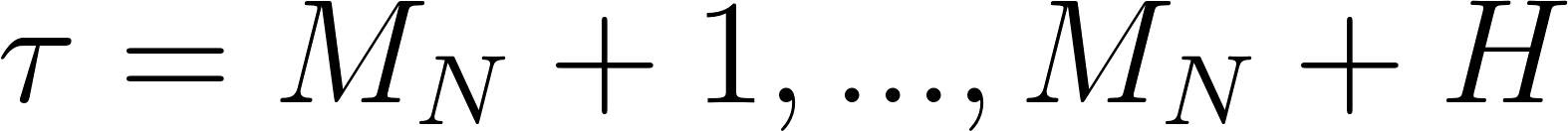
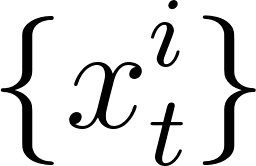
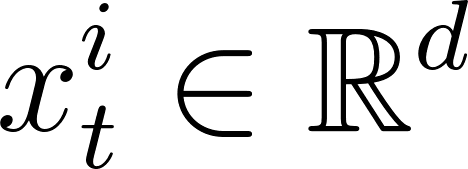
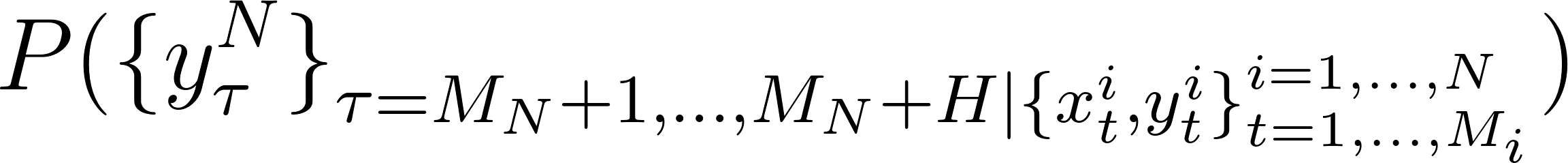
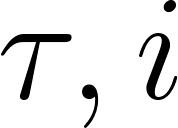
The gradient of the marginal log-likelihood with respect to the hyperparameters-necessary for numerical optimization algorithms—can be expressed as



See Rasmussen and Williams (2006) for details on the derivation of this equation.

### Forecasting Methodology

The main idea of this approach is to avoid representing the whole history as one time series. Each time series is treated as an independent input variable in the regression model. Consider a set of [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=N#0) real time series each of length [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=M_i#0), [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5C%7By%5E%7Bi%7D_%7Bt%7D%5C%7D#0), [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=i%20%3D%201%2C....%2CN#0) and [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=t%20%3D%201%2C....%2CM_i#0). In this application each [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=i#0) represents a different year, and the series is the sequence of a particular prices during the period where it is traded. Considering the length of the stock market year, usually [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=M#0) will be equal to [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=252#0) and sometimes less if incomplete series is considered (for example this year) assuming that the series follow an annual cycle. Thus knowledge from past series can be transferred to a new one to be forecast. Each trade year of data is treated as a separate time series and the corresponding year is used as an independent variable Lin regression model.

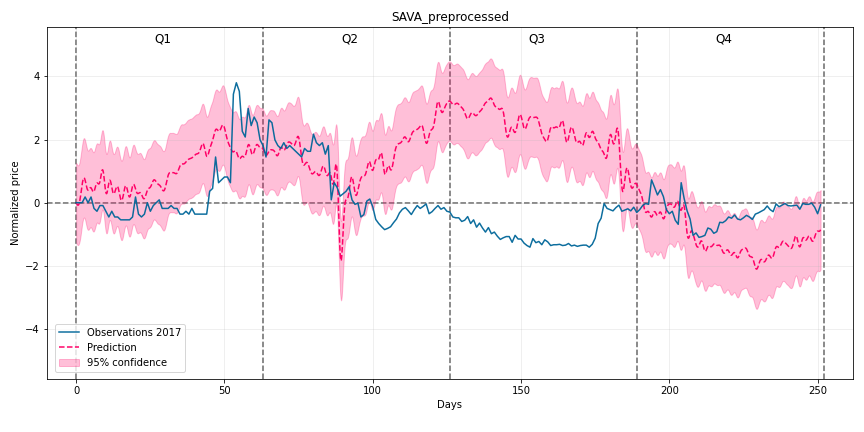
The forecasting problem is that given observations from the complete series [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=i%20%3D%201%2C...%2CN-1#0) and (optionally) from a partial last series [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5C%7By%5EN_t%5C%7D#0) , [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=t%20%3D%201%2C...%2CM_N#0), we want to extrapolate the last series until predetermined endpoint (usually a multiple of a quarter length during a year) - characterize the joint distribution of [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5C%7By%5E%7BN%7D_%7B%5Ctau%7D%5C%7D#0) , [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Ctau%20%3D%20M_N%20%2B%201%2C...%2CM_N%20%2B%20H#0) for some [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=H#0). We are also given a set of non-stochastic explanatory variables specific to each series, [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5C%7Bx%5Ei_t%5C%7D#0), where [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=x%5Ei_t%20%5Cin%20%5Cmathbb%7BR%7D%5Ed#0). Our objective is to find an effective representation of [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=P%5C(%5C%7By%5EN_%7B%5Ctau%7D%5C%7D_%7B%5Ctau%20%3D%20M_N%20%2B%201%2C...%2CM_N%20%2B%20H%20%7C%20%5C%7Bx%5Ei_t%2C%20y%5Ei_t%5C%7D%5E%7Bi%3D1%2C...%2CN%7D_%7Bt%3D1%2C...%2CM_i%7D%7D%5C)#0), with [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5Ctau%2C%20i#0) and [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=t#0) ranging, respectively over the forecasting horizon, the available series and the observations within a series.

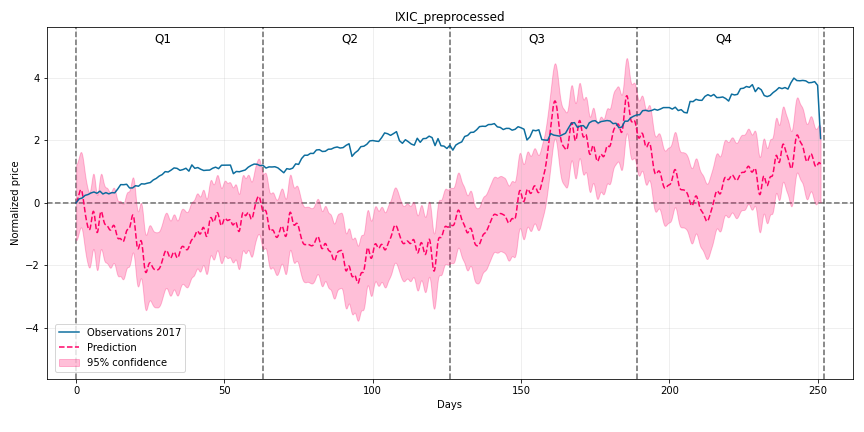
Everything mentioned in this section was implemented in Python using the wonderful library sklearn, mainly the sklearn.gaussian\_process.

## *Result and Comparison with other methods:*

### **Result on Gaussian Process:**

In 2020:





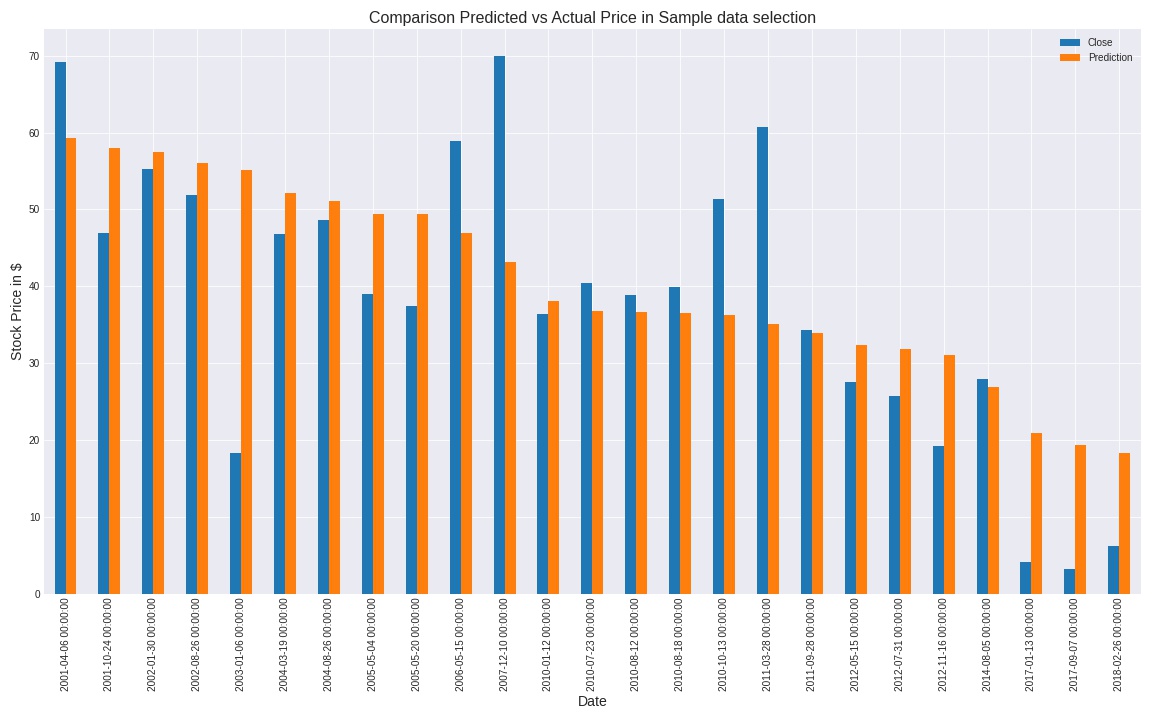
In 2021:

### **Comparison with other methods: Linear Regression, LSTM Network and ARIMA Model:**

#### Linear Regression

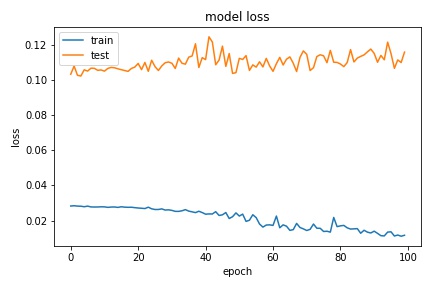
Training Result:

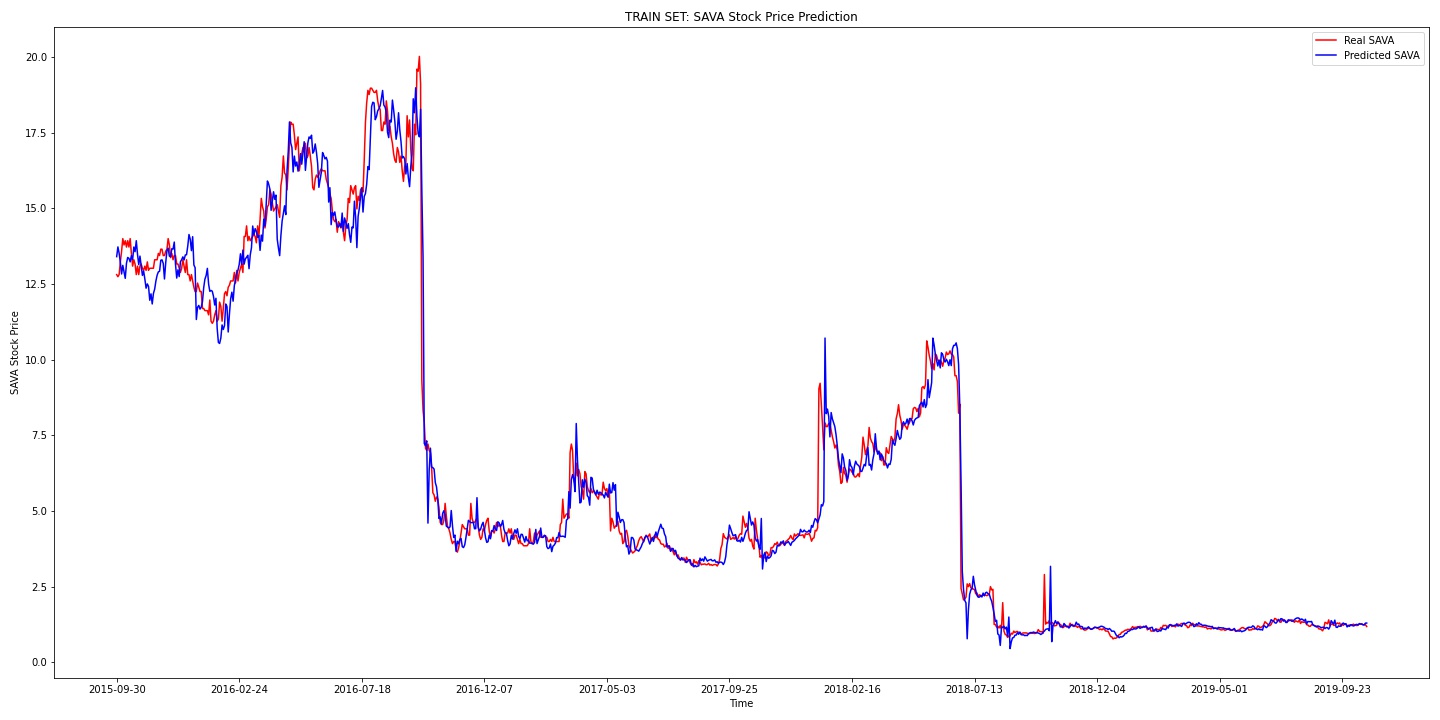


Test Result:

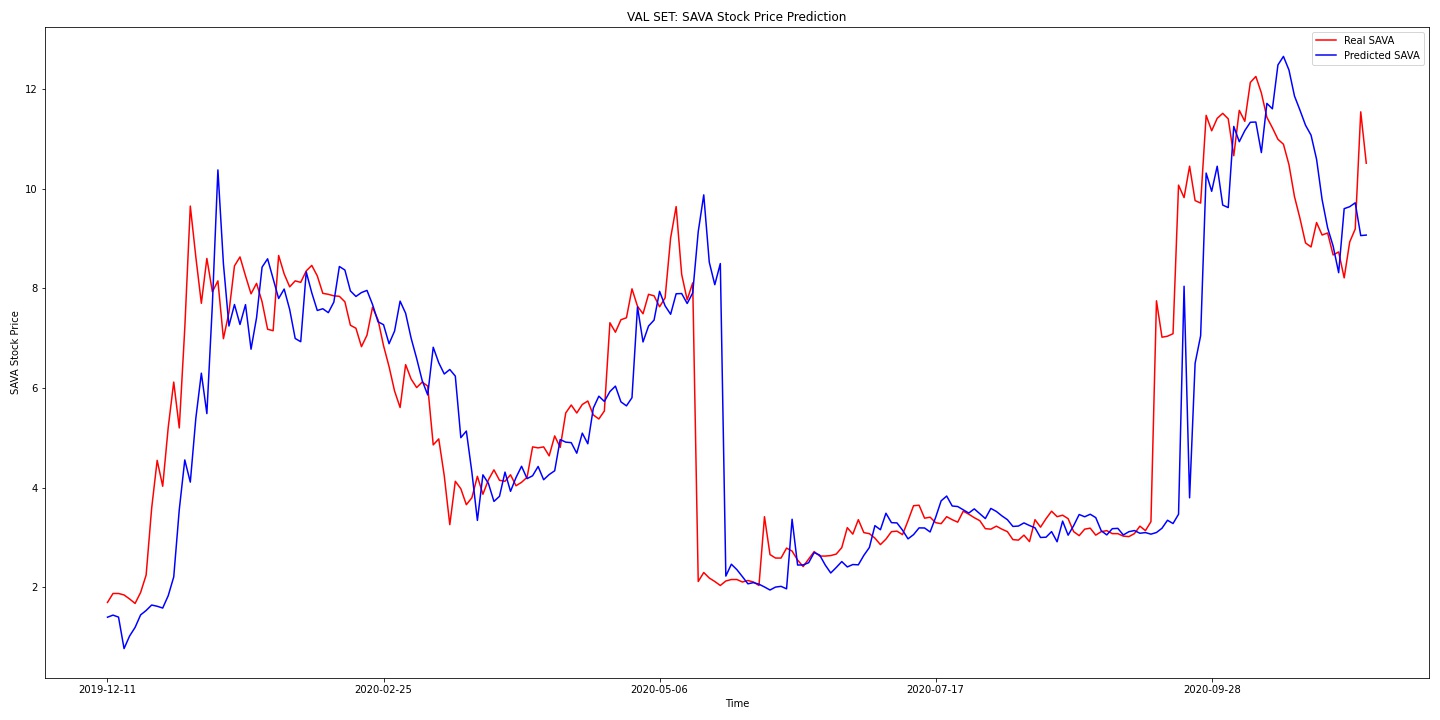
#### ***Long-Short Term Memory Network***

Training Result:

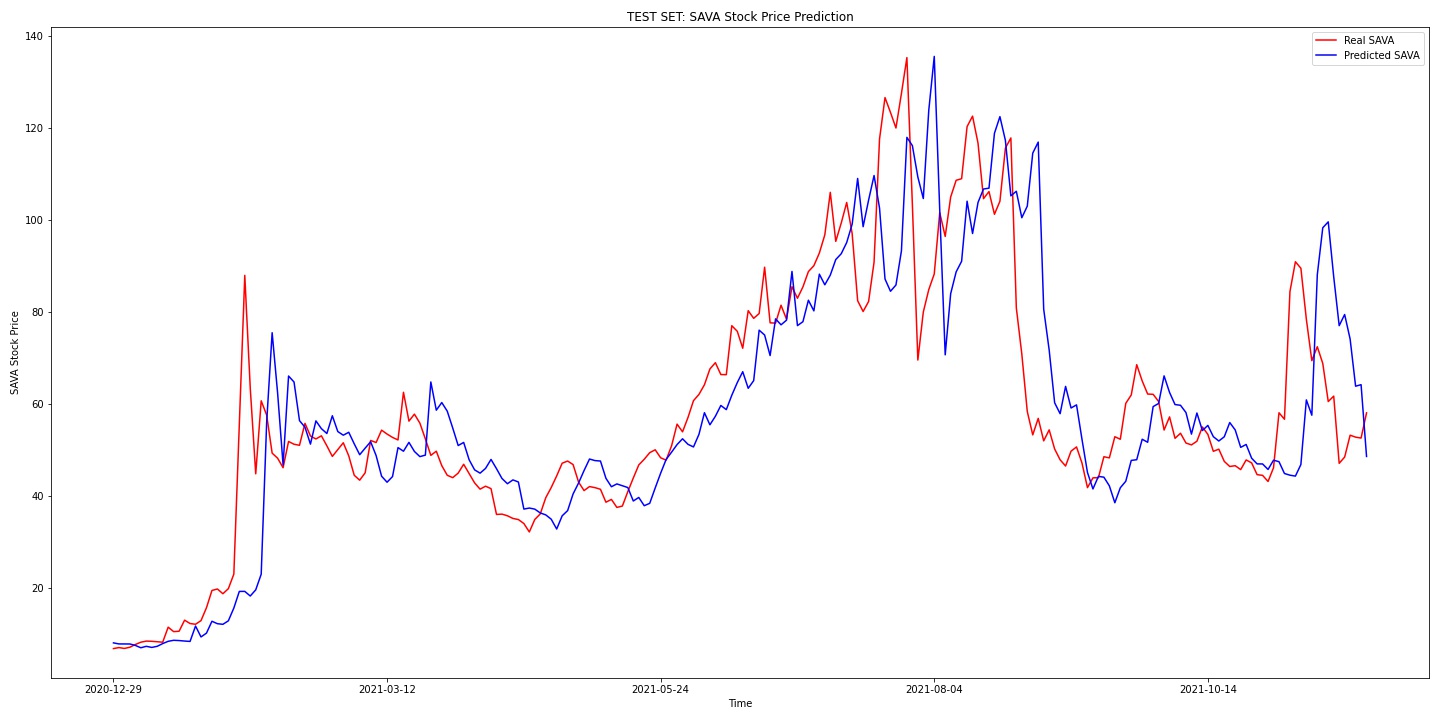




Validation Result:

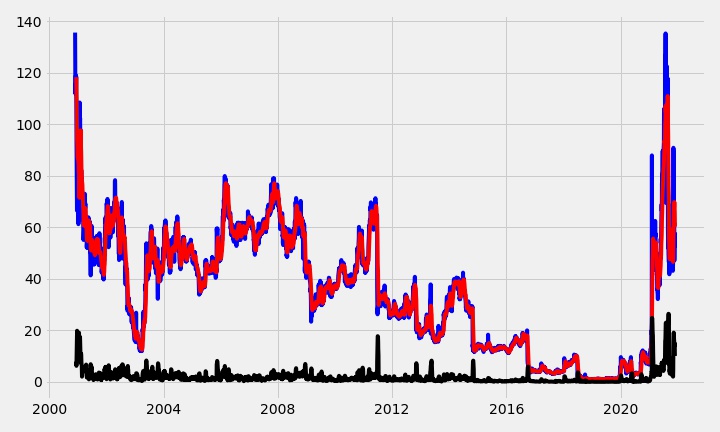


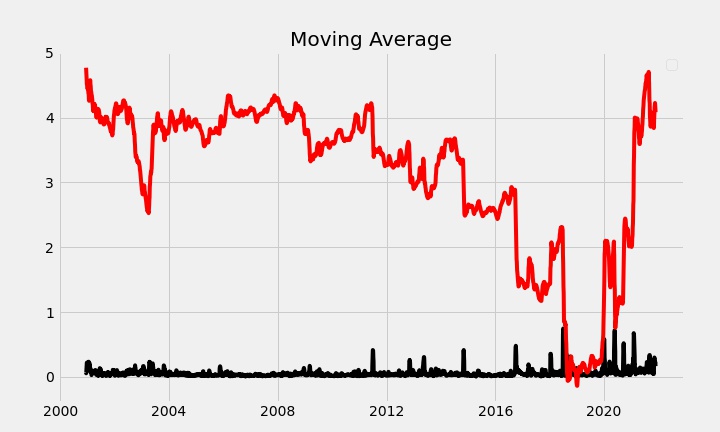
Test Result:



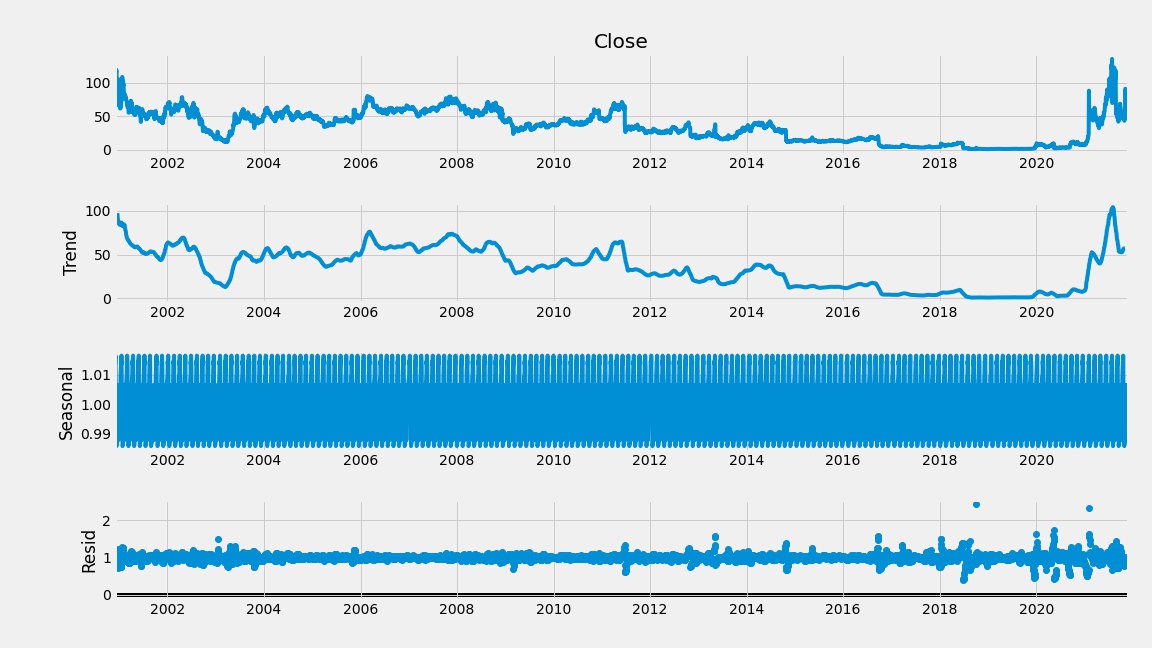
#### ARIMA Model

Testing stationarity:

Elimination Trend:



Seasonal Decompose:



Prediction Result:

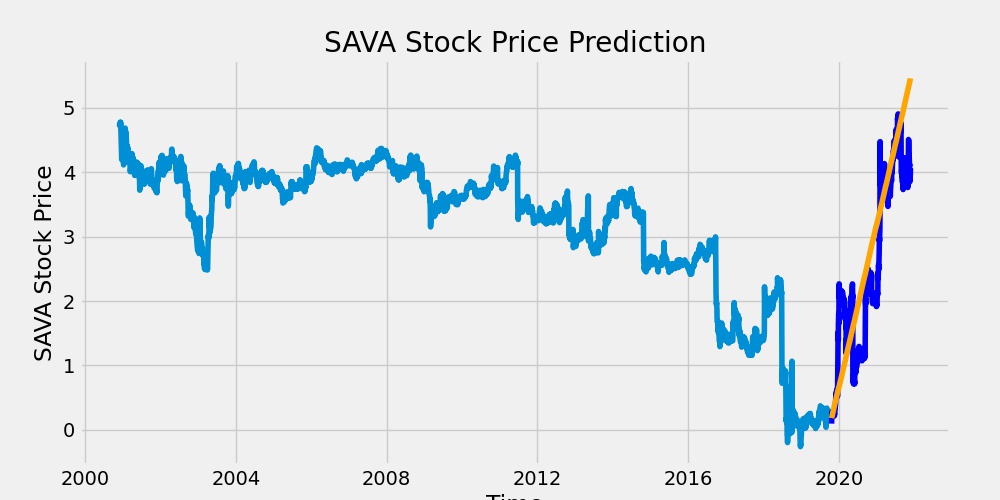


Table of result:

| Type | Gaussian Process | LSTM | Linear Regression | ARIMA model |
| --- | --- | --- | --- | --- |
| MSE | 152.5147257 | 28.7624804 | 282.5947357 | 0.56767490829 |
| MAE | 8.15730316 | 4.172498876 | 12.29430316 | 0.610451537752 |