Data Mining & Machine Learning 1 (MSCDAD_C) Portfolio Project Report

Debayan Biswas
Student ID: 22242821
Master of Science in Data Analytics
National College of Ireland
x22242821@student.ncirl.ie

***Abstract*—This comprehensive project systematically employs machine learning techniques following the structured KDD (Knowledge Discovery in Databases) methodology. It explores three diverse datasets, each presenting unique challenges and scenarios. Dataset 1 focuses on predicting income levels from census data, requiring intricate preprocessing for diverse numeric, binary, and categorical data types. The application of robust evaluation metrics is crucial for assessing algorithm performance accurately. Dataset 2 shifts to hourly averaged responses and gas concentrations, involving complex preprocessing, regression analysis, and time series evaluation with integer, categorical, and date-type data. Dataset 3 aims to estimate room occupancy using environmental sensors, addressing diverse data types. Through classification and regression models, the project gains insights into each dataset's nuances, contributing valuable findings to the effectiveness of machine learning methods. Aligned with KDD's systematic approach, this project advances the understanding of various machine learning techniques and their limitations, promoting informed and effective applications to real-world challenges.**

I. Introduction

Motivated by the escalating prevalence of machine learning techniques and their capacity to tackle real-world challenges, the report endeavors to provide a thorough analysis of the unique characteristics and complexities of the diverse datasets.

A. Dataset 1

This dataset poses the task of predicting income levels from census data using classification, as the output variable is binary. Categorical columns with qualitative data require encoding. The study aims to compare the performance of two classification algorithms for insights by following the systematic approach of KDD. The primary objective is to build a model that can derive patterns from labeled training data and make precise predictions for new data.

B. Dataset 2

Dataset 2 shifts focus to hourly averaged responses and gas concentrations, requiring sophisticated data preprocessing, regression, and time series analysis. Proper time-related visualization and time-series analysis will be implemented in the analysis to predict future outcomes of air quality.

C. Dataset 3

Dataset 3 challenges the estimation of room occupancy with environmental sensors. The goal is to offer valuable insights into the effectiveness of machine learning methods across diverse datasets and domains. Following KDD principles, this introduction sets the stage for a comprehensive investigation to navigate the complexities of different data types, challenges, and objectives in each dataset.
The report details the project's related work, systematically evaluates findings, and describes data mining methodologies using the KDD approach, covering preprocessing, visualization, and transformation for each dataset. The Evaluation section outlines performance measures and results for all three datasets. The document ends with a recap of results, an examination of constraints, and prospects for future exploration. The reference section acknowledges all implemented findings.

II. Related Work

A. Dataset 1

The dataset for the Census data or the Adult dataset examines and predicts the income expectancy based on several parameters. "Predicting Population Income Class through Data Visualization" by G. D. Singh, Himanshi V., and A. Kumar utilizes machine learning and data visualization to assess the impact of population growth on natural resources, and analyze trends in poverty, unemployment, and state growth in a specific region of India. The study employs the Gradient Boosting Classifier Model with high accuracy but lacks precise parameter identification for optimal predictor selection, suggesting a need for refinement and precision to enhance predictive accuracy. "Income Prediction with Support Vector Machine" by A. Lazar highlights supervised learning's accuracy in predicting income on a large dataset. However, the use of data reduction techniques like random subset selection and Principal Component Analysis (PCA) might compromise model accuracy, emphasizing the importance of addressing classification errors. "A Statistical Approach to Adult Census Income Level Prediction" by N. Chakrabarty and S. Biswas explores the growing reliance on data and technological advancements in Data Mining and Machine Learning. The study conducts a comprehensive analysis to identify key factors for improving individual income and addressing societal wealth

disparities. The authors propose exploring hybrid models and advanced preprocessing techniques to enhance results without sacrificing accuracy. "Nowcasting New Zealand GDP using Machine Learning Algorithms" by A. Richardson and T. Mulder enhances GDP forecast accuracy with ML algorithms on a large real-time dataset. However, the machine learning approach lacks interpretability, and efficient outlier handling, and may have high computational complexity, suggesting areas for improvement.

B. Dataset 2

Research on the Air Quality measured by Ilaria Pigliautile et al.'s work on "Investigation of CO2 Variation and Mapping Through Wearable Sensing Techniques for Measuring Pedestrians' Exposure in Urban Areas" investigated Air Quality using wearable sensors and implemented complex algorithms for the estimation of sensor data. The examination indicates constrained correlations between CO2 levels and other environmental factors, suggesting difficulties in establishing clear cause-and-effect relationships. The work by Suhasini V. Kottur and Dr. S. S. Mantha focuses on forecasting air pollutants in Mumbai, using an integrated model with Artificial Neural Networks and Kriging. The model may require recalibration based on geolocation and data types to adapt to evolving pollutant sources and meteorological conditions. Huixiang Liu et al.'s research in Beijing develops a forecasting model using random forest regression (RFR) and support vector regression (SVR). While efficient, considering an ensemble of machine learning algorithms for different demographics is suggested for a comprehensive prediction justification. De Vito's study explores the efficiency of low-cost gas multisensor devices in densifying sparse urban pollution, discussing the feasibility of a multi-gas sensor fusion algorithm for calibration. However, continuous model recalibration and sensor placement biases could pose limiting factors.

C. Dataset 3

The paper "Machine Learning-Based Occupancy Estimation Using Multivariate Sensor Nodes" by A. Singh, Vivek Jain, focuses on accurately determining room occupancy using various sensor nodes and low-cost sensors. While supervised learning with different feature sets is employed, the challenge lies in achieving real-time processing, urging the exploration of more advanced models for complex data. In "An information technology-enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network" by Dong et al., a thorough wireless and wired sensor network is established for occupancy estimation. To enhance its applicability across diverse scenarios, environments, and seasons, the model should be tested more comprehensively. Dr. Timothy Osirike's "Detecting Room Occupancy Using Machine Learning and Sensor Data" utilizes Pycarets and classification models on IoT devices. However, for real-world scenarios with intricate data, the algorithms may need to be substituted with more sophisticated models for efficient occupancy detection. The journal "Accurate occupancy detection of an office room from light, temperature, humidity, and CO2 measurements using statistical learning models" by Luis M. Candanedo and Véronique Feldheim employs statistical models for office room occupancy prediction, achieving high accuracy. However, the absence of advanced machine learning models, like deep learning, may limit its effectiveness in handling more complex data and extreme scenarios.

III. Data Mining Methodology

The data mining methodology comprises data preparation, data visualization, and transformation of data for evaluating the end result of the data. The three datasets are analyzed in this section.

A. DATASET 1 on Adult or Census data

1) Exploratory data analysis and Data Pre-Processing

a) Data summary

The adult dataset consists of 32561 rows and 15 features with a mix of character and numeric type data. In this dataset, 9 columns consist of character-type data most of which are categorical types of data. The categorical data need to be encoded for further data analysis as these are significant for the analysis. The summary for the dataset is checked to find the count, maximum, minimum, mean, and quartiles of all the numeric features present. Length, class, and mode are checked for the categorical type data. The dataset also has missing header values which are added as per the source data.

b) Missing value

The presence of missing or blank values is checked in the dataset as these values have a substantial influence on the end result and analysis of data and need to be removed from the dataset. The blank values are present in the column workclass, occupation, and native_country and are replaced with NA. Replacing with NA allows for easier handling and imputation of missing or blank data during further analysis or modeling. The column native_country which is important for assessing the location of the individual for census data, has been operated to remove all the blank values thus the total count of the dataset falls to 31978. The NA values in workclass and occupation are replaced by mode.

c) Converting Categorical Columns

The categorical columns are type-converted for further operation. They are converted to integer-type data using the integer-factor operator. The character categorical columns are now classified into categories based on numeric data.

2) Visualization

a) Correlation Analysis

The correlation among the numeric columns is checked and a correlation heatmap is plotted which can be seen in Fig.1. The plot clearly depicts that there is no correlation present between the numeric columns of the dataset.

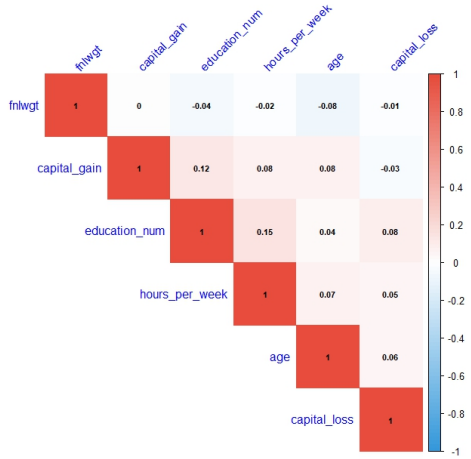

Fig. 1: Correlation Heatmap

b) Barplot

A barplot is plotted for the dependent variable Income and the independent variable Education to check the impact of education on the income status of individuals. The bar plot in Fig.2 clearly shows the distribution as per the respective education of the individual.

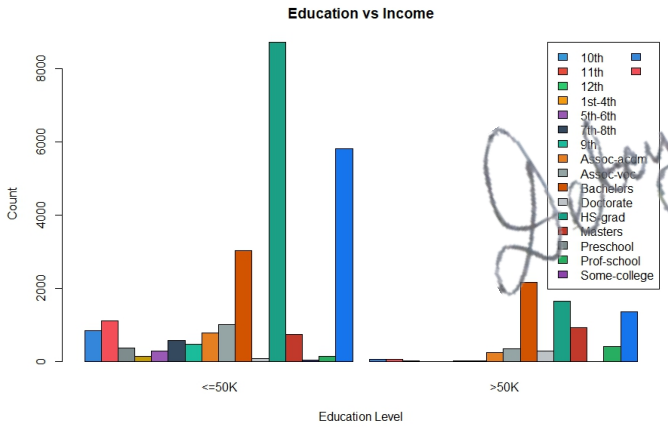

Fig. 2: Barplot Income vs Eductaion

3) Data Preparation and Further Visualization

a) Outliers and Boxplot

Boxplot analysis is performed on the numeric columns to check for any presence of outliers in the data. The Boxplot in Fig.3 confirms the presence of Outliers mostly in the 'fnlwgt' column as it has a wide range of scattered values. This high value can be reduced by normalizing the dataset after outlier removal. The column capital gain and capital loss are significant features and they hold outliers as well which needs to be removed. The Outlier removal using the Z-score gives a better result as compared to the interquartile range method. This outlier needs to be removed before further analysis of data. The dataset is then normalized after outlier removal.

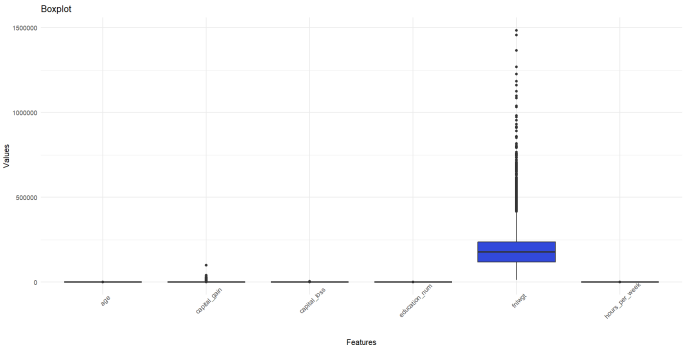

Fig. 3: Boxplot before Outlier removal for Numeric columns

The boxplot in Fig.4 indicates a partial removal of outliers.

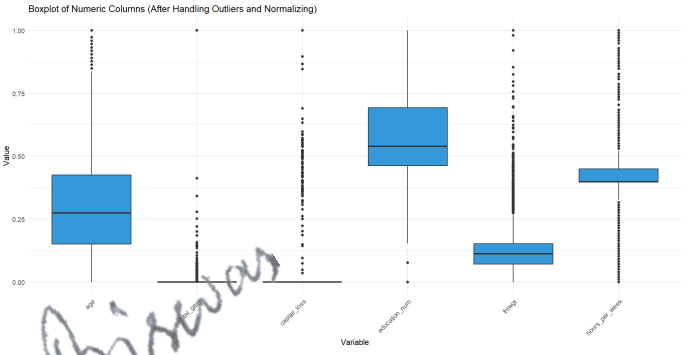

Fig. 4: Boxplot After Outlier Removal and Normalization

Some outliers persist due to the high Y-axis range in Fig.3, where certain outliers were not visible. After normalizing the Y-axis range to a scale of 1, outliers are more apparent. Despite attempting Log transformation to address outliers, it doesn't yield significant improvement, leading to its exclusion from the process.

b) Correlation Analysis for Numeric and all type converted features

The correlation heatmap is again plotted but now for all the numeric and type converted numeric columns. From the heatmap in Fig.5, it is evident that there is no presence of correlation among the dataset which might hinder the proper analytic output.

c) Scatter plot, Boxplot, and Bar plot distribution

A scatter plot is plotted for the two-column capital gain and capital loss depicting that the data concentration is close to both the x and y axis thus giving a clear idea about the scatter. Hence boxplot is plotted as seen in Fig.6 which shows the distribution the both variables.

4) Modeling

a) Splitting data

The dataset is now split into train and test sets using random seed as per student ID of 22242821 with income column considered as split index. The train and test sets are of size 21301 and 9129 respectively.

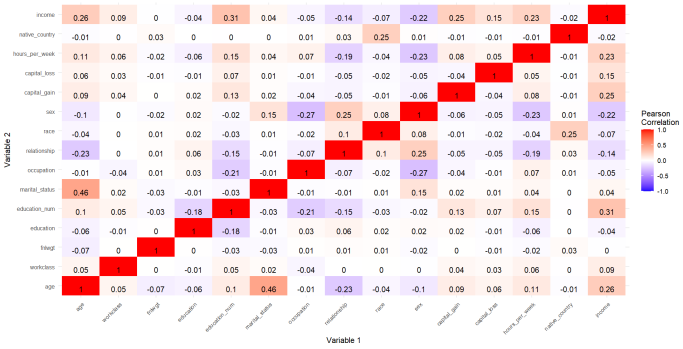

Fig. 5: Correlation Heatmap for all columns

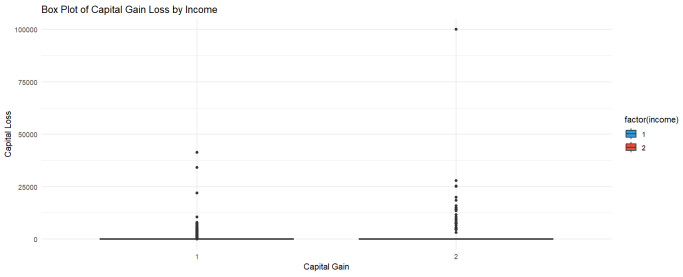

Fig. 6: Box plot for capital gain and loss

b) Random Forest Model

Random Forest, a widely used machine learning technique for classification tasks, excels in enhancing predictive accuracy and managing intricate data relationships by amalgamating multiple decision trees. The Random Forest model is constructed using the 'randomForest' function. The dataset is preprocessed, with the target variable 'income' converted into a factor. The trained model is subsequently employed to predict outcomes on a test dataset, and a confusion matrix is generated to assess the model's efficacy. The resulting accuracy of the Random Forest model is then displayed, offering a numerical gauge of its proficiency in forecasting income classes within the test set.

c) Xgboost Model

Xgboost is applied as a classification model on the cleaned dataset with a random seed of student ID 22242821 and split into train and test sets. The train set is fitted into the Xgboost model to predict the outcome and accuracy of the data. A confusion matrix is generated to assess the model's efficiency. The Evaluation is performed in the Evaluation section depicting the model's accuracy and performance.

B. DATASET 2 on Air Quality

1) Exploratory data analysis

a) Data summary

The data set consists of 9357 rows and 17 features in total having 2 factor type columns which also include a 'date' column, and a 'time' column. The rest of the data columns are of integer type. The time and date column needs to be prepared for time series analysis on the dataset and for generating visualizations.

b) Working on Time feature

For Time-series analysis, the date and time columns are merged forming a single column with a specific date-time format. Hour, Weekday, Month, and Year are extracted from the date-time column which will be an essential step for further analysis.

c) Handling Null or Blank values

The last two columns named 'unknown' do not hold any values and don't impact the analysis. Thus these two columns are dropped. Data present in some of the columns in the dataset has a value of -200 which is considered as bad sensor data. This data is now replaced with NAN (not a number value) which is a correct approach to remove blank value. The rows with blank values will be now removed from the dataset as the presence of NULL or blank values can impact the analysis and performance of the model. The column 'NMHC.GT' on further analysis shows 90 percent NA values present depicting this column does not hold any necessary values that can impact the overall analysis. Thus this column is dropped. The column 'time' is also dropped as it doesn't have any necessary impact on the dataset because it is hourly recorded data for every single day and introducing such data in the analysis will only impact the complexity. The columns are now reordered properly.

2) Visual Exploration and Data Preparation

a) Boxplot

The boxplot is plotted for the numeric columns to check for any outliers present in the dataset. From the boxplot in Fig.7, it can be concluded that there are some outliers present that need to be removed before data modeling. Outliers can greatly impact the data analysis of the data.

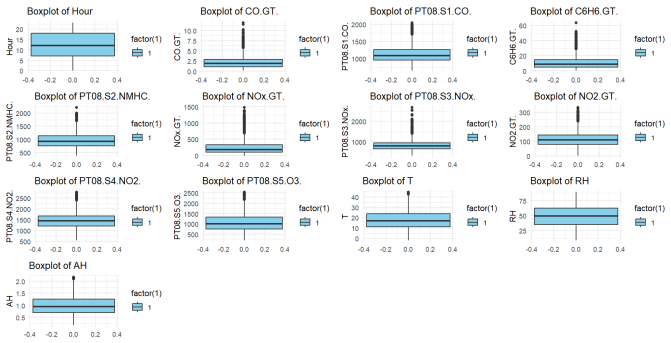

Fig. 7: Box plot

Inter quartile range (IQR) is used as a measure to remove outliers from the dataset. It represents a statistical measure indicating data spread and is computed as the difference between the third quartile (Q3) and the first quartile (Q1). After removing the outliers, the number of rows was reduced to 6099.

A boxplot is further plotted as can be seen in Fig.8 which confirms that most of the outliers have been successfully removed.

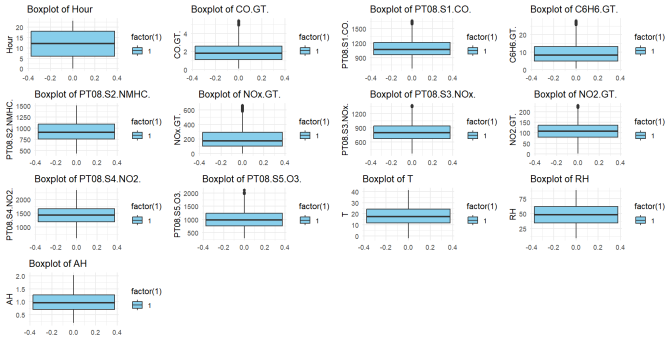

Fig. 8: Box plot after outlier removal

b) Scatter plot

Scatter plots depict the relationship between variables, revealing patterns, correlations, and outliers by providing a quick assessment of the relationship of data. The scatter plot in Fig.9

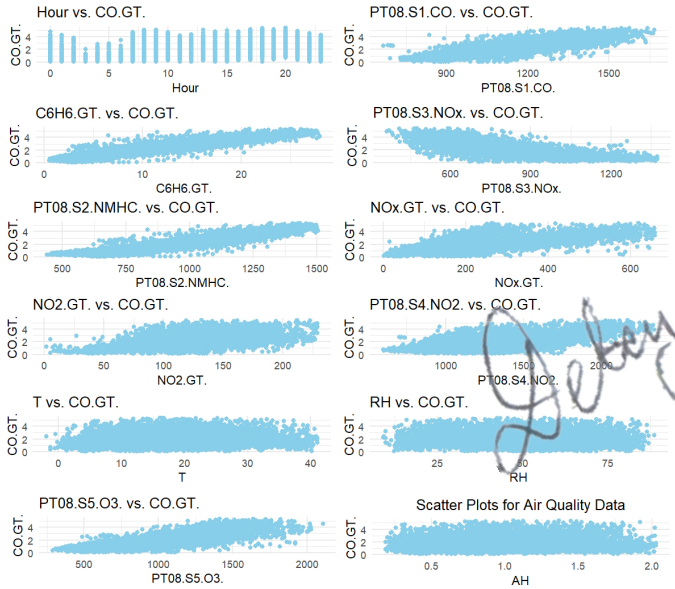

Fig. 9: Scatter plot

shows the relationship between the dependent feature 'CO.GT' concentration with the independent features. From the plot, it can be concluded that most of the features are uniformly distributed except for a few features where the scatter is sparse.

c) Lag plot and Line graph

A lag plot visualization also known as a scatterplot with lag, is a visualization that inspects the autocorrelation of a time-series data. The feature CO.GT is plotted against its previous value at a certain time lag (Date). The plot in Fig.10 shows the Lag plot and gives an insight into the trends, patterns, and correlations in the data across various time intervals.
Similarly a line plot is plotted as can be seen in Fig.11 shows the trends or patterns of dependent feature (CO.GT) over a continuous interval of time.

Fig. 10: Lag plot

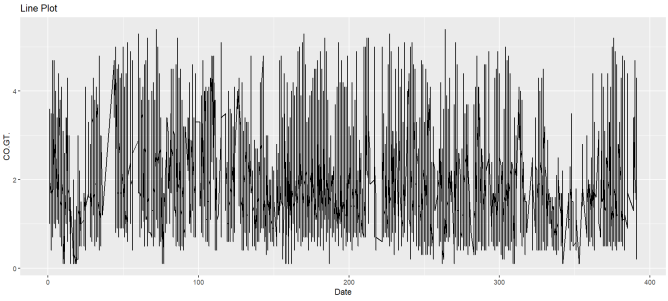

Fig. 11: Line plot

d) Line Plot for hourly averaged sensor response

A line plot is plotted for the hourly averaged sensor response for carbon monoxide which is CO.GT. The line plot is plotted on an Hourly, Weekly, Monthly, and Yearly basis. The plots in Fig.12, show the hourly averaged sensor response count for a daily, weekly, monthly, and yearly basis. The line plot for

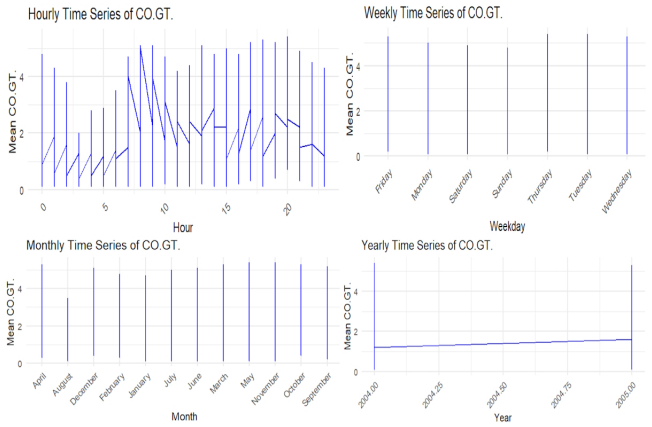

Fig. 12: Line graph for daily, weekly, monthly, and yearly basis

daily response counted per hour shows the variation in the recorded data which can be seen from the crest and trough on the plot. The weekly and monthly plot shows a steady count for the hourly averaged response throughout the time interval.

e) Decomposition of Time-series data

The time-series breakdown illustrated in Fig.13 provides a thorough overview of the dataset. It exhibits the seasonal trend for the specified year range from the start of the value in the dataset till the present day, followed by the representation of trend and seasonal data. Additionally, the figure displays the residuals for the given timeframe.

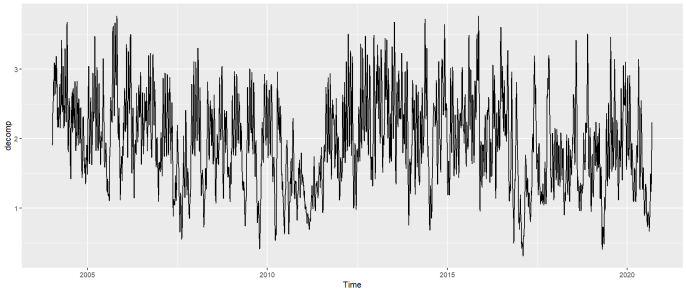

Fig. 13: Decomposition

3) Modeling

a) Data Splitting

The cleaned dataset is now split into train and test sets. The splitting is based on the mentioned train size of 80 percent. Thus the number of data in train and test sets are 4879 and 1220 respectively.

b) Exponential Smoothing

Exponential Smoothing is a method for forecasting time series data that involves assigning changing weights to past observations, with a key focus on progressively decreasing weights over time. It is designed to adapt quickly to changes in data while providing a smooth representation of trends and seasonality. The cleaned dataset is split into train and test sets. The splitting is based on the mentioned train size of 80 percent. Thus the number of data in train and test sets are 4879 and 1220 respectively.The data represents a seasonal pattern which can be observed by checking the Forecast data but as the forecast value has a very low-value difference so when observed over the year in Fig.14, it doesn't provide a proper insight to estimate the forecast. The exponential smoothing is implemented using the Error Trend and Seasonality (ETS) model in the r language. Exponential Smoothing achieves a

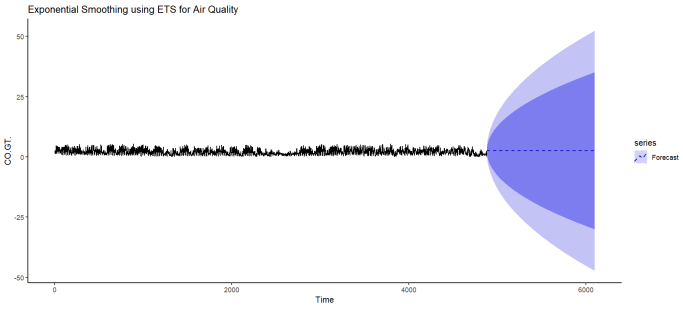

Fig. 14: Exponential smoothing

trade-off between simplicity and precision. However, ETS model applied for exponential smoothing is not providing accurate forecasts and performing poorly, and thus SARIMA will be implemented to check the forecasting. ARIMA is not used as it is a non-seasonal model and will not work efficiently for seasonal data.

c) SARIMA

SARIMA or Seasonal AutoRegressive Integrated Moving Average is a time series forecasting model for predicting future values using historical data or past data records. It exhibits autocorrelation by forecasting future trends through the analysis of past data patterns. The cleaned data is now split into train and test sets, based on the train size of 80 percent. The train and test sets now hold 4879 and 1220 respectively.

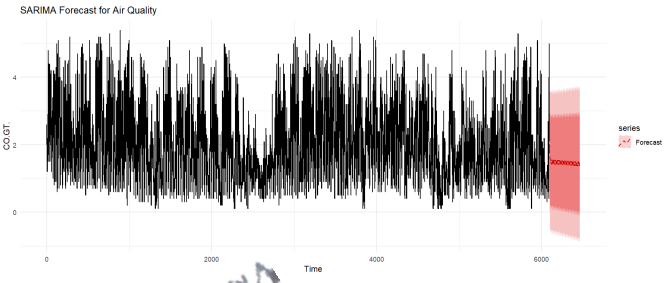

Fig. 15: SARIMA MODEL

SARIMA is implemented using the ARIMA model by introducing seasonality to it as it is for the seasonal model. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) play a crucial role in analyzing data, helping to quantify the correlation between a time series and its lagged values across various time intervals. It helps identify the presence of any systematic patterns or direct relationships between data and its lagged values. The SARIMA plot in Fig.15 predicts the Air quality by forecasting it with the help of the trained data. The SARIMA model gives a proper forecasting value and a lower value of root mean squared error (RMSE), mean absolute error(MAE), and mean absolute percentage error(MAPE), which is discussed in detail in the evaluation section.

C. DATASET 3 on Room Occupancy

1) Exploratory data analysis

a) Data summary

The dataset of room occupancy consists of 10129 rows and 19 columns. The majority of the columns are numeric columns expect for the time and date columns which are present in character format. The dependent variable is the room occupancy column which is influenced by the other independent columns based on the sensor readings. The columns need to be converted into date type before any time-related operation. The column date and time will be further expanded to weekday, month, and hours for implementing further analysis.

b) Time related operations

The date and time column are converted into date type and weekday, month, and hour are extracted from it. After extraction, the column count increases to 22. The time column is dropped from the dataset as the time is in continuously sequential format as the sensors pick the data continuously and including the time data will not only increase complexity but is also unnecessary for the analysis.

c) Null or blank value check

The dataset is checked for the presence of Null or blank values. From the observation, it is evident that the dataset has no Null or blank value present.

d) Re-ordering columns

The columns in the dataset are re-ordered starting with date, month, weekday, datetime, and hour at the beginning for convenience.

2) Data preparation and Visualization

a) Boxplot and outlier

The boxplot is plotted for the numeric features in the dataset to check for the presence of outliers. Outliers can impact the result of the analysis and need to be properly taken care of. The Fig. 16 shows the presence of outliers in the

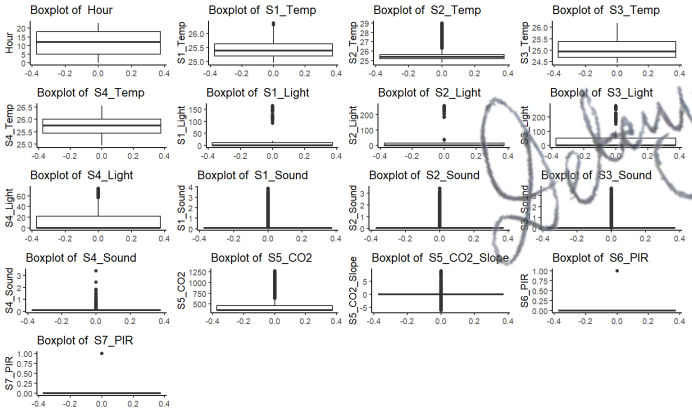

Fig. 16: Box plot before outlier removal

columns. The outlier percentage is also checked which depicts some of the columns consisting of very high value of Outlier percentage. The outliers were operated first using Inter quartile range method of removing outliers but the output generated removes the majority of the data from the room occupancy count column. This is not the correct approach as the data from the dependent feature is important for modeling and analysis. Hence, the Z-score is used as a method to remove the unnecessary outliers from the dataset. The number of rows was reduced to 7549 after the removal of outliers from the dataset. The dependent variable holds numeric category data but upon outlier removal two out of the four categories were removed. This is because the two categories in the dependent variable have very few data for operation. The outlier percentage is also checked which depicts some of the columns still hold outliers. Fig.17 shows that there are still outliers present which should

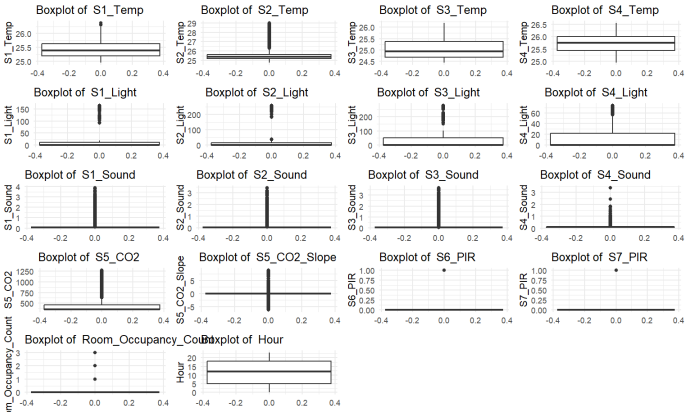

Fig. 17: Box plot after outlier removal

now be removed by the appropriate method of transformation.

b) Distribution of Room occupancy

The room occupancy distribution is checked after the outlier removal. The distribution gives a clear idea that the majority of the data is present for no occupancy count as can be seen in the distribution plot in Fig.18. These data will be worked on during the transformation.

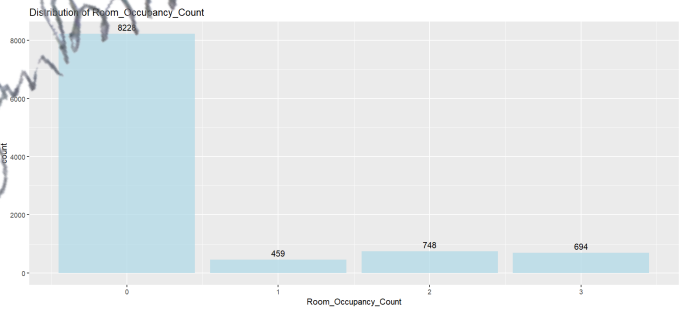

Fig. 18: Distribution of Room Occupancy

c) Correlation Heatmap

Correlation between the numeric columns is checked for the presence of highly correlated columns. The correlation matrix in Fig.19 shows that the column series of temp and column series of light have high collinearity among them individually. This is evident from the fact that the series of four temperature data has almost a similar range of sensor readings to the series of four light data. Thus transformation needs to be applied to them to reduce the collinearity.

Initially log transformation was applied to the dataset but it didn't provide any better result. Hence it is not implemented. Then scaling is implemented to scale the dataset. Scaling is the procedure of adjusting the data to conform to a particular scale range, such as between 0 and 1 or with a mean of 0 and a standard deviation of 1. The result after scaling provides a decent output to carry on with further analysis of the data.

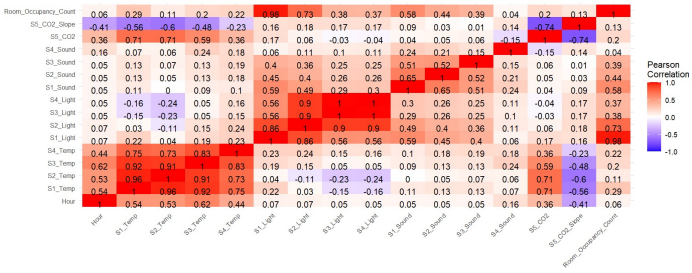

Fig. 19: Correlation Heatmap before Transformation

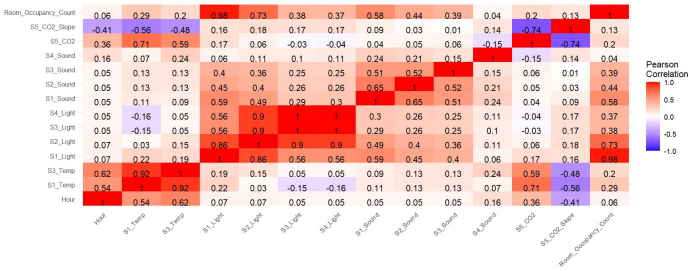

Fig. 20: Correlation Heatmap after Scale transformation

However, checking the Correlation matrix again in Fig.20 shows that some features still hold a correlation among them. These features can not be transformed further as transforming may lead to data loss and discrepancies. This kind of anomaly is due to the fact that the dataset has a similar band of sensor data reading. Thus further operation is ignored.

3) Modeling

a) Data splitting

The transformed data is now split into train and test sets using the seed equivalent to student ID 22242821. The Train and Test sets hold 6040 and 1509 rows respectively.

b) Logistic Regression Model

The logistic regression model is implemented for the analysis of the dataset. The cleaned dataset has the dependent variable which only holds binary values now as the two least important categories with very few data out of the four and got removed while cleaning the data. The train data is fitted into the logistic regression model using the general linear model (GLM) approach. Probability prediction is done on the test set and the probabilities are converted to binary predictions. The evaluation and the analysis are performed in the later section.

IV. Evaluation

A thorough consideration is given to the evaluation methodology for assessing the effectiveness of the utilized machine learning techniques.

A. Dataset 1 Adult Data

The Xgboost and Random Forest Classification Model were used to fit the data and get a proper output for the analysis. The evaluation for the random forest model gives a generous Model accuracy of 86.05 percent. The Precision, Recall, and F1 score stands at 0.63, 0.76, and 0.69 percent respectively which is a fair score for real-life prediction. The Xgboost model evaluation gives a slightly better score than the random forest model with an Accuracy, Precision, Recall and F1 score at 86.33, 0.76, 0.64, and 0.69 percent respectively.

B. Dataset 2 Air quality Data

The findings of the evaluation of the two time-series models in the air quality dataset are analyzed. The exponential smoothing summary for the test set gives a detailed description of the performance of the model. The Exponential Smoothing gives an RSME value of 1.2447 and MAE value of 1.085, and both are fairly low. However, the MAPE which is the mean asboslute percentage error is 125.86 percentage, which is off by a significant percentage from the actual values. This value is very high and may indicate room for improvement in the predictive model. Thus SARIMA is implemented on the dataset.

The test dataset is used to assess the performance of the SARIMA model. Forecasted values are compared with actual observations, and the root mean squared error (RMSE) is calculated to quantify the model's accuracy. The RMSE value of 1.199 suggests a better value compared to exponential smoothing. The MAE stands at 0.866 which is on the lower side and indicates lower mean absolute error. The MAPE value shows a significant improvement over the exponential smoothing model at 52.60 percentage. Thus SARIMA fetches a better value and provides better forecasting and evaluation of the dataset. SARIMA is used here as a machine learning method as data present seasonality which is out of bound for ARIMA evaluation.

C. Dataset 3 Room Occupancy Data

The results of the analysis of the Logistic regression model are evaluated as per the findings for the room occupancy data. The fitted model is tested with the test set and provides a confusion matrix result as can be seen in Fig.21. A confusion matrix serves as a performance evaluation metric, summarizing algorithm performance by presenting counts of true positive, true negative, false positive, and false negative predictions in a structured 2x2 matrix. The high value of True positive and True negative determines that the model performed well in terms of correctly identifying both positive and negative instances. The accuracy, precision, recall value, and F1 score are fairly low at 0.997,1,0.988, and 0.994 percent respectively, thus estimating its efficiency of analysis and performance. This output solely depends on the type of dataset used for operation.

V. Conclusions and Future Work

A. Dataset 1 Adult Data

The Xgboost model demonstrates superior accuracy in precision, recall, and overall F1 score for both classes. Achieving an accuracy of approximately 86.33 reflects a notably

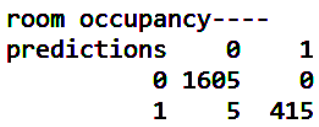

Fig. 21: Confusion matrix

improved performance. The model efficiently assessed the outcomes. The dataset used has limited features that don't fully capture real-life aspects. The evaluation offers insights into income distribution within a population, and this analysis can be expanded to incorporate demographic factors for large datasets. The model is versatile, allowing the examination of occupational, educational trends, and economic disparities. Potential improvements involve creating predictive models for forecasting income trends using historical and time-series data. Additionally, implementing dynamic analysis to comprehend the factors influencing income distribution over time is a future scope.

B. Dataset 2 Air quality Data

The SARIMA machine learning model successfully predicted the dataset with a minimal root mean square error. The model played a significant role in both assessing and predicting the levels of carbon monoxide pollutants in the air. An examination of air quality data over time revealed a seasonal pattern in pollutant concentrations, enhancing our understanding of environmental dynamics. The lower value of RSME and MAE at 1.199 and 0.866 respectively paves the way for better overall forecasting along with a fairly low mean average percentage of error of 52.60 percentage. Improving the capabilities of this model holds the potential for creating advanced predictive models that can forecast upcoming air quality trends. These forecasts would be based on historical data, meteorological conditions, and potential emission scenarios. A potential future avenue is to expand the air quality monitoring network by deploying additional sensors, especially in areas with limited coverage, for a more comprehensive analysis of air quality.

C. Dataset 3 Room Occupancy Data

The Logistic regression model demonstrates excellent accuracy and prediction, nearing 1, which is challenging to achieve in real-world scenarios. The high accuracy is largely attributed to the dataset used for model training, comprising four series of sensor data for temperature, light, sound, and carbon dioxide. To have perfect accuracy and realism in findings, consolidating these individual sensor data series into a single feature is recommended. Refining this model holds the potential to further improve occupancy prediction accuracy for future applications. Additionally, future avenues include integrating room occupancy data with Internet of Things (IoT) technologies to create intelligent environments. This involves implementing sensor networks that communicate with other smart devices to optimize lighting, heating, and ventilation based on real-time occupancy patterns.

ACKNOWLEDGMENTS

Professor Musfira Jilani played a crucial role in the completion of this project, offering extensive guidance, support, valuable advice, and providing critical feedback.

REFERENCES

[1] Applied Statistics: Deterministic and forecast-adaptive time-dependent models, by Abraham, B. and G. E. P. Box, edition 1, Oxford University Press
[2] Time Series Analysis, by James D. Hamilton, edition 1, 1994, Princeton University Press ISBN: 0691042896
[3] Forecasting: principles and practice, by Rob Hyndman and George Athanasopoulos, Edition 2
[4] The R Book, by Crawley, Michael J., edition, 2013, A John Wiley Sons, Ltd.
[5] Scaling Up Machine Learning: Parallel and Distributed Approaches, by R. Bekkerman, M. Bilenko, and J. Langford, edition, 2011, Cambridge University Press
[6] Logistic Regression: A Primer, by Fred C. Pampel, Second edition, 2020, Sage Publications,Inc
[7] Datacamp *Time Series Forecasting*. Footnote:1
[8] Geeksforgeeks website *Topic: Logistic Regression in Machine Learning*. Footnote:2
[9] Python for Data Analysis, by Dr. W. McKinney, 2nd Edition, 2017, O'Reilly Media publishers

[1] [2]

[1]https://www.datacamp.com/tutorial/tutorial-time-series-forecasting
[2]https://www.geeksforgeeks.org/understanding-logistic-regression/