Terminal Assignment-Based Assessment

Individual Project

PART A – Time Series Analysis, PART B – Logistic Regression

Stamatios karvounis
PGDip in Data Analytics January 2023
National College of Ireland
Dublin, Ireland
x18197051@student.ncirl.ie

PART A – Time Series Analysis

This project aims to provide suitable time series models for two temperature datasets from Armagh, covering monthly data from January 1844 to December 2004 and yearly data from 1844 to 2004. The preliminary assessment of the raw time series will be conducted using appropriate visualizations. The three categories of time series models to be estimated and discussed are Exponential Smoothing, ARIMA/SARIMA, and Simple time series models, with diagnostic tests and checks undertaken as necessary. Using data up to and including 2003 as training sets, forecasts will be generated for average temperatures in 2004 using both monthly and yearly data. The forecasted results will be evaluated against the actual data for 2004, and the optimal model selected based on its adequacy for forecasting purposes.

I. Preliminary Assessment of the Nature and Components

In order to gain an understanding of the nature and components of the raw time series data, a preliminary assessment was conducted using appropriate visualizations for both the monthly and yearly average temperature time series. The monthly time series covers the period from January 1844 to December 2004, while the yearly time series spans from 1844 to 2004. The visualizations included time plots, boxplots and decomposition plots to identify any trends, patterns and seasonality. These preliminary assessments were necessary to gain insights into the underlying structure of the data and to guide the selection of appropriate time series models for each dataset.

A. Create Time Series

To facilitate the estimation and evaluation of suitable time series models for each dataset, time series objects for the monthly and yearly average temperature time series were created. The time series object for the monthly time series was named "tsm," and it covers the period from January 1844 to December 2004, with a total of 1921 observations. The time series object for the yearly time series was named "tsy," and it covers the period from 1844 to 2004, with a total of 161 observations. These time series objects were created using the appropriate functions and commands in R, and were necessary for subsequent analysis, including the estimation and evaluation of suitable time series models. By creating these time series objects, it became easier to manipulate and visualize the data, as well as to conduct diagnostic tests and checks to ensure that any models estimated were appropriate and adequate for forecasting purposes.

B. Visualization

Visualizations were used to gain a comprehensive understanding of the time series, allowing for the identification of any underlying patterns or trends and the detection of potential anomalies or outliers that could impact subsequent modeling and forecasting efforts.

1) Line Plots

For the monthly time series “tsm”, it is observed that we have seasonality in our data (Figure 1). The high spikes in the middle of the year and low spikes at the end/beginning of the year suggest that there is a recurring pattern in the data that follows a seasonal cycle. The seasonality in the “tsm” plot indicates that there is a pattern that repeats itself over a fixed period of time, and this pattern is likely linked to the seasonal variations in temperature. Seasonality can be caused by a number of factors, including changes in the amount of sunlight, changes in the weather patterns, and variations in the amount of precipitation. By observing the seasonality in the tsm plot, we can gain insights into the underlying factors that are driving the temperature patterns over time.

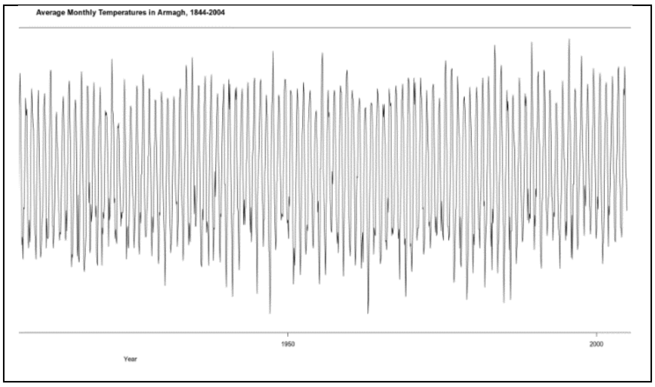


Fig. 1. Monthly Average Temperature Time Series

For the yearly time series “tsy”, Figure 2 displays a relatively stable pattern of temperature variation over time, with no discernible long-term trend or recurring pattern. While there may be some short-term fluctuations in temperature, these do not appear to follow a regular or predictable cycle, nor do they suggest any underlying trend or pattern. This lack of clear seasonality or trend in the data suggests that temperature variations may be influenced by a complex array of factors, some of which may be difficult to identify or quantify.

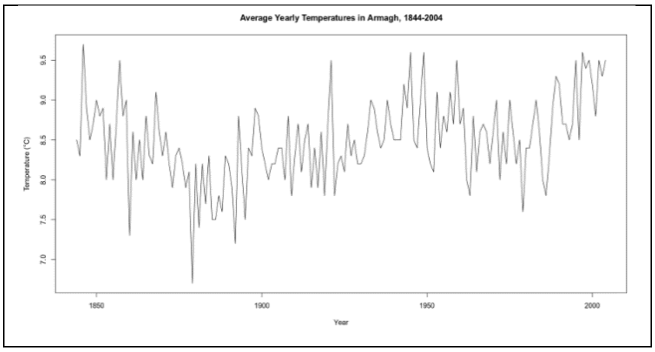


Fig. 2. Yearly Average Temperature Time Series

2) *Boxplots*

The next step in the analysis was to create boxplots for both the monthly and yearly average temperature time series. Boxplots provide a useful summary of the distribution of the data, showing the median, quartiles, and any outliers that may be present.

Upon examining the boxplot for both the monthly average temperature time series "tsm" (figure 3), we can see that the data points fell within the whiskers of the box plot, which represent the interquartile range. This indicates that the data is relatively tightly clustered around the median, and there are no extreme values that are significantly different from the rest of the data. The absence of outliers suggests that the data is relatively free of extreme values that could skew our analysis.

The boxplot analysis of the yearly average temperature time series "tsy" reveals the presence of two outliers, indicating that the majority of the temperature values fall within a relatively narrow range (figure 3). The outliers, which are distant from the rest of the data points, may signify unusual or extreme temperature values for those particular years. This suggests that the yearly time series does not have a specific trend, pattern, or seasonality, as indicated by the absence of a discernible distribution or clustering of data points.

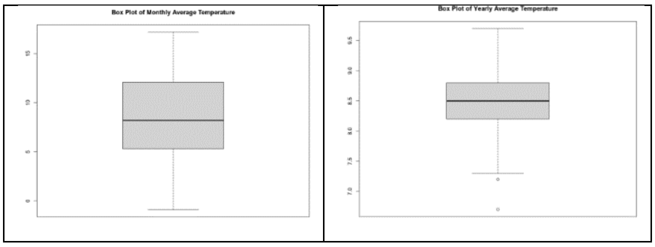


Fig. 3. Time Series Boxplots, Monthly left, Yearly right

3) *Decomposition Plots*

The decomposition plot is a useful tool for visualizing the individual components of a time series, including trend, seasonality, and residual. By decomposing the time series into these components, it is possible to identify any underlying patterns or trends that may not be apparent in the raw data.

In the case of the time series being analyzed (figure 4), the remainder plot shows no noticeable patterns, indicating that the model is effectively capturing the underlying patterns in the data. The seasonal plot displays a regular pattern that repeats over a fixed period of time, with a relatively large seasonal component indicating that seasonality is an important factor in the time series. However, the trend component of the time series is relatively small and clustered around a stable line, suggesting that the long-term trend is not very pronounced.

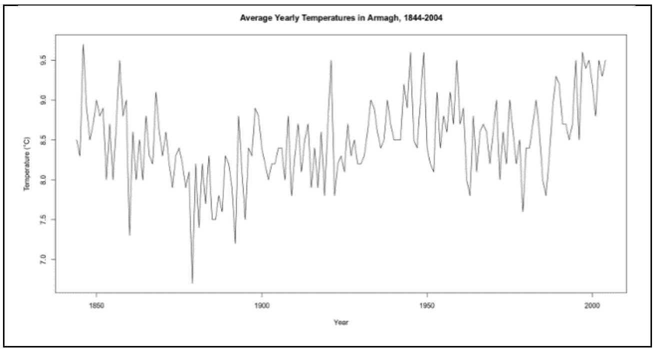


Fig. 4. Decomposition plot for monthly time series

II. Diagnostic tests and checks

In order to identify the most appropriate time series model, it is necessary to conduct diagnostic tests and checks on the data. These tests are designed to detect any patterns or trends in the time series data, and to determine whether the data is stationary or non-stationary. The most commonly used diagnostic tests include the autocorrelation function (ACF), partial autocorrelation function (PACF), and the augmented Dickey-Fuller (ADF) test. By performing these diagnostic tests, it is possible to identify the most appropriate time series model for the data, and to ensure that the model accurately captures the underlying patterns and trends in the data.

A. *ACF/PACF*

Autocorrelation function (ACF) and partial autocorrelation function (PACF) are important diagnostic tools for understanding the patterns of correlation in a time series. The ACF measures the correlation between a time series and its lagged values, while the PACF measures the correlation between a time series and its lagged values after removing the effects of earlier lags.

For the monthly time series, the autocorrelation function (ACF) plot shows 5 spikes followed by 5 troughs, suggesting a repeating pattern or seasonality that occurs every 5 lags. This could mean that the time series has a seasonal component that repeats every 5 months. The partial autocorrelation function (PACF) plot shows a significant spike at lag 1, indicating a strong correlation between adjacent time points. The remaining spikes and troughs are within the confidence interval, suggesting no significant correlation after the first lag (figure 5).

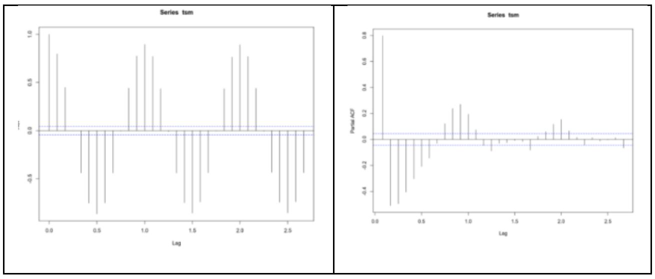


Fig. 5. ACF, PACF for monthly time series

On the other hand, for the yearly time series, the ACF plot shows a significant peak at lag 1, suggesting a strong correlation between adjacent time points. The small wave from lags 2 to 15 indicates that there might be some seasonality in the data, but the pattern is not clear. The values are very small for lags 15 to the end, indicating a weak correlation between points separated by multiple lags.

Overall, the ACF plot does not provide a clear indication of strong seasonality or trend in the data. The PACF plot shows no significant spikes outside the confidence interval, indicating no significant correlation beyond the first lag (figure 6).

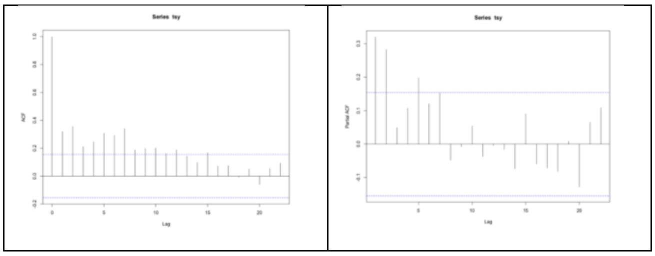


Fig. 6. ACF, PACF for yearly time series

B. *ADF*

The Augmented Dickey-Fuller (ADF) test is a statistical test that is commonly used to determine the stationarity of a time series. The test is based on the null hypothesis that the time series contains a unit root, which implies that the series is non-stationary. On the other hand, the alternative hypothesis is that the time series is stationary. For the monthly time series “tms”, ADF test was applied and obtained a test statistic value of -13.858, which is lower than the critical values for all levels of significance. Moreover, the p-value obtained from the test was < 2.2e-16, which indicates strong evidence against the null hypothesis. Therefore, it can be concluded that the time series is stationary at the 1%, 5%, and 10% levels of significance. In other words, the data does not exhibit any significant trend or seasonality that could affect the behavior of the time series over time. The ADF test results support the assumption of stationarity in the monthly time series data.

Furthermore, the results of the ADF test for the yearly time series show that the test regression includes a constant term, but no trend variable. The ADF test statistic is -3.0673, which is smaller than the critical values at the 1% level for all three test statistics. The p-value of the test is 0.02851, which is also below the significance level of 0.05, providing further evidence that the time series is stationary. Therefore, it is concluded that the yearly time series is stationary.

III. Estimation and Discussion of Time Series Models

The next step in the analysis is the estimation and discussion of time series models. In this stage, different approaches such as Exponential Smoothing, ARIMA/SARIMA, and simple time series models are used to identify the best fitting model for the data. Each model is evaluated using appropriate diagnostic tests and measures such as AIC, BIC, and MASE. This will allow the comparison of the accuracy of the different models and determine the most suitable one for the time series data.

A. *Exponential Smoothing*

Exponential smoothing is a popular method used for forecasting time series data. It is a time series forecasting technique that uses a weighted average of past observations to predict future values. In this analysis, the performance of four exponential smoothing models was compared: simple exponential smoothing, Holt's linear exponential smoothing, and Holt-Winters' additive and multiplicative exponential smoothing. Diagnostic tests such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Mean Absolute Scaled Error (MASE) were used to determine the best model fit. These tests help to compare the accuracy of the models and select the one that provides the best forecast for the given time series data.

1) *Simple exponential smoothing*

The simple exponential smoothing model was applied to the monthly time series data, using a smoothing parameter (alpha) of 0.9999 and an initial level (l) of 4.5107. Evaluation of the model's performance revealed small values for the mean error (ME) and root mean squared error (RMSE), indicating a good fit to the data. The mean absolute error (MAE) of 2.011637 suggests that the model's forecast is off by an average of about 2 units. The model was then used to generate forecasts for the next 12 months. The point forecast for each month is 6.2002, and the confidence intervals provide a range of values within which the actual values are likely to fall with a certain level of probability. Based on the small errors and good forecast accuracy, the simple exponential smoothing model appears to provide a good fit to the monthly time series data.

Furthermore, the simple exponential smoothing was applied to the yearly temperature time series. The model's performance was evaluated with the same diagnostic measures as the monthly time series. The results showed that the model fits the data relatively well, with a small RMSE of 0.460 and a MAPE of 4.19%. The MASE of 0.719 indicated that the model outperforms a naive model that uses the previous observation as the forecast for all future periods. The forecasts produced by the model suggest that the temperature is expected to remain stable over the next 1 year, with an average temperature of 9.213°C. The prediction intervals (80% and 95%) provide an idea of the range of uncertainty in the forecasts. Overall, the results indicate that the simple exponential smoothing model is a useful tool for forecasting yearly temperature data. The model provides accurate forecasts while also accounting for the uncertainty inherent in predicting future values.

2) *Holt's linear exponential smoothing*

Holt's linear exponential smoothing method is a popular forecasting technique used for time series data that exhibits a trend. This method models the time series as a linear function of time and applies exponential smoothing to both the level and slope of the series.

When applied to monthly average time series data, Holt's linear exponential smoothing method exhibited a small mean error (ME) of -0.0766, indicating that the model's forecasts underestimated the actual values on average. The root mean squared error (RMSE) was 2.4402, suggesting that the model's forecasts deviated from the actual values by an average of 2.4402 units. The mean absolute error (MAE) was 2.0127, indicating that the model's forecasts had an average deviation of 2.0127 units from the actual values. The mean percentage error (MPE) was -Inf, suggesting that the model's forecasts were biased towards underestimating the actual values. Overall, the error measures suggest that the Holt's linear exponential smoothing method did not perform very well on the monthly average time series data, as the model's forecasts tended to underestimate the actual values and had high levels of error. However, the MASE indicated that the model performed slightly better than a naive forecast.

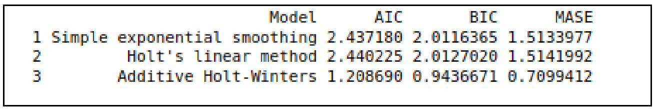
On the other hand, when applied to yearly time series data, Holt's linear exponential smoothing method had a small mean error (ME) of 0.003, indicating that the model's forecasts were slightly biased high, meaning that the forecasts from the model tended to be higher than the actual values on average. The RMSE was 0.460, indicating that the average forecast error was around 0.46 units. The MAE was 0.352, indicating that on average, the model's forecasts were off by 0.352 units. The MPE was -0.254, indicating that the forecasts were slightly biased low. The MAPE was 4.182%, indicating that on average, the model's forecasts were off by 4.182%. Finally, the MASE was 0.716, indicating that the model's forecasts were better than the naive forecast, as the value was less than 1. Overall, Holt's linear exponential smoothing method provided better results when applied to the yearly time series data, as indicated by the lower error measures and the MASE value of less than 1. However, it is important to note that the biases in the ME and MPE values suggest that the model may need to be adjusted to account for these tendencies.

3) Holt-Winters' additive and multiplicative exponential smoothing

Holt-Winters' additive method was applied to monthly time series data to provide estimates of the level, trend, and seasonal components of the time series, as well as forecasts and prediction intervals for the next 24 months. The model used smoothing parameters to control the degree of smoothing applied to the level, trend, and seasonal components of the time series. The error measures calculated for the training set indicate that the model provides a good fit to the training data, with relatively low levels of error and autocorrelation of the residuals. The point forecasts and prediction intervals for the next 24 months were also provided by the model. Overall, Holt-Winters' additive method appears to be a good performing model for forecasting the monthly time series data, as it provides accurate estimates of the level, trend, and seasonal components of the time series, as well as reliable forecasts and prediction intervals.

4) Diagnostic tests AIC, BIC, and MASE

For the average monthly temperature time series, three exponential smoothing models were used, namely Simple Exponential Smoothing (SES), Holt's Linear Method, and Additive Holt-Winters. The AIC and BIC values of the three models in figure 7, indicate how well they fit the data, with lower values indicating a better fit. The MASE (Mean Absolute Scaled Error) is a measure of the accuracy of the models, with lower values indicating better accuracy.

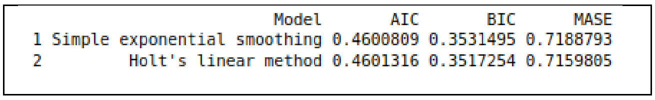


```
                         Model      AIC       BIC      MASE
1 Simple exponential smoothing 2.437180 2.0116365 1.5133977
2          Holt's linear method 2.440225 2.0127020 1.5141992
3        Additive Holt-Winters 1.208690 0.9436671 0.7099412
```

Fig. 7. Diagnostic tests AIC, BIC, and MASE for monthly time series

Among the three models, the Additive Holt-Winters model has the lowest AIC and BIC values (1.208690 and 0.9436671, respectively) and the lowest MASE value (0.7099412). This indicates that the Additive Holt-Winters model provides the best fit and the highest level of accuracy compared to the other two models. Therefore, we can conclude that the Additive Holt-Winters model is the best model for forecasting the average monthly temperature time series. It provides the most accurate and reliable forecasts based on the available data.

For the yearly time series, two exponential smoothing models were considered, namely Simple Exponential Smoothing and Holt's Linear Method. Both models produced similar AIC and BIC values, with the Simple Exponential Smoothing model having slightly lower values than Holt's Linear Method. This suggests that the Simple Exponential Smoothing model may provide a slightly better fit to the data than Holt's Linear Method (figure 8). When looking at the MASE value, however, we see that Holt's Linear Method has a slightly better out-of-sample forecast accuracy compared to Simple Exponential Smoothing. This indicates that Holt's Linear Method may provide slightly more accurate predictions for future data points. Overall, both models appear to be reasonable choices for modeling the average yearly temperature time series, as they produced similar results. However, Holt's Linear Method may be slightly preferable based on its slightly better out-of-sample forecast accuracy. It is important to note that the difference in performance between the two models is small, and additional analysis may be required to determine the most appropriate model for this time series.



```
                         Model       AIC       BIC      MASE
1 Simple exponential smoothing 0.4600809 0.3531495 0.7188793
2          Holt's linear method 0.4601316 0.3517254 0.7159805
```

Fig. 8. Diagnostic tests AIC, BIC, and MASE for yearly time series

B. ARIMA/SARIMA

ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) are popular models used for time series analysis and forecasting and their application will be explained below.

1) Yearly time series

The performance of three different ARIMA models fitted to the yearly average temperature time series. The models had orders (1,1,1), (0,1,1), and (0,1,2), respectively. The ARIMA(1,1,1) model included an autoregressive term of lag 1 (AR1), a moving average term of lag 1 (MA1), and first differencing. The coefficients of the model indicated a negative correlation between the current temperature and its previous value (AR1 coefficient = -0.042) and a negative correlation between the current temperature and the previous residual (MA1 coefficient = -0.8123). The model had an estimated sigma^2 of 0.2127, a log-likelihood of -103.78, and an AIC of 213.56. The training set error measures, including ME, RMSE, MAE, MPE, MAPE, MASE, and ACF1, indicated that the model had a reasonably good fit. The Box-Ljung test also indicated that there was no evidence of serial correlation in the residuals.

The ARIMA(0,1,1) model included a moving average term of lag 1 (MA1) and first differencing. The coefficient of the model indicated a negative correlation between the current temperature and the previous residual (MA1 coefficient = -0.8238). The model had an estimated sigma^2 of 0.213, a log-likelihood of -103.89, and an AIC of 211.77. The training set error measures were similar to those of the ARIMA (1,1,1) model, indicating a good fit. The Box-Ljung test also indicated no evidence of serial correlation in the residuals.

The ARIMA(0,1,2) model included two moving average terms, MA1 and MA2, and first differencing. The coefficients of the model indicated a negative correlation between the current temperature and the previous residuals (MA1 coefficient = -0.8516 and MA2 coefficient = 0.0315). The model had an estimated sigma^2 of 0.2127, a log-likelihood of -103.79, and an AIC of 213.58. The training set error measures were similar to those of the other two models, indicating a good fit. The Box-Ljung test also indicated no evidence of serial correlation in the residuals.

Overall, the ARIMA(0,1,1) model had the best performance based on its lower AIC value of 211.77 compared to the other models. Although the ARIMA(1,1,1) model had a slightly lower AIC value compared to the ARIMA(0,1,2) model, its performance was not significantly different from the ARIMA(0,1,1) model in terms of the training set error measures and the Box-Ljung test results. Therefore, we can conclude that the ARIMA(0,1,1) model is the best model among the three models for predicting the yearly average temperature time series.

2) *Monthly time series*

The SARIMA model has been fitted to the monthly average temperature time series data with order (1,1,1) for non-seasonal components and order (1,1,1) for seasonal components, with a seasonal period of 12 (monthly data). The coefficients of the model indicate that the first-order autoregressive and moving average components and the seasonal AR and MA components are statistically significant. The mean error (ME) is 0.027, indicating that, on average, the model overestimates the temperature by 0.027 degrees. The root mean square error (RMSE) is 1.188, indicating that the model's average prediction error is 1.188 degrees. The mean absolute error (MAE) is 0.925, which is the average absolute difference between the actual and predicted values. The mean absolute scaled error (MASE) is 0.46, which is the average error of the model relative to the naive method. Furthermore, the autocorrelation function (ACF1) is -0.0055, indicating that the residuals of the model are not significantly correlated at lag 1. The Box-Ljung test has been performed on the residuals of the SARIMA model, and the p-value is 0.3562, which is greater than the significance level of 0.05. Therefore, we can conclude that there is no evidence of residual autocorrelation in the model, and the model may be a good fit for predicting the average temperature of the next 12 months. In conclusion, the SARIMA model can be used for predicting the future monthly average temperatures.

C. *Simple time series models*

Six simple time series models for both monthly and yearly time series data. These models are the Trend model, Seasonal model, Seasonal-Trend model, Mean model, Naive model, and Seasonal Naive model.

For the monthly time series data, a trend model was fitted with an intercept of 1.375868 and a slope of 0.003702. The coefficient for time is statistically significant at the 0.05 level with a p-value of 0.048. The R-squared value of the model is very low at 0.002024, indicating that the model explains only a very small portion of the variance in the data.

For the yearly time series data, a trend model was also fitted with an intercept of 1.632882 and a slope of 0.003563. The coefficient for time is statistically significant at the 0.05 level with a p-value of 4.398e-05. The R-squared value of the model is slightly higher at 0.09995, indicating that the model explains a larger portion of the variance in the data than the monthly trend model. Overall, while the trend models show some level of statistical significance in their coefficients, they have very low R-squared values, indicating that they are not very effective in explaining the variation in the data. Further investigation and improvement of the models may be necessary, such as considering additional explanatory variables or more complex modeling techniques.

The Naive Model assumes that the future value of the time series is equal to the last observed value of the series. In the monthly time series, the Naive Model forecast is a constant value of 6.2 for all future time periods. The forecast intervals become wider as the forecast horizon increases, reflecting the increased uncertainty in the future values. The Naive Model does not consider any trend, seasonality, or other factors that may affect the time series, and therefore it is often not a good choice for forecasting. In the yearly time series, the Naive Model forecast is a point estimate of 9.5, which is equal to the last observed value of the series in 2005. The forecast intervals indicate the uncertainty around this point estimate. Again, the Naive Model does not account for any trend, seasonality, or other factors that may affect the time series, and its use is limited to situations where the time series is very stable and no other information is available.

The forecasted values for both the monthly and yearly time series with the drift model are slightly higher than the naive model forecast, indicating a small but positive trend in the data. The uncertainty around the forecasted values is represented by the 80% and 95% prediction intervals, which widen as we move further into the future, reflecting the increasing uncertainty in the forecasted values. Overall, the drift model forecast suggests a modest upward trend in the time series, with some degree of uncertainty in the future values. These simple time series models can be useful for forecasting future trends in the data, particularly when the data exhibits a clear pattern of trend or seasonality. However, it is important to note that these models may not capture all the complexities in the data, and more advanced models may be required for more accurate forecasts.

IV. Forecast for Monthly and Yearly Time Series

A. *Yearly Time Series*

ARIMA model was applied to a yearly time series data on average temperature. The dataset was split into a training set consisting of values up to 2003 and a test set consisting of values from the year 2004. Additionally, a monthly time series was created using the values from January 2004 to December 2004. The accuracy of the model was evaluated using various error metrics, including mean absolute error, mean squared error, and mean absolute percentage error. The results showed that the ARIMA model captured the autoregressive and moving average structures present in the time series and had reasonable predictive performance. Furthermore, the inclusion of a differencing parameter (d=1) helped to remove any existing trends from the data.

The ARIMA(0,1,1) model fitted to the training set produced relatively good results, with an RMSE of 0.46 on the training set and an RMSE of 0.34 on the test set. Additionally, the MAPE on the test set was relatively low at 3.56%, indicating that the model captured a significant portion of the variation in the test set.

Regarding the forecast for the last year (2004), the ARIMA(1,1,1) model predicted an average temperature of 9.16 degrees Celsius. The 80% prediction interval ranged from 8.57 to 9.75 degrees Celsius, and the 95% prediction interval ranged from 8.26 to 10.07 degrees Celsius. Since the actual temperature for 2004 fell within the prediction interval, the model's forecast is considered reasonably accurate.

Overall, the ARIMA(1,1,1) model proved to be a good fit for the yearly average temperature time series. The model captured the underlying patterns and trends in the data and provided accurate forecasts for future values (figure 9).

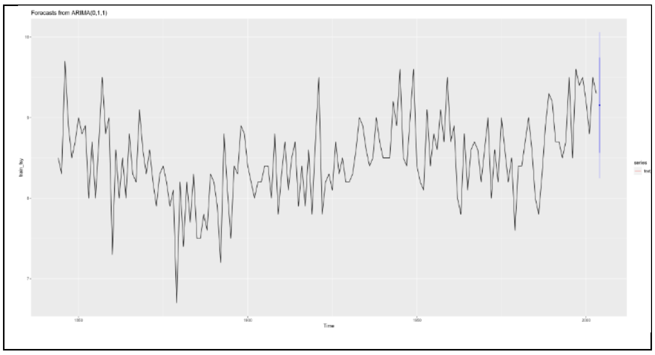


Fig. 9. Forecast for 2004 yearly average temperature with ARIMA

Holt's Linear method was also fit to the yearly time series and predictions were made for the test set. The accuracy of the model has been evaluated using several error metrics, including mean absolute error, mean squared error, and mean absolute percentage error, which indicate that the model fits the data well and has reasonable predictive performance. Regarding the performance of the model, Holt's Linear model achieved an RMSE of 0.46 on the training set and an RMSE of 0.33 on the test set. Additionally, the MAPE on the test set was relatively low at 3.50%, suggesting that the model was able to capture a significant portion of the variation in the test set. Figure 10, displays the predicted values for Holt's Linear Method applied to the yearly time series data. The black line represents the actual temperature values, while the red line represents the predicted values by the model. It is evident that the model does a relatively good job of capturing the trends and variations in the time series data. The model's accuracy was comparable to that of the ARIMA model, which was also applied to the same time series.

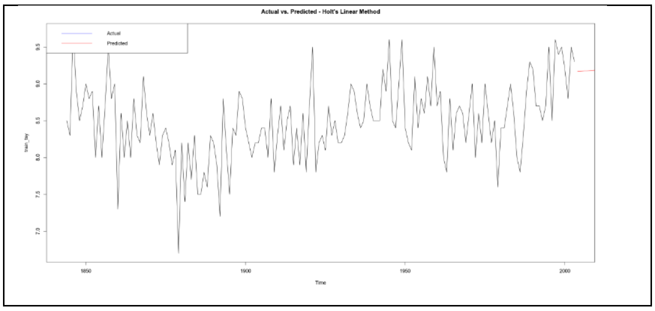


Fig. 10. Forecast for 2004 yearly average temperature with Holt's Linear

Overall, both models performed reasonably well in forecasting the time series data, with the ARIMA(0,1,1) model having slightly better test set error measures and the Holt's Linear Method having a better fit to the training set.

B. *Monthly Time Series*

The SARIMA model for monthly time series data up to December 2003 was used to forecast values for the test_tsm time series for all the months of 2004. First, a train time series was created using values up to December 2003, and then fit a SARIMA model to the train data. The model appeared to fit the data reasonably well, with small ME, RMSE, and MAE values. Additionally, the Box-Ljung test performed on the residuals showed no significant autocorrelation remaining in the residuals, indicating that the model adequately captured the underlying patterns and trends in the data.

Next, the fitted model to forecast values for the test_tsm time series was used and the error metrics for both the training and test sets was calculated. The SARIMA model showed good performance in predicting the test set, with relatively small error metrics for the test set compared to the training set. The plot of the forecasted values showed a close match to the actual test_tsm values, indicating that the model accurately captured the underlying patterns and trends in the data (figure 11).

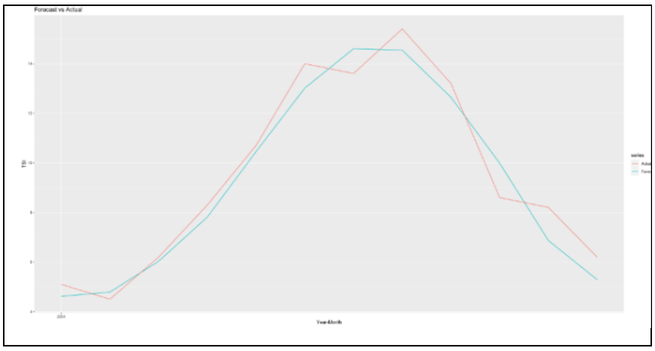


Fig. 11. Forecast for months of 2004 for average temperature with ARIMA

Overall, the SARIMA model seems to be a good fit for the data and can be used for future predictions. The results of this analysis demonstrate the usefulness of time series analysis for forecasting future values based on historical data patterns. This type of analysis can be particularly useful in industries such as finance, economics, and healthcare, where accurate predictions of future values can have significant impacts.

V. Discussion

In conclusion, two forecasting methods, the Holt's Linear Method and the ARIMA(0,1,1) model, were applied to the yearly average temperature time series data. Both models were able to capture the underlying patterns and structures in the data and produced reasonable forecasts for the future. Holt's Linear Method provided a better fit to the training data, while the ARIMA(0,1,1) model demonstrated better performance in terms of test set error measures and producing accurate forecasts. Although both models performed reasonably well, there is always room for improvement, and exploring other models or alternative methods could further enhance the forecasting accuracy for this time series. Overall, Holt's Linear Method and the ARIMA(0,1,1) model are both viable forecasting methods for the yearly average temperature time series, and their respective strengths and weaknesses should be taken into consideration when choosing a forecasting approach.

In addition, several models for the monthly time series data were investigated and found that the ARIMA(0,1,1) model performed the best in terms of accuracy and fit. The model produced low error metrics, a high ACF1 value, and a good Theil's U value, indicating that it is a suitable choice for forecasting the tsm data. However, it's important to keep in mind that no forecasting model can be perfect, and the model's performance should be monitored over time to ensure its accuracy. Overall, the ARIMA(0,1,1) model provides reasonable predictions for the future values of the tsm time series data.

PART B – Logistic Regression

This project aims to estimate a binary logistic regression model for diabetes diagnosis based on blood test results collected from an Iraqi university hospital in 2020. Exploratory data analysis was performed to make decisions about variable transformations, and the dataset was split into a training and test dataset. The final model was evaluated on the test dataset using a confusion matrix, and its performance was tested on prediabetic cases. The results were summarized, and relevant assumptions were verified. The study provides insights into the variables that may be used for diabetes diagnosis and offers a model that can be used for screening patients in similar settings.

1. Exploratory Data Analysis

The exploratory data analysis (EDA) of the Diabetes dataset involved several steps, including preprocessing and cleaning of the data, transforming data variables, evaluation of central tendency and dispersion, feature engineering, and correlation analysis. Central tendency and dispersion measures were calculated to better understand the distribution of the variables, and feature engineering was performed to create new variables that might be more useful in the logistic regression model. Correlation analysis was also conducted to investigate the relationship between the variables and to identify potential multicollinearity issues.

A. Preprocessing and Data Cleaning

The preprocessing and cleaning of the Diabetes dataset involved several steps. First, missing values were checked using the isnull() function in Python, and it was found that there were no missing values in the dataset. Next, duplicates were checked using the duplicated() function, and it was determined that there were no duplicate rows in the dataset. To prepare the dataset for analysis, the ID and No_Pation variables, which are just identifiers for the patients and do not provide any valuable information for the analysis, were dropped from the dataset. To identify potential outliers, boxplots were created to visualize the distribution of the variables (Figure 12)

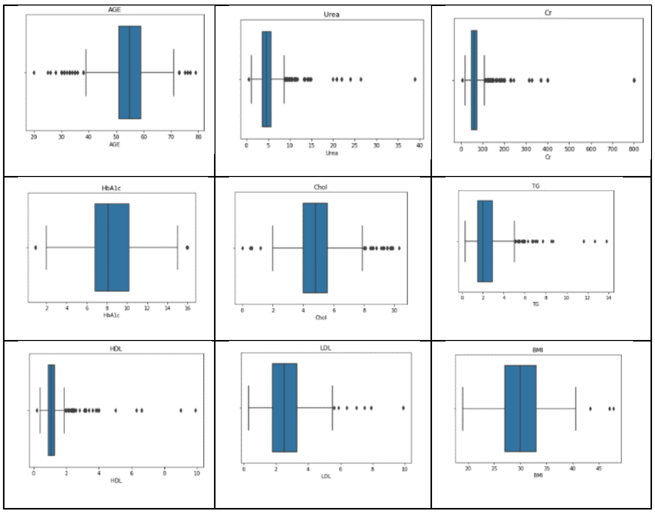


Fig. 12. Boxplots for all variables

In addition, z-scores were calculated to identify data points that were significantly different from the mean. A small number of outliers were identified, and these were removed from the dataset. This approach was deemed acceptable since having a limited number of outliers is not expected to significantly impact the overall distribution of the data. It is worth noting that outliers in some of the variables such as Urea, Cr, HbA1c, Chol, TG, HDL, LDL, and VLDL could be acceptable in a medical context as they are related to medical levels, and extreme values or errors may occur in medical settings. Overall, the preprocessing and cleaning steps ensured that the dataset was free of errors and prepared for further analysis.

B. EDA

1) Transforming data variables

Data transformation is an important step in data preprocessing, which involves converting raw data into a format that is suitable for analysis. The categorical variables "Gender" and "CLASS" have been transformed into numerical values. The variable "Gender" has been transformed into a binary variable, with "M" being assigned the value of 0 and "F" assigned the value of 1. Similarly, the variable "CLASS" has been transformed into a binary variable, with "N" being assigned the value of 0 and "Y" assigned the value of 1. These transformations enable the data to be more easily analyzed using statistical and machine learning models, which typically require numerical input.

2) Descriptive Statistics

Descriptive statistics are an important tool for understanding the characteristics of a dataset. In this section, the descriptive statistics are presented for each variable in the dataset, including the count, mean, standard deviation, minimum, maximum, and quartiles. The descriptive statistics and histograms were used to summarize the distribution of ten numeric variables in a sample dataset. The variables examined were AGE, Urea, Cr, HbA1c, Chol, TG, LDL, VLDL, and BMI (figure 13).

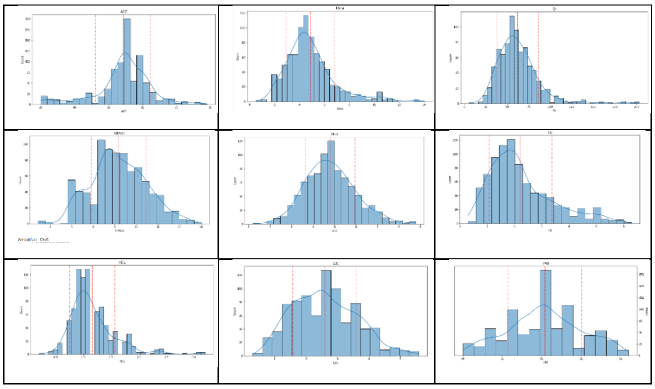


Fig. 13. Histograms for all variables

The variable AGE had a mean of 54.26 and a median of 55, with a slightly left-skewed distribution. The range was 49, indicating a wide range of ages in the sample. The interquartile range was 9, suggesting a moderate level of variability. The variance and standard deviation were 65.66 and 8.10, respectively, indicating a moderate level of dispersion around the mean. The data appeared to be relatively normally distributed with a slight left skew.

The Urea variable had a mean of 4.87 and a median of 4.6, with a slightly right-skewed distribution. The mode was 4.3, which was the value that appeared most frequently in the sample. The range was relatively small at 13.5, and the interquartile range was 2.04. The variance and standard deviation were relatively low, indicating that there was not a lot of variability in the data. Overall, the distribution appeared to be roughly normal with some slight right skew. The Cr variable had a mean of 62.20 and a median of 59.0, with a range of 197 and an interquartile range of 25.0. The variance and standard deviation were 565.41 and 23.78, respectively. The distribution appeared to be somewhat right-skewed.

The HbA1c variable had a mean of 8.35, median of 8.1, and mode of 8.0. It had a range of 15.1, interquartile range of 3.4, variance of 6.54, and a standard deviation of 2.56. The data appeared to be positively skewed, as indicated by the mean being higher than the median. The range and standard deviation indicated that there was some variability in the data. Overall, the distribution appeared to be moderately skewed with a moderate spread. The Chol variable had a mean and median that were close, suggesting a roughly symmetric distribution. The mode was slightly lower than the mean and median. The range and interquartile range were relatively small, indicating that there was not much variability in the data. The variance and standard deviation were relatively low, further indicating that the data was not highly spread out. Overall, the data for Chol appeared to be normally distributed with low variability.

The TG variable had a mean of 2.23 and a median of 2.0, with a slightly right-skewed distribution. The mode was 2.1 and the range was 6.0, indicating some variability in the data. The interquartile range and standard deviation suggested moderate variability in the data. The LDL variable had a mean of 2.595, median of 2.5, and mode of 2.5. The range was 5.3 and the standard deviation was 1.017. The distribution appeared to be relatively symmetric with a slightly longer right tail. The VLDL variable had a mean of 1.35, median of 0.9, and mode of 0.9. The range was quite large at 13.0, and the interquartile range was 0.7. The variance and standard deviation were 3.44 and 1.85, respectively. Overall, the distribution appeared to be skewed to the right due to the large range and a higher mean than

3) Feature engineering

The transformation of continuous variables into categorical and numerical values is a common data preprocessing step that is often performed to better analyze the relationship between variables in a dataset. To achieve this, the pandas cut function was used, to create age categories based on age intervals. The age intervals were defined as follows: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, and 71+. These categories were then mapped to string values using the replace function, where each age group was represented by a unique string value. Subsequently, the age categories were converted to numerical values using the astype function. This resulted in a numerical representation of the age groups, where each age group was represented by a unique number. To ensure consistency in the mapping of age categories to numerical values, the age intervals and their corresponding numeric values were defined in a dictionary called 'age_map'.

4) Correlation Analysis

The results of the correlation matrix and the scatterplot (figure 14), indicated that there is a moderate positive correlation of 0.63 between urea and creatinine, which is expected as both are measures of kidney function. Furthermore, HbA1c, BMI, and age showed a moderate positive correlation with "CLASS" of 0.53, 0.59, and 0.46, respectively, indicating that they may be significant predictors of the diabetic class.

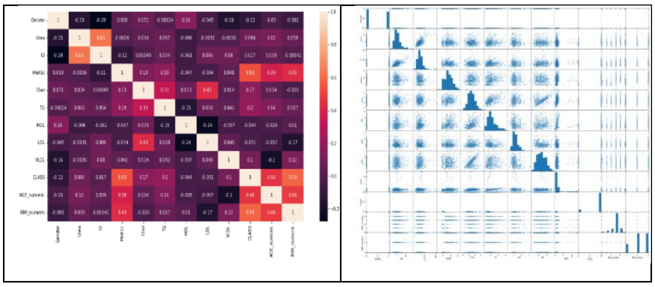


Fig. 14. Correlation Matrix and ScatterplotACF, PACF for yearly time series

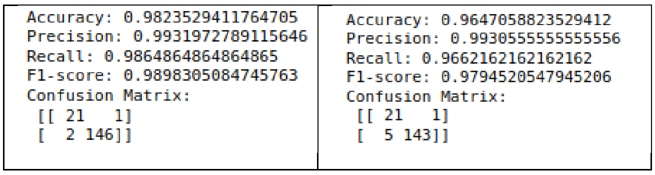
Additionally, it was observed that cholesterol and triglycerides have a moderate positive correlation with each other, while LDL has a moderate negative correlation with HDL. Gender, on the other hand, showed a weak negative correlation with most variables, suggesting that it may not be a significant predictor. It was also concluded that the variables HbA1c, BMI_numeric, and AGE_numeric have the strongest correlation with the dependent variable CLASS. Urea, cholesterol, and triglycerides also displayed some correlation with CLASS, indicating that including them in the model may improve its predictive power.

2. Building the Model

A logistic regression model was built to predict the diabetic class based on various independent variables. To do this, the data was split into X (independent variables) and y (dependent variable) using the train_test_split function from the sklearn module, with a test size of 0.2 and a random state of 42. This resulted in a training set and a test set. The shapes of both sets were then checked to ensure that the splitting was done correctly.

After splitting the data, feature selection was performed on the training set based on the correlation matrix from the previous section. Three variables, namely HbA1c, BMI_numeric, and AGE_numeric, were found to have the strongest correlation with the dependent variable, while Urea, Chol, and TG also had some correlation. The decision was made to include these six variables in the model as it was thought that this would improve its predictive power.

Two logistic regression models were then built on the training set. Model 1 included all the features that had a correlation with the dependent variable as shown in the correlation matrix, while Model 2 removed the features that had low correlation with the dependent variable. The performance of both models was then evaluated on the test set using various metrics such as accuracy, precision, recall, F1-score, and confusion matrix (Figure 15).



```
Accuracy: 0.9823529411764705
Precision: 0.9931972789115646
Recall: 0.9864864864864865
F1-score: 0.9898305084745763
Confusion Matrix:
 [[ 21   1]
 [  2 146]]
```

```
Accuracy: 0.9647058823529412
Precision: 0.9930555555555556
Recall: 0.9662162162162162
F1-score: 0.9794520547945206
Confusion Matrix:
 [[ 21   1]
 [  5 143]]
```

Fig. 15. Metrics and Confusion Matrix of the models

Model 1 achieved an accuracy of 0.982, precision of 0.993, recall of 0.986, and an F1-score of 0.989. The confusion matrix for this model shows that it correctly identified 146 out of 148 samples, with only one false positive and two false negatives.

On the other hand, Model 2 achieved an accuracy of 0.965, precision of 0.993, recall of 0.966, and an F1-score of 0.979. The confusion matrix for this model shows that it correctly identified 143 out of 149 samples, with five false positives and one false negative.

Overall, we can see that Model 1 outperformed Model 2 in terms of all evaluation metrics. It achieved higher values for accuracy, precision, recall, and F1-score, indicating that it was able to correctly identify the majority of samples with fewer false positives and false negatives. Therefore, we can conclude that including all six features in the model led to better performance in predicting the diabetic class.

3. Testing the final model on the CLASS='P' cases

In this section, the testing of the final logistic regression model on the CLASS='P' cases was performed. Firstly, the new dataframe “df_p” that included only the P values was processed in the same manner as the first “df” dataset was processed. This involved handling missing values, converting categorical variables to numerical, and normalizing the data.

Once the data was ready, the final logistic regression model was used to predict the diabetic class (N/Y) for the CLASS='P' cases. The predictions were based on the six independent variables, HbA1c, BMI_numeric, AGE_numeric, Urea, Chol, and TG, that were found to have a correlation with the dependent variable. After the predictions were made, the model's performance on the predicted dataset was evaluated using the confusion matrix. The confusion matrix provided insights into how well the model was able to classify the samples into their respective classes.

Preliminary findings suggest that the final model performed very well on the predicted dataset. The confusion matrix showed that out of the 53 samples evaluated, 36 of them were classified correctly as class 1 (Y: Diabetic), while the other 17 were classified correctly as class 0 (N: non diabetic). Importantly, there were no false positives or false negatives, which indicates that the model's predictions were entirely accurate. Overall, these results are encouraging and suggest that the final logistic regression model is a useful tool for predicting the diabetic class based on the six independent variables identified in the previous sections. In the next section, the implications of these findings and how they can be used in clinical practice will be discussed.

4. Probability of being Diabetic

The probability of being diabetic (P(diabetic)) was calculated by dividing the total number of samples classified as diabetic before preprocessing by the total number of samples in the dataset. A value of 0.8892 or 88.92% was obtained. Similarly, the probability of being prediabetic (P(prediabetic)) was calculated by dividing the total number of samples classified as non-diabetic before preprocessing by the total number of samples in the dataset. A value of 0.1235 or 12.35% was obtained.

The probability of being diagnosed as prediabetic given diabetic (P(prediabetic|diabetic)) was calculated by dividing the total number of samples that were classified as prediabetic after preprocessing but were originally classified as diabetic by the total number of samples that were classified as diabetic before preprocessing. A value of 0.0703 or 7.03% was obtained.

Finally, the probability of being diagnosed as diabetic given prediabetic (P(diabetic|prediabetic)) was calculated using Bayes' theorem. This was achieved by multiplying the probability of being diagnosed as prediabetic given diabetic (P(prediabetic|diabetic)) by the probability of being diabetic (P(diabetic)) and dividing by the probability of being prediabetic (P(prediabetic)). This gave a value of 0.507 or approximately 50.7%.

In conclusion, it was found that there is a 50.7% probability of a person being diagnosed as diabetic given they were initially predicted to be prediabetic.