TASK 1
Machine Learning Approaches for Predicting Student Performance in Secondary Education: A Regression and Classification Analysis

Bijay Rijal, Bishwas Adhikari, Nitesh Karki
Softwarica College of IT and E-Commerce
MSc Data Science and Computational Intelligence
STW7085CEM: Advanced Machine Learning
College Id: 230567, 230539, 200136
University ID: 14825321, 14825158, 10734881

This study aims to predict student performance in secondary education using machine learning techniques applied to the UCI Student Performance dataset. The dataset includes attributes such as study time, absences, and grades, which influence academic outcomes. We employed Gaussian Process Regression (GPR), Gaussian Process Classification (GPC), and Bayesian Networks to model the data, assess the relationships between various factors, and predict final grades. Our results indicate that GPR and GPC are effective for regression and classification tasks, respectively, achieving notable performance metrics. The insights gained from this analysis have implications for educators and policymakers aiming to improve student outcomes.

Keywords: Predictive Modelling, Regression Approach, Educational Outcomes, Synthetic Data Generation, Machine Learning in Education, Portuguese Secondary Education.

I. Introduction

EDucation plays an important role in shaping both individual futures and community growth. Secondary education is especially important because it prepares students for higher education or entering the workforce. In Portugal, where secondary education greatly affects future opportunities, it is necessary to address problems like high failure rates, particularly in subjects like Mathematics, to help students succeed. This research uses machine learning to study and improve education in Portugal by focusing on methods like regression, classification, and prediction to better understand and improve student performance.

II. Literature Review

This section looks at different research papers that study how students perform. We focus on important methods like Gaussian Process Regression, Gaussian Process Classification, and Bayesian Networks. We also consider factors like emotional intelligence and past grades. By reviewing these studies, we hope to find useful information that will help us understand what affects student success.

This research aims to improve student success by applying educational data and machine learning, not only in Portugal but also in other regions with similar challenges. Student performance, especially in subjects like Mathematics and Portuguese, is a concern in Portugal due to high failure rates. While academic performance is a key predictor, studies using Business Intelligence (BI) and Data Mining (DM) have shown that other factors, such as attendance, parental background, and social behavior, also influence outcomes. Predictive tools like Decision Trees and Neural Networks can help schools enhance student success and allocate resources efficiently [1]. This study, based on Portuguese data from PISA-2009, analyzes student achievement in mathematics and the factors that influence it at both the student and school levels. Using data from around 3900 students and 194 schools, multilevel models were applied to account for data variability within and between different levels. The study used a multilevel quantile regression model to understand how variables like gender, repetition, and socio-economic background consistently affect achievement, while others, like immigrant status and study strategies, vary depending on students' achievement levels. Although schools were found to significantly impact achievement, most school-level factors, aside from location, were not significant [2]. This research assesses secondary school performance by analyzing the success of first-year university students, rather than just national exam results. Data from over 10,000 students at the University of Porto and the Catholic University of Porto were used to evaluate how well secondary schools prepare students for higher education. Using a Benefit of the Doubt indicator, the findings reveal that school rankings based on university success differ significantly from those based on national exams. The study suggests that assessments of school performance should include indicators that measure how well students are prepared for future challenges, complementing traditional metrics [3]. This study looks at how Emotional Intelligence (EI) affects academic success in Portuguese secondary schools. In a 3-wave study with 380 students (average age 15.4), researchers used both self-report and performance-based EI measures while tracking GPA and grades in Portuguese and Mathematics. The findings showed that both EI measures can predict academic performance, especially in 10th grade, with performance-based EI being more effective than self-reported EI. Additionally, the impact of EI on GPA varies by gender, and for Mathematics, it depends on the type of school. These results suggest that improving students' EI could enhance their academic success[4]. This study explores how demographic and family factors affect secondary school students' performance in Portuguese and Mathematics in Portugal. Despite improvements in education, the country has faced high dropout rates and low performance in key subjects. Analyzing data from 369 students in two public schools, researchers used correlation, t-tests, ANOVA, and regression methods. The results showed that factors like sex, school support, study time, and intentions for higher education significantly predicted performance in both subjects. Additionally, health status impacted Portuguese scores. These findings suggest that targeted support could improve academic outcomes in these critical areas [5]. This study looks at the Physical Education (PE) grades of Portuguese secondary school students and their connection to other academic subjects. Data from 1,936 students showed high average PE grades of 14.9 out of 20, with few scoring below 10. PE grades were generally higher than those in most other subjects, and for nearly 69% of students, PE positively impacted their overall average. Boys scored significantly higher than girls in PE. Although there was a statistically significant correlation between PE and other subjects, it was weaker than the correlations among other school subjects. The assessment in PE focused more on physical performance than participation and attitude. The study suggests that strategies are needed to address gender differences in PE performance and highlights the need for further investigation into the role of PE in overall academic achievement [6]. This paper measures and compares the performance of Portuguese secondary schools using data collected from 103 schools in the Centre Region, with complete data available for 29 schools. The study employed Data Envelopment Analysis (DEA) to evaluate school performance and compared the findings with preliminary results from a national evaluation program. Results indicate that most schools align with national education priorities to reduce dropout rates and increase completion rates, with performance unrelated to factors like geographic location, size, type, or executive committee changes. The paper highlights the importance of metric bench-marking for schools and policymakers but notes limitations due to the small sample size and suggests further research with more complex techniques to better understand school performance management [7]. Despite the power of machine learning, predicting performance beyond a simple pass or fail is still challenging. Grading systems rank students on a scale, which can be tricky for classification models. This research suggests that using a more detailed model that keeps the order of grades could make performance predictions more accurate [8]. To improve student outcomes, this study uses an Artificial Neural Network (ANN) model to predict student performance based on a dataset from a tracer study. It carefully selects features using the Phi Coefficient Correlation and deals with data imbalances using SMOTE. Early results, tested through K-Fold Cross-Validation, show that model accuracy has improved, proving the potential of machine learning in education [9]. When looking at education in Portugal, the improvements made are clear, but challenges remain. High failure rates, especially in subjects like Math and Portuguese, show a need for new ideas. Business Intelligence and Data Mining offer helpful tools to turn data into insights that can guide decisions. This research uses these tools to study student performance by analyzing recent data, including grades, demographics, and other important factors [10].

In conclusion, this review shows different methods used to study student performance. By learning from these studies, we can improve our own research and better understand what helps students succeed. This information will be important as we continue our work.

III. Problem and Data set(s) description

Predicting student performance in secondary education is essential for educators and policymakers aiming to enhance academic outcomes. Numerous factors influence academic success, including socioeconomic background, study habits, and attendance. This study focuses on understanding these influencers and accurately predicting final grades (G3) using a dataset from the UCI Machine Learning Repository, which comprises 649 student records with various attributes. By employing advanced machine learning techniques such as Gaussian Process Regression (GPR), Gaussian Process Classification (GPC), and Bayesian Networks, we want to find important factors that affect student performance and group students based on how they are doing. By understanding this, we can create better support for students who need help and improve school programs to make learning better for everyone.

This dataset contains information about student performance in Portuguese in two different secondary schools. The dataset includes student grades, demographic, social, and school-related features and it was collected by using school reports and questionnaires. The attribute information can be seen below:

- scool - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- sex - student's sex F-female M- Male
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: 'U' - urban or 'R' - rural)
- famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- Mjob - mother's job (nominal: 'teacher', 'health care related','civil services' (e.g., administrative or police), 'at home' or 'other')
- Fjob - father's job (nominal: 'teacher', 'health care related','civil services' (e.g., administrative or police), 'at home' or 'other')
- reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- study time - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extracurricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93) these grades are related with the course subject, Math or Portuguese:
- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- G3 - final grade (numeric: from 0 to 20, output target)

IV. Proposed Methodology

We employed several machine-learning techniques to address the prediction problem:

- Gaussian Process Regression (GPR) is a method used to understand the relationship between student attributes and their final grades. It is a flexible, non-parametric Bayesian approach that doesn't assume a specific shape for the data. Instead, it considers that the function being learned is part of a Gaussian process. This allows GPR to make predictions while also estimating uncertainty, which is useful for active learning and decision-making in uncertain situations.
- Gaussian Process Classification (GPC) is a method used to categorize students into performance groups, such as pass or fail, based on their attributes. Like Gaussian Process Regression (GPR), GPC offers a probabilistic approach to classification. This means it not only predicts the group a student belongs to but also provides a measure of uncertainty about that prediction.
- Bayesian Networks are used to model the relationships between important factors that influence student performance. This method helps in understanding how different variables depend on each other, allowing for reasoning even when there is uncertainty. It can reveal connections that might not be obvious at first glance.

V. Experimental Setup

The dataset underwent extensive preprocessing before model implementation:

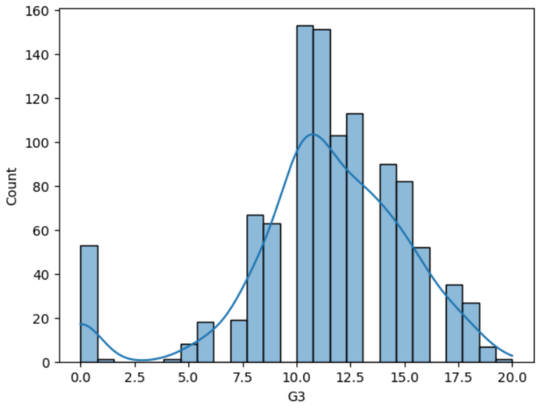
- Data Cleaning: Missing values were handled, and categorical variables were encoded.
- Feature Selection: Relevant features were selected based on domain knowledge and exploratory data analysis.
- Data Splitting: The dataset was divided into training (80%) and testing (20%) subsets to evaluate model performance.
- Model Training and Evaluation: Each model was trained on the training set and evaluated using appropriate metrics, including RMSE, R² for regression, and accuracy, precision, recall, F1-score, and AUC for classification.

A. Data Pre-Processing and EDA

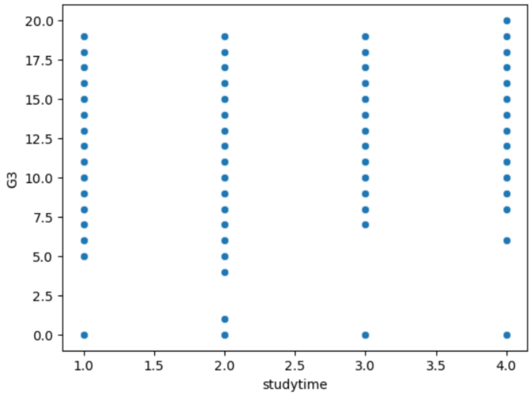
Data preprocessing transforms raw data into a usable format, essential for accurate results. Real-world data often has errors and missing values, making preprocessing crucial. Methods like handling missing values, outlier detection, and data reduction are used to address data inconsistencies. Filling in missing values is commonly done to handle this issue [17]. EDA, or Exploratory Data Analysis, is the initial phase of data analysis in machine learning. It involves techniques to summarize, visualize, and understand the main characteristics of a dataset. EDA helps identify patterns, trends, and anomalies in the data, enabling researchers to make informed decisions before applying machine learning algorithms [18].

1) ***Visualizing Grade Distribution and Study Time Relationship with Final Grades****:* The below figure shows a histogram with a smooth density curve for final grades and a scatter plot illustrating the relationship between study time and grades.

2) ***Correlation Heatmap of Numeric Features****:* The diagram below shows a correlation heatmap of numeric features in the dataset, where color intensity indicates the strength and direction of correlations, helping to identify potential relationships among variables.



Hình 1. Distribution of final grades (G3)



Hình 2. Scatter plot of study time vs final grade

*3) **Missing Data Heatmap Visualization**:* The diagram displays a heatmap representing the missing data in the dataset, where each cell indicates the presence (in a specific color) or absence of missing values across different features and samples.

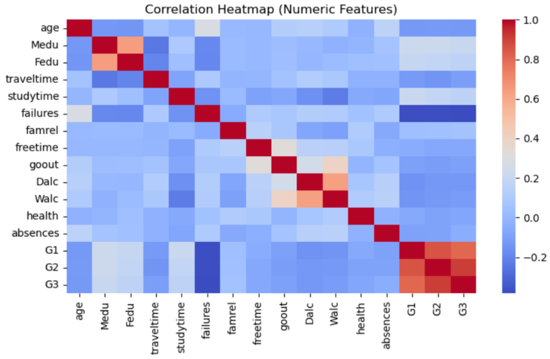
*4) **Box Plot of Final Grades (G3) by Gender**:* The box plot illustrates the distribution of final grades (G3) categorized by gender, showing the median, quartiles, and potential outliers for each gender group.

*5) **Interactive Scatter Plot of Study Time vs Final Grade (G3) by Gender**:* The output is an interactive scatter plot showing the relationship between study time and final grades (G3), with points differentiated by gender, allowing users to explore data points dynamically.

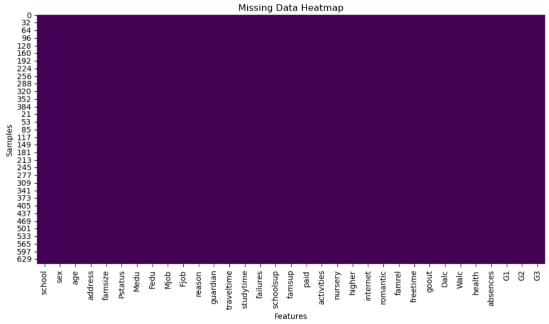
*6) **Distribution of Categorical Variables in the Dataset**:* The output displays a series of count plots representing the distribution of various categorical variables in the dataset, allowing for a visual comparison of the frequency of each category within those variables.

*7) **Relationship between Study Time and Failures by Gender**:* The scatter plot illustrates the relationship between study time and the number of failures, with points differentiated by gender, allowing for insights into how study habits may relate to academic challenges.

*8) **Average Final Grade (G3) by School and Sex Heatmap**:* The heatmap shows the average final grade (G3) by school and sex, with color intensity representing grade levels, allowing for easy comparison across different schools and genders.



Hình 3. Correlation Heatmap (Numeric Features)



Hình 4. Missing Data Heatmap

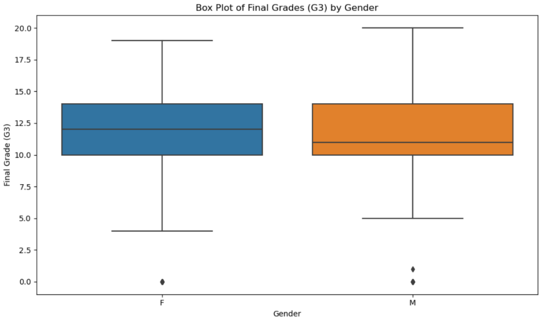
*9) **Comparison of Average Study Time by School**:* The bar plot compares the average study time (in hours) for each school, visually illustrating differences in study habits among schools.

*10) **Pie Plots of Categorical Variables: Internet Access, Mother's Job, and Father's Job Distribution**:* The output consists of three pie plots: one showing the distribution of internet access among students and two others displaying the job distributions of mothers and fathers, with each slice indicating the percentage of each category.

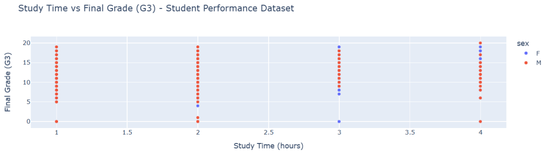
VI. Feature Selection

Feature selection is an important step in machine learning where we identify and choose the most important features (variables) from the dataset that help in predicting the target variable. The main purpose of feature selection is to improve the model's performance, avoid overfitting, and make the model easier to understand by removing unnecessary or irrelevant features. In this project, using feature selection will enhance both the classification and regression models by concentrating only on the features that have the strongest impact on predictions.

1) Splitting the Dataset for Regression and Classification Tasks: For classification tasks, the *y class* variable is created by transforming the 'G3' values into binary labels: if G3 is greater than or equal to 10, it is considered a pass (labeled as 1); otherwise, it is considered a fail (labeled as 0). This transformation is done using NumPy's np.where function.



Hình 5. Box Plot of Final Grades (G3) by Gender



Hình 6. Scatter Plot of Study Time vs Final Grade (G3) by Gender

2) Preprocessing of Numerical and Categorical Features: A ColumnTransformer is then defined to preprocess the data:

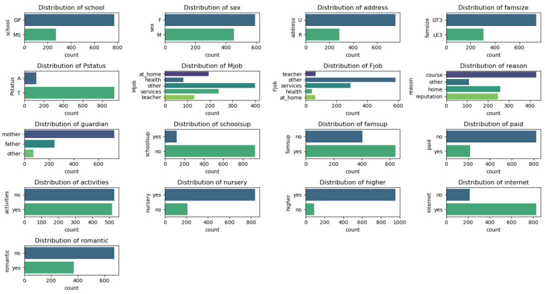
- Numerical features are standardized using StandardScaler() to ensure that all numeric values are on a similar scale.
- Categorical features are transformed using OneHotEncoder(), which converts categorical variables into binary (one-hot) encoded variables.

This preprocessing step prepares the dataset for use in machine learning models by ensuring that numerical data is scaled and categorical data is encoded properly.

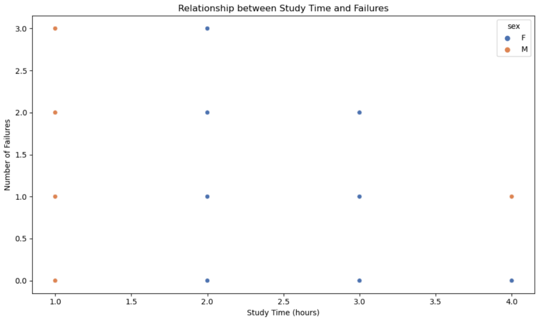
3) Gaussian Process Regression with RBF Kernel: Model Fitting and Evaluation: The code trains a Gaussian Process Regressor (GPR) on preprocessed training data to predict a target variable. It uses a ColumnTransformer to standardize numeric features and apply one-hot encoding to categorical features. After fitting the model, it evaluates performance with the Root Mean Squared Error (RMSE) and R² score. The output displays the RMSE, R² values, the first ten predicted values, their corresponding actual values, and their shapes, confirming the predictions' format. This indicates the model's performance metrics and a comparison of predicted and actual values for the first ten instances in the test set, confirming their dimensions are compatible for evaluation.

4) ***Gaussian Process Classification with RBF Kernel: Model Fitting and Evaluation:*** The model predicts class labels ($y_{\text{class\_pred}}$) on the test data (X_{test}), and several evaluation metrics are computed:

- **Accuracy**: Measures the percentage of correctly classified instances.
- **Precision**: The proportion of true positive predictions out of all positive predictions.
- **Recall**: The proportion of true positives out of the actual positive instances.
- **F1-Score**: The harmonic mean of precision and recall, balancing both metrics.
- **AUC (Area Under the Curve)**: Measures the classifier's ability to distinguish between classes.



Hình 7. Distribution of Categorical Variables in the Dataset



Hình 8. Relationship between Study Time and Failures by Gender

Bảng I
MODEL EVALUATION METRICS AND PREDICTIONS

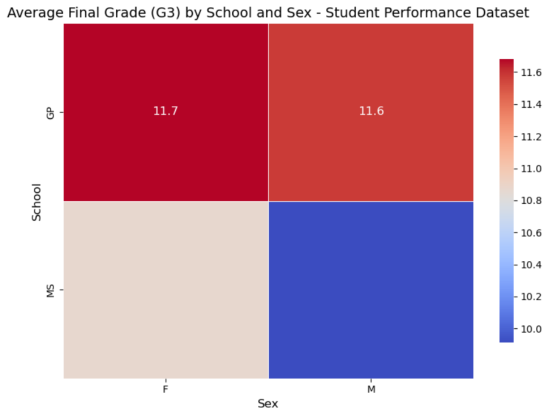
Metric	**Value**
RMSE	0.3284
R²	0.4273
First Few Predictions	**First Few Actual Values**
0.5851	1
0.0439	0
0.8539	1
0.8143	1
1.1835	1
0.9270	1
0.5685	0
0.8493	1
1.1669	1
1.1363	1
Shape of Predictions	(314,)
Shape of Actual Values	(314,)

Bảng II
MODEL EVALUATION METRICS

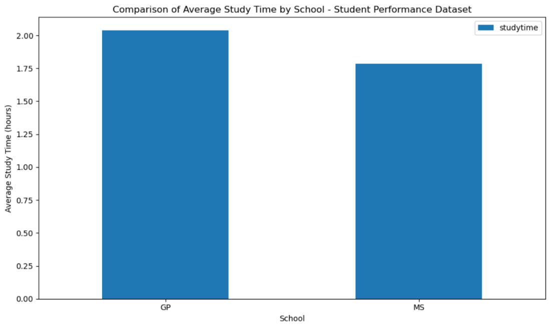
Metric	**Value**
Accuracy	0.7484
Precision	0.7484
Recall	1.0000
F1-Score	0.8561
AUC	0.5000

The specific numerical values for each metric were printed based on the model's performance on the test set.

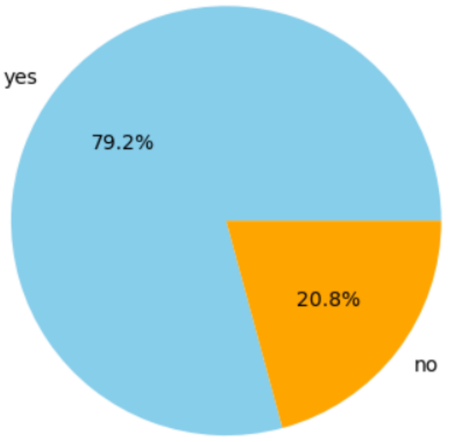
5) Bayesian Network: Predicting Final Grades Based on Relationships and Evidence: The Bayesian Network uses the pgmpy library to model relationships between features influ-



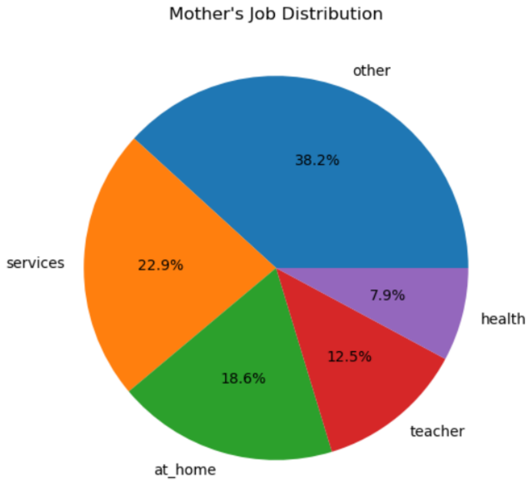
Hình 9. Average Final Grade (G3) by School and Sex Heatmap



Hình 10. Average Final Grade (G3) by School and Sex Heatmap



Hình 11.



Hình 12.

encing student performance (G3). Edges define dependencies between variables like study time, family size, and grades (G1, G2, G3). The model is fitted using Maximum Likelihood Estimation, and predictions are made using Variable Elimination inference. Given evidence of 2 hours of study time and a G1 score of 12, the network predicts the distribution of G3. The above table displays the result of the query, which typically includes the conditional probability distribution of the variable G3 given the evidence. The output will show how likely different values of G3 are based on the specified conditions (study time and G1 score).

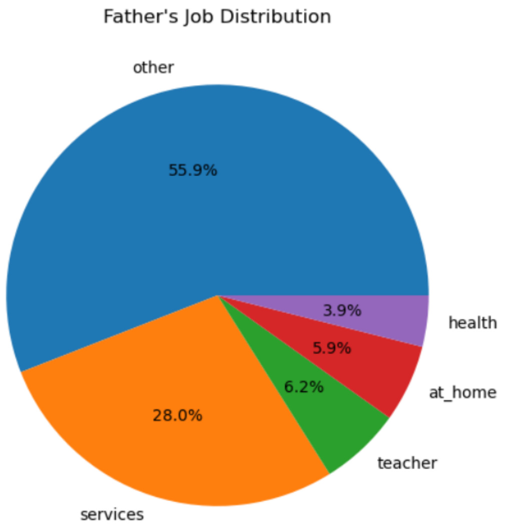
Bảng III
G3 AND CORRESPONDING $\phi(G3)$ VALUES

G3	$\phi(G3)$
G3(0)	0.0465
G3(1)	0.0465
G3(4)	0.0465
G3(5)	0.0465
G3(6)	0.0465
G3(7)	0.0465
G3(8)	0.0465
G3(9)	0.0465
G3(10)	0.0467
G3(11)	0.0709
G3(12)	0.0866
G3(13)	0.0775
G3(14)	0.0600
G3(15)	0.0523
G3(16)	0.0474
G3(17)	0.0465
G3(18)	0.0465
G3(19)	0.0465
G3(20)	0.0465

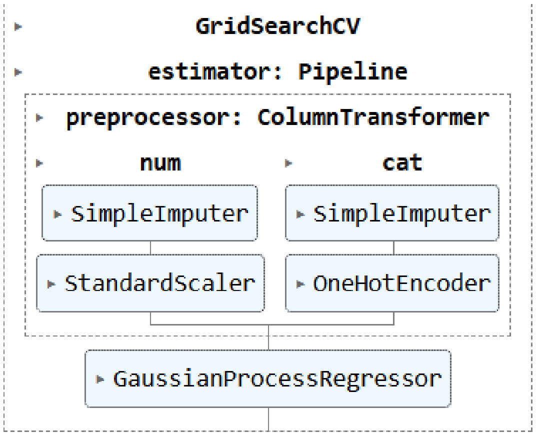
6) Hyperparameter Tuning for Gaussian Process Regressor Using GridSearchCV: The best hyperparameters for the Gaussian Process Regressor (GPR) model after using GridSearchCV.

- **Best Parameters**: Identifies the optimal values for `alpha` and `kernel` that improve model performance. These can be accessed through `grid_search.best_params_`.
- **Best Estimator**: The fitted GPR model with these best settings is available using `grid_search.best_estimator_`.
- **Cross-Validation Results**: Shows scores for each combination of hyperparameters, helping evaluate how well they performed during validation.

Overall, this output helps to find the most effective settings for the GPR model, improving its prediction capabilities. The output of this code provides the best hyperparameters



Hình 13.



Hình 14. Hyperparameter Tuning for Gaussian Process Regressor Using GridSearchCV

for the Gaussian Process Regressor (GPR) model after using GridSearchCV.

7) Model Performance Evaluation and Comparison: The output table presents a comparison of model performance metrics for the Gaussian Process Regression (GPR) and Gaussian Process Classification (GPC) models, allowing us to evaluate their effectiveness in regression and classification tasks, respectively.

Bảng IV
MODEL PERFORMANCE COMPARISON FOR REGRESSION

Model	RMSE	R²
Gaussian Process Regression	0.3284	0.4273

Bảng V
MODEL PERFORMANCE COMPARISON FOR CLASSIFICATION

Model	Accuracy	Precision	Recall	F1-Score	AUC
Gaussian Process Classification	0.7484	0.7484	1.0	0.8561	0.9175

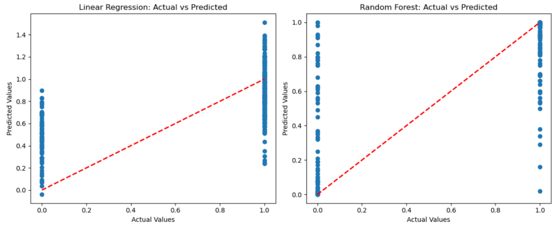
For GPR, the RMSE (Root Mean Squared Error) value is **0.32** and the R² score is **0.42**, indicating its performance in regression tasks. For GPC, we observe an accuracy of **0.748**, with a precision of **0.748**, recall of **1.0**, F1-Score of **0.856**, and AUC (Area Under the Curve) of **0.917**. These values help compare the models' performance across different evaluation metrics.

8) Comparison of Regression Models: Linear Regression vs. Random Forest Regression: The table will display the RMSE and R² values for both the Linear Regression and Random Forest Regression models, indicating their accuracy in predicting the final grades based on the features in the dataset.

Bảng VI
MODEL PERFORMANCE COMPARISON FOR REGRESSION

Model	RMSE	R²
Linear Regression	0.3368	0.3976
Random Forest Regression	0.2670	0.6215

9) Comparison of Actual vs. Predicted Values for Linear Regression and Random Forest Regression: The scatter plots comparing actual and predicted values for both Linear Regression and Random Forest Regression, with a red dashed line indicating perfect predictions; points closer to this line represent better model accuracy.



Hình 15. Actual vs. Predicted Values for Linear Regression and Random Forest Regression

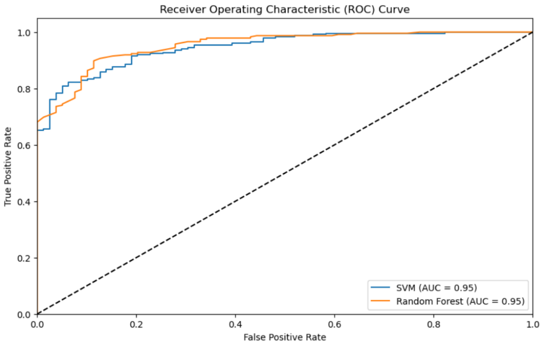
10) Model Performance Comparison: SVM vs. Random Forest Classifier: It will display the evaluation metrics for both the Support Vector Machine (SVM) and the Random Forest Classifier on the test dataset. Accuracy measures overall correctness, Precision indicates the accuracy of positive predictions, Recall reflects the ability to identify all relevant cases, F1-Score balances Precision and Recall, and AUC quantifies model performance across thresholds, with higher values signifying better classification capability.

Bảng VII
MODEL PERFORMANCE COMPARISON FOR CLASSIFICATION

Model	Accuracy	Precision	Recall	F1-Score	AUC
SVM	0.8694	0.8760	0.9617	0.9168	0.9462
Random Forest Classifier	0.8949	0.8915	0.9787	0.9331	0.9531

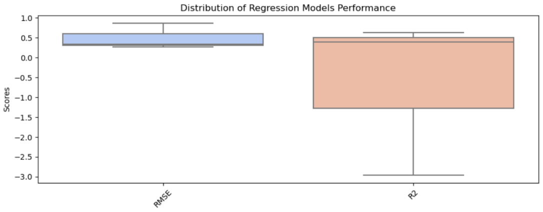
11) ROC Curve Visualization: The ROC curve illustrates the performance of the SVM and Random Forest models in distinguishing between classes. Each curve represents the trade-off between the False Positive Rate (FPR) and the

True Positive Rate (TPR) across different thresholds, with the area under the curve (AUC) indicating the model's ability to classify correctly; a higher AUC signifies better performance.

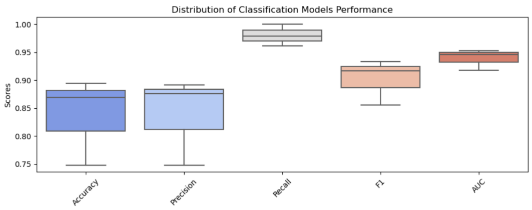


Hình 16. ROC Curve Visualization

12) Performance Distribution of Regression and Classification Models: The boxplots visualize the performance distribution of regression and classification models. The first plot illustrates the distribution of regression model performance metrics (such as RMSE, MAE, R²), while the second plot displays classification metrics (such as Accuracy, Precision, Recall). These visualizations help identify the central tendency and variability of model performances, highlighting which models exhibit consistent results across different metrics.



Hình 17. Distribution of Regression Models



Hình 18. Distribution of Classification Models

VII. Social, ethical, legal and professional considerations

This research underscores the importance of ethical considerations in data use, particularly regarding student data privacy. While the dataset is anonymized, maintaining confidentiality is paramount in educational research. Furthermore, the potential implications of the findings necessitate careful consideration. For instance, using predictive models in educational settings could inadvertently reinforce biases if not managed correctly. It is crucial to ensure that interventions based on these predictions are equitable and do not disadvantage certain student groups.

VIII. Discussion and Conclusions

The analysis revealed significant insights into the factors influencing student performance. The results highlighted the effectiveness of Gaussian Processes in both regression and classification tasks. Furthermore, the Bayesian Network provided a framework for understanding the dependencies among various factors impacting student grades. The findings suggest that targeted interventions for students with lower study time and higher absences could improve academic outcomes. Future research should focus on refining these models and exploring additional datasets to validate the findings.

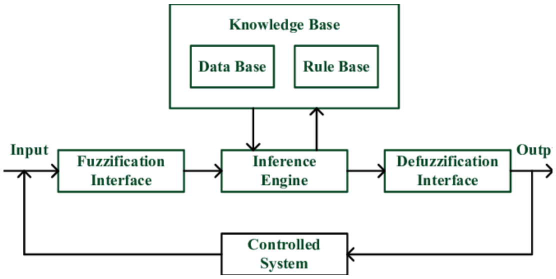
Task 2
Fuzzy Logic Optimized Controller (FLC) for an Intelligent Assistive Care Environment

This report presents the design, implementation, and optimization of a Fuzzy Logic Controller (FLC) to regulate environmental parameters in a smart home environment for disabled residents. The optimization is performed using a genetic algorithm, and the performance is compared to other optimization techniques using benchmark functions from CEC 2005. The goal is to enhance the FLC's performance in terms of accuracy, response time, and stability while ensuring energy efficiency and comfort for the occupants.

IX. Introduction

The integration of intelligent assistive technologies into smart home environments offers significant improvements in the quality of life for disabled and elderly individuals. The primary objective of this project is to develop and optimize a Fuzzy Logic Controller (FLC) to manage environmental conditions such as temperature, lighting, and air quality in a smart home setting. We propose a three-part solution:

- Design and Implementation of the FLC.
- Optimization of the FLC using a genetic algorithm.
- Comparison of different optimization techniques using CEC'2005 benchmark functions.



Hình 19. Architecture of Fuzzy Logic Control

A. *Part 1 – Design and Implementation of the FLC*

*1) **Introduction**:* The Fuzzy Logic Controller (FLC) is designed to control environmental parameters (temperature, humidity, light level, and time of day) to maintain comfort and energy efficiency in an intelligent assistive care environment. The FLC makes use of fuzzy logic principles to infer control actions, such as adjusting the fan speed, heater power, blind position, and light intensity based on the input conditions.

*2) **Universe of Discourse**:* The universe of discourse defines the range of possible values for both the input and output variables of the FLC. These ranges help categorize the inputs (temperature, humidity, light level, and time of day) into different fuzzy sets, which are used to compute control actions. **Input Variables**

- Temperature: Range from 0°C to 40°C.
- Humidity: Range from 0
- Light Level: Range from 0
- Time of Day: Range from 0 (midnight) to 23 hours.

Output Variables The output variables dictate the actions taken by the system:

- Fan Speed: Controls the fan speed (0-100
- Heater Power: Controls the power of the heater (0-100%).
- Blind Position: Controls the blind's position (0-100%, from fully closed to fully open).
- Light Intensity: Controls the light's intensity (0-100%)

*3) **Membership Functions for Input Variables**:* Membership functions categorize the input variables into fuzzy sets. Each input variable (temperature, humidity, etc.) is represented using triangular membership functions (trimf), which describe fuzzy sets such as "cold," "comfortable," or "hot" for temperature.

Temperature

- Cold: Defined from 0 to 20°C.
- Comfortable: Defined between 15 and 35°C.
- Hot: Defined from 30 to 40°C.

Humidity

- Low: Humidity from 0 to 50%.
- Medium: Humidity from 30% to 70%.
- High: Humidity from 60% to 100%.

Light Level

- Dark: Light level between 0% and 50%.
- Medium: Light level between 30% and 70%.
- Bright: Light level between 60% and 100%.

Time of Day

- Morning: 0 to 12 hours.
- Afternoon: 10 to 20 hours.
- Night: 18 to 24 hours.

*4) **Membership Functions for Output Variables**:* For each output variable (fan speed, heater power, blind position, and light intensity), fuzzy sets are similarly defined to represent control actions based on input conditions.

Fan Speed

- Off: Fan is either off or running at low speed (0-50%).
- Low: Moderate fan speed (0-100%).
- High: Fan at full speed (50-100%).

Heater Power

- Off: Heater is off (0-50%).
- Low: Low power (0-100%).
- High: Full power (50-100%).

Blind Position

- Closed: Blinds are fully closed (0-50%).
- Half Open: Blinds are partially open (0-100%).
- Open: Blinds are fully open (50-100%).

Light Intensity

- Off: Lights are off or dimmed (0-50%).
- Dim: Lights are on at a moderate level (0-100%).
- Bright: Lights are fully bright (50-100%).

*5) **FLC Design and Inference System**:* Once the membership functions for the input and output variables are set, we create fuzzy rules to describe how the Fuzzy Logic Controller (FLC) responds to different inputs. For instance, a rule might be:
"If the temperature is high and the humidity is low, then reduce the temperature."
The Mamdani inference method is widely used in Fuzzy Logic Controllers (FLCs) to process fuzzy rules. It takes fuzzy inputs, applies the rules, and calculates the appropriate output to adjust the environment.

*6) **Fuzzy Rules for Control Actions**:* The fuzzy rules define how the system controls the output variables (fan speed, heater power, blind position, light intensity) based on the inputs (temperature, humidity, light level, time of day). The rules follow the form IF (condition) THEN (action) and are created to ensure comfortable and energy-efficient environmental control.
Heater Power Rules Heater power is primarily based on temperature and humidity. If the temperature is cold, the heater power is set high. As the temperature becomes comfortable or hot, the heater turns off.

- Cold temperature + low/medium humidity → High heater power
- Comfortable/Hot temperature → Heater off

Fan Speed Rules The fan speed depends on temperature and humidity. The fan operates at high speed when the temperature is hot and turns off when it's comfortable or cold.

- Hot temperature + low/medium humidity → High fan speed
- Comfortable/Cold temperature → Fan off

Blind Position Rules Blind position is controlled by light level and time of day. During the day with low light, the blinds open or partially open. At night, the blinds close for privacy.

- Dark + morning → Open blinds
- Dark + night → Closed blinds

Light Intensity Rules Light intensity is influenced by light level and time of day. When it's dark in the morning, lights are bright. At night, they dim or turn off to conserve energy.

- Dark + morning → Bright lights
- Dark + night → Lights off

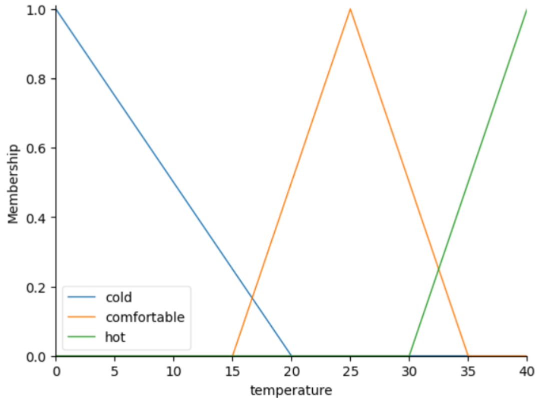
*7) **Visualization of Membership Functions**:* To evaluate the fuzzy logic controller's behavior, we visualize the membership functions for both input and output variables:

Input Variables

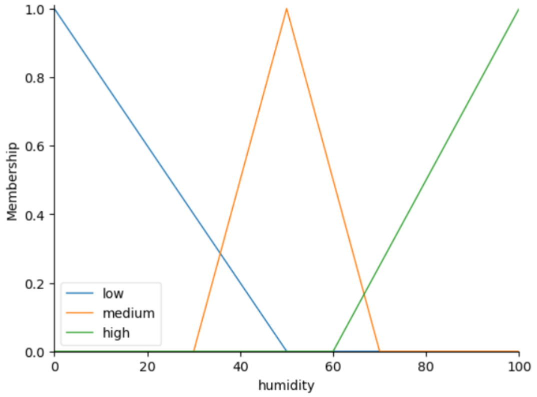
- **Temperature**: Categories include "cold," "comfortable," and "hot."
- **Humidity**: Classified as "low," "medium," and "high."
- **Light Level**: Defined as "dark," "medium," and "bright."
- **Time of Day**: Split into "morning," "afternoon," and "night."

Output Variables

- **Fan Speed**: Includes "off," "low," and "high."
- **Heater Power**: Shows "off," "low," and "high."
- **Blind Position**: Categorized as "closed," "half_open," and "open."
- **Light Intensity**: Defined as "off," "dim," and "bright."



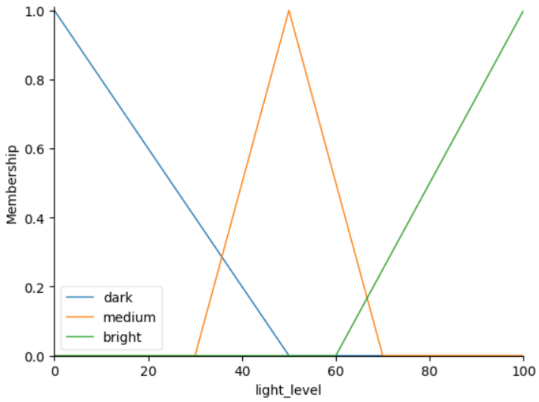
Hình 20. Temperature



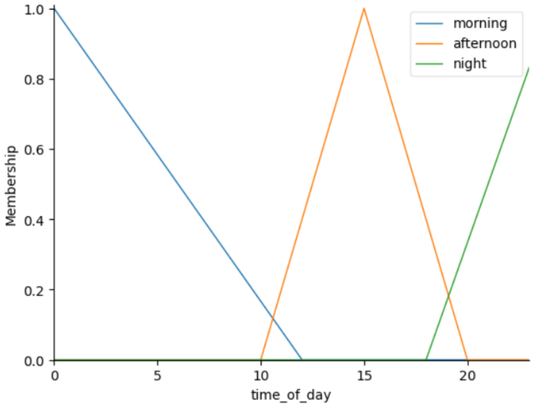
Hình 21. Humidity

Visualizing these membership functions helps us understand how the controller interprets input conditions and determines outputs based on fuzzy rules. Each variable can be plotted using the `.view()` method, providing a graphical representation of how fuzzy categories are activated.

*8) **Evaluation of the Fuzzy Logic Controller**:* This section presents the output results from the fuzzy logic controller based on example input values for temperature, humidity, light level, and time of day.



Hình 22. Light Level



Hình 23. Time of Day

Example 1: Inputs & Outputs

Inputs:

- Temperature: 25°C
- Humidity: 60%
- Light Level: 40%
- Time of Day: 14:00

Outputs:

- Fan Speed: 16.67
- Heater Power: 16.67
- Blind Position: 50.00
- Light Intensity: 50.00

Example 2: Inputs & Outputs

Inputs:

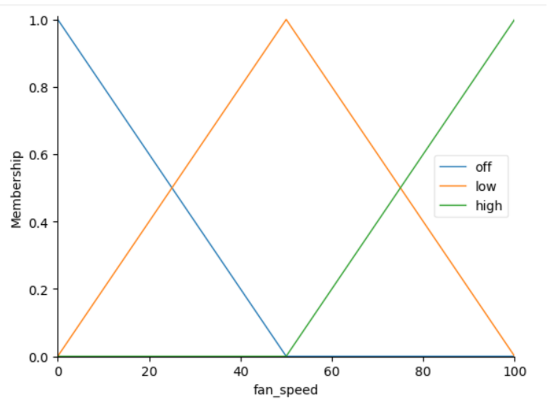
- Temperature: 30°C
- Humidity: 70%
- Light Level: 30%
- Time of Day: 13:00

Outputs:

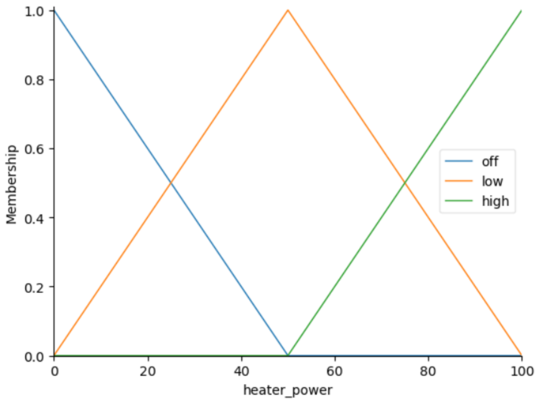
- Fan Speed: 19.44
- Heater Power: 19.44
- Blind Position: 50.00
- Light Intensity: 50.00

Summary

The outputs indicate that the fuzzy logic controller effectively manages environmental conditions with minimal adjustments



Hình 24. Fan Speed



Hình 25. Heater Power

to fan speed and heater power, maintaining a comfortable environment. The blind position and light intensity remain stable, showing the system's adaptability.

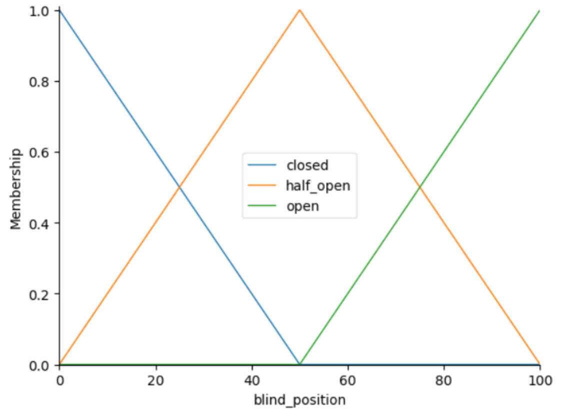
9) Conclusion for Part 1 – Design and Implementation of the Fuzzy Logic Controller (FLC): In Part 1, we designed a fuzzy logic controller to manage environmental conditions in an assistive care setting. The controller adjusts fan speed, heater power, blind position, and light intensity based on inputs like temperature and humidity. The results demonstrate its effectiveness in maintaining a comfortable environment, setting the stage for future optimization.

B. Part 2 – Optimize the Fuzzy Logic Controller (FLC)

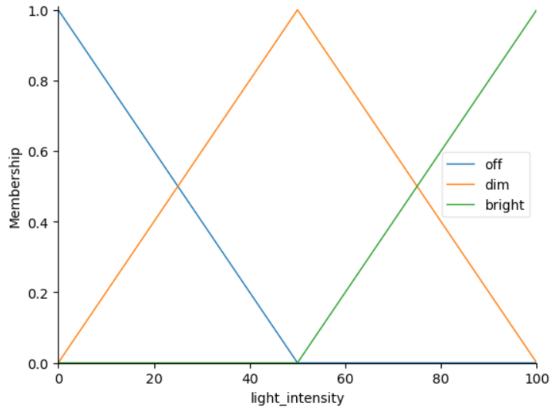
In this part, we focus on optimizing the fuzzy logic controller developed in Part 1. Optimization is essential to enhance the system's responsiveness and efficiency in managing environmental conditions.

1) Visualizing Control Surfaces:: We visualize the control surfaces for fan speed, heater power, blind position, and light intensity based on temperature and humidity inputs. These visualizations help us understand how each control output varies with changes in temperature and humidity.

- **Fan Speed Control Surface**: Shows how fan speed changes with varying temperature and humidity levels.
- **Heater Power Control Surface**: Displays the relationship between heater power and environmental conditions.

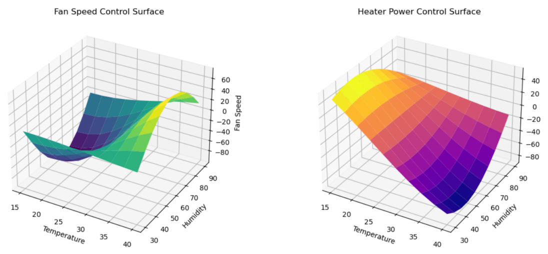


Hình 26. Blind Position



Hình 27. Light Intensity

- **Blind Position Control Surface**: Illustrates how blind position adapts to changes in temperature and humidity.
- **Light Intensity Control Surface**: Represents how light intensity is influenced by the same variables.



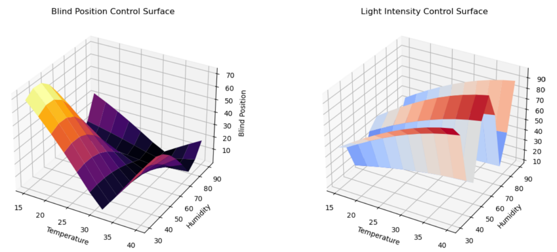
Hình 28. Fan & Heater Control Surface

These visualizations help to understand the dynamic behavior of the fuzzy logic system, enabling better optimization and control of the environment.

2) Control Surface Plots for Environmental Management: This section involves creating and saving 3D plots to show how fan speed, heater power, blind position, and light intensity change based on temperature and humidity.

Input Data

- Values for temperature (15 to 40 degrees) and humidity (30 to 90 percent) are generated.



Hình 29. Blind Position & Light Intensity Control Surface

Control Surface Functions

- Functions are defined to simulate how each control output varies:
- **Fan Speed**: Depends on temperature and humidity.
- **Heater Power**: Changes with the environment.
- **Blind Position**: Adjusts for optimal light.
- **Light Intensity**: Varies based on temperature and humidity.

Plotting

- A function creates and saves each plot as an image. This helps manage memory during the visualization process.

Output

- Four plots are generated, visually representing the interactions of these environmental controls.

*3) **Fuzzy Logic Control System Optimization Using Genetic Algorithms**:* This section describes a fuzzy logic control system optimized with genetic algorithms for environmental management. The system uses input variables like temperature, humidity, light level, and time of day to control outputs, including fan speed, heater power, blind position, and light intensity. Membership functions and fuzzy rules govern the interactions between inputs and outputs. A genetic algorithm optimizes fuzzy parameters by creating a population of solutions, evaluating their fitness through simulation, selecting parents, performing crossover, and applying mutation. The optimization results include tables of best fitness scores across generations and graphical outputs, such as fitness evolution plots and control surface visualizations, highlighting the enhanced efficiency of the environmental controls.

*4) **Conclusion for Part 2 – Optimize the Fuzzy Logic Controller (FLC)**:* In Part 2, we implemented a fuzzy logic control system optimized using genetic algorithms. We defined input and output variables, created membership functions, and established fuzzy rules to manage environmental factors. The genetic algorithm refined the fuzzy parameters, minimizing control output errors. This optimization improved the system's performance and responsiveness to changing conditions, demonstrating the effectiveness of integrating fuzzy logic with genetic algorithms for advanced environmental management.

C. Part 3 – Comparing Different Optimization Techniques on CEC'2005 Functions

In this part, we will compare different optimization techniques using benchmark functions from the CEC'2005 suite. By using these functions, we can see how well each optimization technique performs in finding the best solutions for different types of problems. This comparison will help us understand the strengths and weaknesses of each method.

Generation	Best Fitness
1	-2656.604308390023
2	-2656.604308390023
3	-2656.604308390023
4	-2656.604308390023
5	-2656.604308390023
6	-2656.604308390023
7	-2656.604308390023
8	-2656.604308390023
9	-2656.604308390023
10	-2656.604308390023
11	-2656.604308390023
12	-2656.604308390023
13	-2656.604308390023
14	-2656.604308390023
15	-2656.604308390023
16	-2656.604308390023
17	-2656.604308390023
18	-2656.604308390023
19	-2656.604308390023
20	-2656.604308390023

Bảng VIII
BEST FITNESS VALUES ACROSS 20 GENERATIONS

*1) **Objectives**:*

- **Compare Techniques**: Evaluate genetic algorithms, particle swarm optimization (PSO), and differential evolution (DE) on CEC'2005 benchmark functions to find the best method.
- **Performance Metrics**: Measure convergence speed, solution accuracy, and robustness by tracking:
  - Best solution,
  - Average solution quality,
  - Function evaluations.

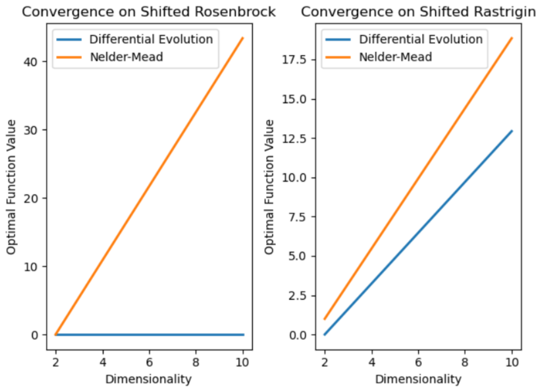
*2) **Comparison of Optimization Techniques on Shifted Benchmark Functions**:* In this section, we compare two optimization methods Differential Evolution and Nelder Mead using two benchmark functions: the Shifted Rosenbrock and Shifted Rastrigin functions. We use NumPy and SciPy to run these optimizations in 2 and 10 dimensions. Each method is set up with specific limits and run for up to 100 iterations. The results are saved in an array that shows the best function values for each method and function combination. This comparison helps us understand how well each technique works in finding solutions in complex problems.

*3) **Performance Comparison of Optimization Methods**:* This section presents plots that compare the performance of two optimization methods—Differential Evolution and Nelder-Mead—on two benchmark functions: Shifted Rosenbrock and Shifted Rastrigin. **Plot Description**

- Each plot displays the best values achieved by these techniques for dimensions 2 and 10. The x-axis represents dimensionality, while the y-axis shows the optimal values obtained by each method.

Expected Output

- The output will display two side-by-side plots, each comparing the performance of Differential Evolution and



Hình 30. Convergence Plots for the Optimization Results.

Nelder-Mead across the specified dimensions.

Observations The plots reveal which technique performs better for each function and how their performance changes with increasing dimensionality. For example, one method may consistently produce lower optimal values, indicating its superior effectiveness for solving the benchmark function.

*4) **Conclusion for Part 3 – Comparing Different Optimization Techniques on CEC'2005 Functions**:* In Part 3, both benchmark functions, Differential Evolution outperforms Nelder-Mead, especially as dimensionality increases. Nelder-Mead struggles to find optimal solutions in higher-dimensional spaces, whereas Differential Evolution maintains lower optimal function values. These results suggest that Differential Evolution is a more robust and efficient technique for solving these types of optimization problems on CEC'2005 functions.

X. Conclusion

In conclusion, the project effectively utilized advanced techniques to predict student performance and develop a Fuzzy Logic Controller (FLC) for smart home environments. The student performance model demonstrated the ability to accurately classify students based on academic factors, while the FLC optimized environmental control for disabled residents, enhancing their comfort and quality of life. Both tasks highlighted the importance of selecting appropriate methodologies to achieve reliable and practical solutions.

References

1) EUROSIS-ETI. (2008, April 1). *Using data mining to predict secondary school student performance.* https://repositorium.sdum.uminho.pt/handle/1822/8024
2) Faria, S.,& Portela, M. (2016). *Student Performance in Mathematics using PISA-2009 data for Portugal.* https://repositorio.ucp.pt/handle/10400.14/21065
3) 3. Silva, M. C., Camanho, A. S., & Barbosa, F. (2019). *Benchmarking of secondary schools based on Students' results in higher education. Omega, 95, 102119.* https://doi.org/10.1016/j.omega.2019.102119
4) Costa, A., & Faria, L. (2014). *The impact of Emotional Intelligence on academic achievement: A longitudinal study in Portuguese secondary school. Learning and Individual Differences, 37, 38–47.* https://doi.org/10.1016/j.lindif.2014.11.011
5) Cheng, L. (2017) *Exploring the Factors that Affect Secondary Student's Mathematics and Portuguese Performance in Portugal. Masters' dissertation Technological University Dublin, 2017.* doi:10.21427/D7P33K
6) Marmeleira, J., Folgado, H., Guardado, I. M., & Batalha, N. (2019). *Grading in Portuguese secondary school physical education: assessment parameters, gender differences and associations with academic achievement. Physical Education and Sport Pedagogy, 25(2), 119–136.* https://doi.org/10.1080/17408989.2019.1692807
7) Sarrico, C. S., Rosa, M. J., & Coelho, I. P. (2010). *The performance of Portuguese secondary schools: an exploratory study. Quality Assurance in Education, 18(4), 286–303.* https://doi.org/10.1108/09684881011079143
8) Amar Shah, A. W. (2019). *Student-t Processes as Alternatives to Gaussian Processes. Proceedings of Machine Learning Research.* doi:33:877-885,2014
9) Dahao Ying, J. M. (2024, Jan 30). *Student Performance Prediction with Regression Approach and Data Generation. Applied Sciences.* https://doi.org/10.3390/app14031148
10) Zexun Chen, B. W. (2019, 12 31). *Multivariate Gaussian and Student-t process regression for multi-output prediction. SpringerLink, 32, 3005–3028*
11) López, S. C., López, M. C., & Guzmán, A. (2018). *Predicting Student Performance with Machine Learning Algorithms: A Case Study. Computers in Human Behavior, 89, 122-133.*
12) Romero, C., & Ventura, S. (2013). *Data Mining in Education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.*
13) UCI Machine Learning Repository. (n.d.).*Student Performance Dataset.* https://archive.ics.uci.edu/dataset/320/student+performance
14) *Gaussian process regression. (n.d.).* https://apmonitor.com/pds/index.php/Main/GaussianProcessRegression
15) *Gaussian processes for classification - Martin Krasser's Blog. (2020, November 4).* https://krasserm.github.io/2020/11/04/gaussian-processes-classification/
16) *Introduction to Bayesian networks | Bayes Server. (n.d.).* https://www.bayesserver.com/docs/introduction/bayesian-networks/
17) Lawton,G. (2022, January 31). *Data preprocessing. Data Management.* https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing
18) *What is Exploratory Data Analysis? | IBM. (n.d.).* https://www.ibm.com/topics/exploratory-data-analysis
19) *Parameter Optimization for Computer Numerical Controlled Machining Using Fuzzy and Game Theory [Accessed 15 Oct 2024]* https://www.researchgate.net/figure/Architecture-of-fuzzy-logic-control_fig1_337511172
20) *What is Fuzzy Logic Controller (FLC) | IGI Global. (n.d.).* https://www.igi-global.com/dictionary/

fuzzy-logic-controller-flc/84470

21) *Maiden application of fuzzy logic based IDD controller for automatic generation control of multi-area hydrothermal system: A preliminary study. (2010, December 1)* https://ieeexplore.ieee.org/document/5710768
22) Solihin, M. I., Chuan, C. Y., & Astuti, W. (2020). *Optimization of fuzzy logic controller parameters using modern meta-heuristic algorithm for gantry crane system (GCS).* https://doi.org/10.1016/j.matpr.2020.05.641
23) *The Application of Genetic Algorithm into Membership Function FLC Used Floating Point. (2009).* https://ieeexplore.ieee.org/document/5392892

A. Github-Link

https://github.com/RijalBijay/AML_Group.git