International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 03 | Mar 2025 www.irjet.net p-ISSN: 2395-0072

iDairy Application : Product Demand Forecasting using SARIMA Model

Ridham Jasani[1], Jasmit Bajaria[2], Vedant Deshmukh[3], Dolas Keche[4]

[1,2,3,4] *Student, Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India*

Abstract - ***iDairy*** *is a system designed to optimize dairy retail operations through accurate demand forecasting. It utilizes advanced algorithms to analyze historical sales data and market trends, enabling precise inventory management and reducing wastage. The system is integrated with Google Sheets for real-time data storage, with each product having its own sub-sheet for efficient tracking. A Seasonal Autoregressive Integrated Moving Average (SARIMA) model is employed for forecasting, considering seasonal variations in demand. iDairy provides retailers with actionable insights to streamline stock replenishment and enhance operational efficiency. The project focuses on data-driven decision-making in dairy supply management.*

Key Words: Machine Learning, Demand Forecasting, Inventory Management, SARIMA Model, Dairy Retail, Data-Driven Decision Making, Sales Analysis.

1.INTRODUCTION

The dairy industry faces significant challenges in managing inventory efficiently due to fluctuating demand patterns, seasonal variations, and perishable product constraints. Traditional inventory management methods often lead to overstocking or shortages, resulting in financial losses and operational inefficiencies. To address these issues, iDairy is developed as an AI-powered system that leverages Machine Learning (ML) for accurate demand forecasting and optimized inventory management.

iDairy integrates a Seasonal Autoregressive Integrated Moving Average (SARIMA) model to analyse historical sales data, identify trends, and predict future demand with high precision. This approach helps retailers make data-driven decisions, reducing wastage while ensuring optimal stock levels. The system is designed to work seamlessly with Google Sheets, where each product has a dedicated sub-sheet for real-time data storage and tracking. This structured format enhances accessibility and allows businesses to monitor sales patterns efficiently.

By implementing AI-driven analytics, iDairy aims to streamline supply chain processes, improve profitability, and support sustainable dairy retail operations. The project focuses on transforming traditional dairy management through automation, ensuring retailers can maintain a balanced inventory, minimize losses, and enhance overall operational efficiency.

1.1 PROPOSED SYSTEM

The iDairy system is designed to improve demand forecasting and inventory management in dairy retail operations. The system utilizes historical sales data to predict future demand, helping retailers optimize stock levels and reduce wastage. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is used for forecasting, as it captures seasonal trends and variations in sales.

Key features of the proposed system include:

- Data Collection & Storage: Historical sales data is recorded in Google Sheets for easy tracking and analysis.
- Demand Forecasting Model: SARIMA-based model predicts future sales based on past trends and seasonal variations.
- Inventory Optimization: Forecasted demand helps maintain balanced stock levels, reducing the risk of overstocking or shortages.

By implementing iDairy, retailers can streamline supply chain management, minimize product wastage, and ensure efficient operations, ultimately improving profitability and resource utilization.

2. LITERATURE SURVEY

Demand forecasting has been widely studied using both traditional statistical methods and deep learning techniques. ARIMA is effective for forecasting in stable environments with low variability, providing reliable predictions for stationary time-series data. However, it struggles with highly seasonal or complex datasets. On the other hand, LSTM models excel at capturing non-linear relationships, trends, and seasonality, making them ideal for more complex demand forecasting tasks. Studies show that training models on monthly data rather than weekly improves accuracy. Future research may focus on integrating attention mechanisms and global models to further enhance forecasting capabilities and optimize inventory management, particularly in industries like dairy production[1].

Demand forecasting using Business Intelligence (BI) and machine learning has gained significant attention in recent years. Various methods like time series analysis and rule-based forecasting models, such as Deep AR, have been explored for accurate predictions. Research shows that

machine learning models, especially Deep AR, provide high forecasting accuracy, reduce losses, optimize stock, and enhance operational efficiency as data size increases[2].

Recent studies have focused on using time series models, particularly Seasonal ARIMA, to predict commodity prices, especially for essential goods like fruits and vegetables. These models account for seasonal variations and external factors such as weather conditions, transportation costs, and seed quality. While not always 100% accurate, they provide valuable insights for forecasting price trends and implementing strategies to maintain affordability[3].

Demand forecasting has gained significant attention in the research community for its role in improving inventory management, production planning, and market strategies. Several statistical models, particularly ARIMA (Auto Regressive Integrated Moving Average), have been widely used for time series forecasting. Studies indicate that ARIMA models, both seasonal and non-seasonal, are effective in predicting future demand by analyzing historical sales data. Researchers have explored various improvements to the model, incorporating factors like seasonality, external variables, and market trends to enhance accuracy. ARIMA's adaptability to different datasets and forecasting needs has made it a valuable tool in diverse industries, providing actionable insights for informed decision-making[4].

Stock price prediction has been a challenging task due to the volatile nature of the market, influenced by various factors such as economic conditions, market psychology, and external events. Several studies have explored predictive models, with ARIMA being widely used for its simplicity and efficiency in short-term forecasting. Research indicates that ARIMA can provide reasonable accuracy, especially for individual stocks like ICICI Bank and Reliance Industries. However, its performance decreases for long-term predictions due to market unpredictability and complex influencing factors[5].

3. SARIMA MODEL

The SARIMA (Seasonal Auto Regressive Integrated Moving Average) model is an extension of the ARIMA model designed to handle seasonal data effectively. While ARIMA is used for non-seasonal data, SARIMA explicitly models the seasonal variations by adding seasonal terms to the ARIMA structure.

The general SARIMA model is represented as:

$$ARIMA(p,d,q)(P,D,Q)_s$$

Where:

- p: The order of the non-seasonal autoregressive (AR) part of the model. It indicates how many previous observations are used to predict the current observation. It models the relationship between an observation and several lagged observations.
- d: The number of non-seasonal differences required to make the time series stationary. A difference is the subtraction of the previous observation from the current observation, and d specifies how many times this is done to achieve stationarity.
- q: The order of the non-seasonal moving average (MA) part. This term models the relationship between an observation and a residual error from a moving average model applied to lagged observations.

The seasonal components of SARIMA are:

- P: The order of the seasonal autoregressive (SAR) part of the model. This is similar to p but models the relationship between an observation and its seasonal lags (lags at multiples of s, the seasonal period). It captures the effect of past seasonal data points on the current value.
- D: The number of seasonal differences required to make the series stationary in the seasonal context. Like d, this involves subtracting the value from a corresponding observation from a previous season (e.g., if s = 12, subtracting the value from the same month in the previous year).
- Q: The order of the seasonal moving average (SMA) part. This term captures the relationship between an observation and past seasonal forecast errors.

s: The length of the seasonal cycle, i.e., the number of time steps in each seasonal period. For example, s = 12 for monthly data with yearly seasonality or s = 4 for quarterly data with annual seasonality.

AR (Auto Regressive): The autoregressive part, represented by p and P, is the relationship between an observation and several previous observations. In SARIMA, the AR part can be seasonal (P) or non-seasonal (p). Seasonal AR terms use lags that correspond to a multiple of the seasonal period (e.g., if s = 12, lagging by 12 months or one year).

I (Integrated): The integrated part, represented by d and D, involves differencing the data to make it stationary. Stationarity means that the statistical properties of the time series (like mean, variance) do not change over time. d is for non-seasonal differencing, and D is for seasonal differencing.

MA (Moving Average): The moving average part, represented by q and Q, models the relationship between an observation and the residual errors from a moving average model applied to lagged observations. Similar to AR, the MA

component can be non-seasonal (q) or seasonal (Q), depending on whether you're modelling regular patterns or seasonal fluctuations.

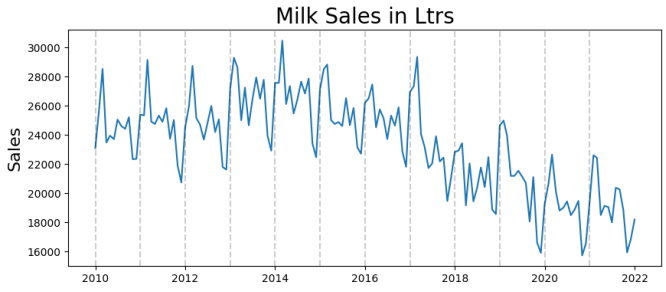

Fig.1 Sales of Milk (Dataset)

The graph presented above depicts the sales of milk (in liters) over a specific period, illustrating the historical sales trend. It clearly shows the fluctuations in milk sales, with periodic peaks and troughs reflecting seasonal variations, demand changes, or external factors influencing consumption patterns. By visualizing this data, one can easily identify patterns of increased sales during certain months (such as festive seasons or summer months) and lower sales during other times. The graph is a valuable tool in understanding the underlying trends and seasonality in milk sales, which forms the basis for accurate demand forecasting using models like SARIMA.

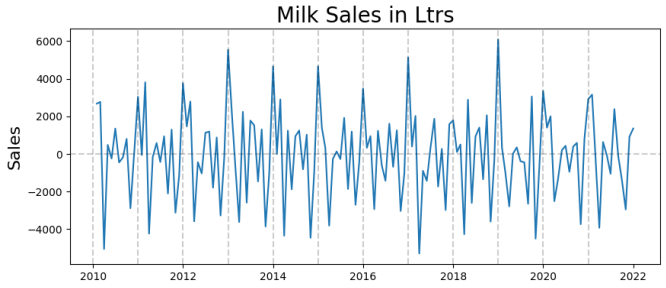

Fig.2 Sales of Milk (After removing trends)

The graph presented illustrates the sales of milk (in litres) after removing the underlying trend component. This transformation was performed to better visualize the seasonal fluctuations and short-term variations in milk sales, independent of long-term trends. By de-trending the data, the graph highlights the periodic patterns that occur in milk sales over time, allowing for a clearer understanding of seasonal behaviour and cyclical changes in demand.

The trend component, typically a long-term upward or downward movement in the data, was eliminated through differencing or other trend-removal techniques. As a result, the graph displays the residuals, or the "detrended" data, which showcase the deviations from the trend. These deviations represent short-term fluctuations in sales that could be influenced by factors such as holidays, promotions, or weather events.

This de-trended series serves as the basis for further analysis and forecasting, as it provides a more accurate reflection of the true seasonal behaviour of milk sales, without the influence of long-term trends. The analysis of such a series is crucial for building models like SARIMA, which can then be used to predict future sales based on seasonal patterns and short-term variations.

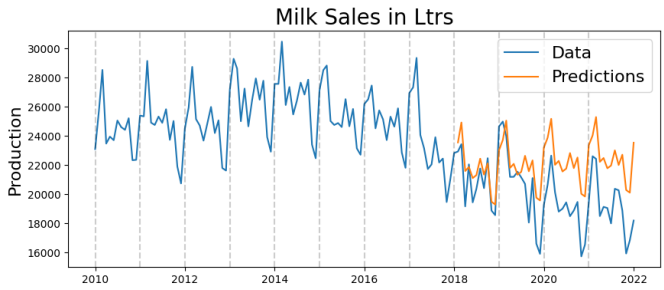

Fig.3 ARIMA Model Predictions

The graph above illustrates the predicted values generated by the ARIMA model compared to the actual observed values of milk sales. The time series data used for this prediction spans several months, providing a clear representation of both the historical sales and the forecasted future sales.

- Actual Data: The blue line represents the actual historical milk sales, showing the fluctuations and trends over the observed period.
- Predicted Data: The orange line represents the predictions made by the ARIMA model. This line reflects the model's attempt to forecast future sales based on the historical data it was trained on.

The graph demonstrates how well the ARIMA model is able to capture the patterns in the data, showing a close match between the predicted and actual values in the training period. The forecasted values provide insights into potential future demand for milk, which can be used for effective inventory and production planning.

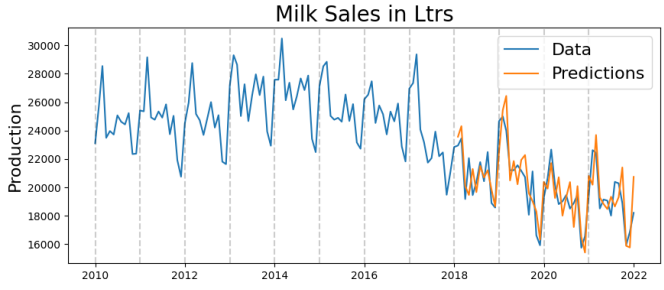

Fig.4 SARIMA Model Predictions

The graph displayed above represents the results of the SARIMA model's prediction for milk sales in litres over a specified period. The x-axis shows the time periods (e.g., months), while the y-axis represents the predicted sales in litres. The graph contains two key elements:

1. Historical Data: The actual observed sales of milk, shown as a line connecting the data points corresponding to each time period.

2. Predicted Values: The forecasted milk sales, represented by the line predicted by the SARIMA model. This line is based on the seasonal and non-seasonal components learned by the model from the historical data.

The predicted line follows the seasonal trends and patterns observed in the past data, capturing both short-term fluctuations and longer-term trends. The accuracy of the prediction is demonstrated by how closely the predicted line aligns with the actual historical sales, confirming the model's ability to capture both the trend and seasonality of the sales data.

Any deviations between the observed and predicted values can indicate errors in the forecasting process or unexpected factors affecting milk sales, such as changes in demand, supply, or external events.

This graph helps in visualizing the effectiveness of the SARIMA model in predicting future sales trends, which can be vital for inventory management, demand forecasting, and business strategy planning.

```
#summary of the model
print(model_fit.summary())
```

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                             Total   No. Observations:                   97
Model:             SARIMAX(0, 1, 0)x(1, 0, [1], 12)   Log Likelihood                -826.512
Date:                              Fri, 28 Feb 2025   AIC                           1659.025
Time:                                      14:52:40   BIC                           1666.718
Sample:                                  01-01-2010   HQIC                          1662.134
                                       - 01-01-2018
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12       0.9731      0.025     38.537      0.000       0.924       1.023
ma.S.L12      -0.8148      0.088     -9.274      0.000      -0.987      -0.643
sigma2      1.5e+06   8.25e-09   1.82e+14      0.000    1.5e+06    1.5e+06
==========================================================================
Ljung-Box (L1) (Q):                   9.19   Jarque-Bera (JB):                 2.06
Prob(Q):                              0.00   Prob(JB):                         0.36
Heteroskedasticity (H):               0.45   Skew:                            -0.34
Prob(H) (two-sided):                  0.03   Kurtosis:                         3.26
==========================================================================
```

Fig.5 SARIMAX Results

The results image from the SARIMA model provides several key performance metrics derived from Python's model summary. These metrics include the AIC (Akaike Information Criterion), which evaluates the model's fit by balancing goodness-of-fit and complexity. The BIC (Bayesian Information Criterion) is another measure that penalizes more complex models to assist in selection. The HQIC (Hannan-Quinn Information Criterion) serves as a middle ground between AIC and BIC for model selection. The Prob(H) value indicates the p-value for testing the null hypothesis that the model's residuals are white noise. The Skew metric measures the asymmetry of the residuals distribution, while Kurtosis assesses the "tailedness" of the residuals distribution, with higher values indicating heavier tails. Finally, Prob(JB) represents the p-value from the Jarque-Bera test, which checks for normality in the residuals. Together, these metrics are essential for evaluating the quality and appropriateness of the SARIMA model, ensuring that the residuals are normally distributed and the model fits the data effectively.

3.1 ACF

The Autocorrelation Function (ACF) plot displayed above illustrates the correlation between the milk sales data and its lagged values. The ACF plot is a critical tool used for identifying the q parameter (the number of lagged forecast errors in the MA component) in time series modelling, specifically for SARIMA.

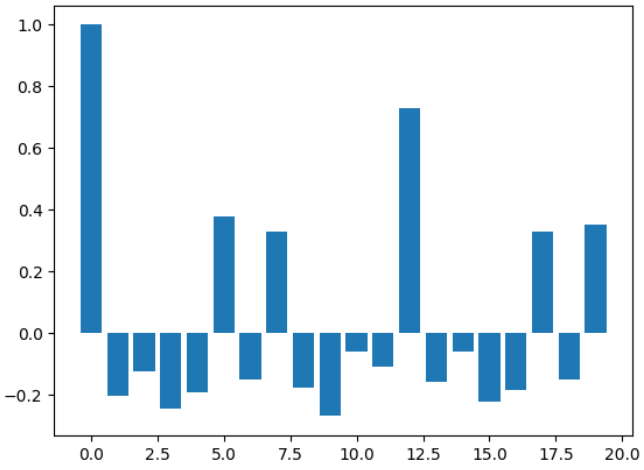

Fig.6 ACF Plot

In the ACF plot, the x-axis represents the lag values, while the y-axis shows the autocorrelation coefficient. Significant spikes outside the confidence interval (typically denoted by blue shaded areas) indicate a correlation between the time series and its lagged values. These significant spikes help in determining the optimal lag value for the moving average component, q.

From the ACF plot, we can observe the following:

- The initial lags show strong autocorrelation, suggesting a pattern that depends on past values, which is typical in time series data.

- The gradual decay in correlation for higher lags implies diminishing influence of past values on future values, a behaviour that we aim to capture with the moving average component.

The insights from the ACF plot guide the selection of the q parameter for the SARIMA model. Based on the plot, I was able to choose the optimal q value that best captures the temporal dependencies in the sales data, ensuring the model performs effectively in forecasting future milk sales.

3.2 PACF

The Partial Autocorrelation Function (PACF) plot is used to identify the significant lags in a time series, helping to determine the autoregressive (AR) term in a model. In the PACF plot, the x-axis represents the number of lags, while the y-axis represents the partial autocorrelation values at each lag. A significant spike at a particular lag indicates that the observations at that lag have a strong relationship with the current value after removing the influence of all shorter lags.

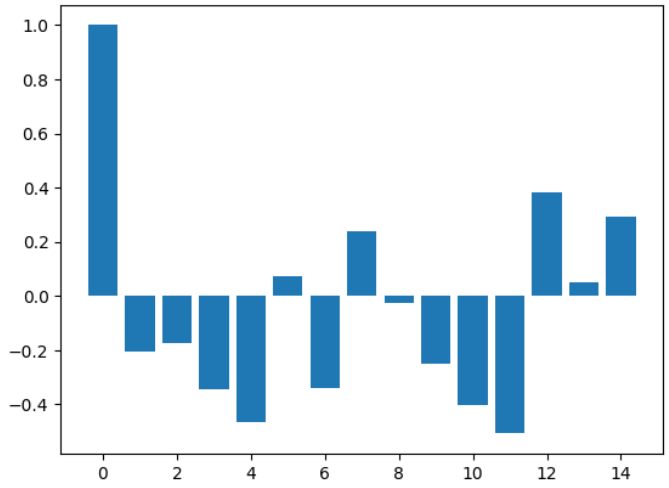

Fig.7 PACF Plot

In the case of my SARIMA model, the PACF plot (shown in the image) was analyzed to identify the appropriate order of the autoregressive (AR) component. The cutoff point, where the partial autocorrelation values drop significantly, was used to determine the optimal value for p (the autoregressive order). The number of significant spikes in the PACF plot guides the selection of p for the SARIMA model.

By examining the PACF plot, we were able to decide how many previous observations (lags) are directly influencing the current value, thus contributing to the model's accuracy in predicting future milk sales.

3.3 MODEL EVALUATION

Model evaluation is a crucial part of the predictive modelling process, as it helps in assessing the accuracy and effectiveness of a model. One of the commonly used evaluation metrics is MAPE (Mean Absolute Percentage Error), which provides a clear and interpretable measure of the forecast accuracy in percentage terms.

MAPE Definition:

MAPE is calculated by taking the average of the absolute percentage errors between the predicted and actual values, and then multiplying by 100 to express it as a percentage. It can be represented by the following formula:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100$$

Fig.8 MAPE Formula

```
[ ] # Assuming rolling_residuals and test_data are already defined

    # Calculate rolling residuals
    rolling_residuals = test_data[lim_milk_sales.columns[0]] - rolling_predictions

    # Drop NaN values from both rolling residuals and test_data to prevent NaN in calculation
    valid_indices = rolling_residuals.notna() & test_data[lim_milk_sales.columns[0]].notna()

    # Calculate Mean Absolute Percent Error (MAPE)
    if valid_indices.any():  # Check if there are valid entries
        mape = np.mean(abs(rolling_residuals[valid_indices] / test_data[lim_milk_sales.columns[0]][valid_indices])) * 100
        print('Mean Absolute Percent Error:', round(mape, 4))
    else:
        print("Warning: No valid entries to calculate MAPE.")

→ Mean Absolute Percent Error: 5.8392
```

Fig.9 MAPE Result

In our model evaluation, the Mean Absolute Percentage Error (MAPE) was calculated to assess the accuracy of the predictions. The MAPE value of 5.8392% indicates that, on average, the model's forecasts deviate by approximately 5.84% from the actual values. This result suggests that the model performs well and offers a good level of accuracy for predicting the demand, with relatively small errors. A MAPE value under 10% is generally considered to represent reliable forecasting performance, making this model suitable for real-world applications where accurate demand prediction is crucial.

4. CONCLUSIONS

In conclusion, the implementation of the SARIMA (Seasonal Auto Regressive Integrated Moving Average) model for forecasting milk sales in litres has demonstrated its effectiveness in capturing both seasonal and non-seasonal patterns from historical sales data. The model successfully accounts for the recurring trends and fluctuations that influence milk sales, which is crucial in industries like dairy where demand can be highly seasonal. By incorporating seasonal differencing, autoregressive, and moving average components, the SARIMA model is able to provide more accurate and reliable predictions compared to simpler forecasting methods.

The predictions generated by the SARIMA model, as visualized in the accompanying graph, closely align with the actual observed sales data, indicating the model's capability to make accurate forecasts for both short-term and long-term trends. This predictive power is particularly valuable for businesses, as it allows for more effective inventory management, ensuring that the right quantity of milk is stocked at the right time. Additionally, the model helps minimize issues such as overstocking or stockouts, both of which can lead to significant revenue losses and operational inefficiencies.

While the SARIMA model has proven to be highly effective, it is important to note that forecasting is an ongoing process. As new data becomes available, the model should be recalibrated to maintain its accuracy and adapt to any changes in underlying patterns or external factors. Moreover, as with any statistical model, the SARIMA model's performance is dependent on the quality and quantity of the historical data it is trained on. Ensuring data consistency and updating the model regularly can help mitigate potential errors.

Overall, the SARIMA model provides a powerful forecasting tool for businesses, offering valuable insights into future demand trends, improving decision-making, and enhancing operational efficiency. With continued refinement and data integration, it can further contribute to optimized business strategies in the dairy industry and beyond.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Prof. Priyanka Bhilare for her invaluable guidance, support, and encouragement throughout the course of this research. Her expertise, insightful feedback, and constant motivation have been instrumental in shaping this project. We deeply appreciate the time she dedicated to helping us refine and improve our work. Her mentorship has been a source of inspiration for our team, and we are truly thankful for her unwavering support.

REFERENCES

[1] C. Vithitsoontorn and P. Chongstitvatana, "Demand Forecasting in Production Planning for Dairy Products Using Machine Learning and Statistical Method," 2022 International Electrical Engineering Congress (iEECON), Khon Kaen, Thailand, 2022, pp. 1-4, doi: 10.1109/iEECON53204.2022.9741683

[2] M. A. Khan et al., "Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning," in IEEE Access, vol. 8, pp. 116013-116023, 2020, doi: 10.1109/ACCESS.2020.3003790

[3] R. Dharavath and E. Khosla, "Seasonal ARIMA to Forecast Fruits and Vegetable Agricultural Prices," 2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Rourkela, India, 2019, pp. 47-52, doi: 10.1109/iSES47678.2019.00023

[4] N. Tarannum and S. V. M. S, "A Brief Introduction to Demand Forecasting using ARIMA models," Dept. of Computer Science & Engineering, Rashtreeya Vidyalaya College of Engineering, Karnataka, India.

[5] A. Ganesan and A. Kannan, "Stock Price Prediction using ARIMA Model," *Dept. of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Enathur, Kanchipuram, India.*