ALBUKHARY INTERNATIONAL UNIVERSITY

ALBUKHARY INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTING AND INFORMATICS

COURSE DETAILS	
SCHOOL	SCHOOL OF COMPUTING AND INFORMATICS
COURSE NAME	Machine Learning
COURSE CODE	(CCS2213)
LECTURER	**Prof Dr. Zurinahni binti Zainol**
SEMESTER & YEAR	SEMESTER 2, YEAR 2
ACTIVITY	Compare supervised and unsupervised
STUDENT NAME	Areyan Muhemed Ali Hussein
ID	AIU21102179

1. Introduction

Background and Motivation

Predicting student absenteeism and academic success can significantly impact educational institutions by enabling them to take proactive measures. Institutions can identify at risk students early and implement interventions to improving retention and support academic achievement, For instance understanding the patterns and factors contributing to absenteeism can helps schools design targeted programs to engage students better and address issues before they escalate into significant problems. This not only improving individual student outcomes but also enhances the overall educational environment.

Similarly, understanding weather patterns can provides insights into climate behavior, aiding in various applications such as agriculture, disaster management and environmental planning. By analyzing historical weather data, we can detect patterns and trends that are critical for forecasting and decision making in climate-sensitive activities. For example, farmers can use weather predictions to plan their planting and harvesting schedules, optimizing crop yield and reducing losses due to unexpected weather changes. In disaster management, accurate weather forecasts and pattern recognition can improving preparedness and response strategies potentially saving lives and reducing property damage.

2. Project Idea

This project involves two main components:

Supervised Learning This component compares various supervised learning models to predict student absenteeism. The goals is to identify the most accurate model and understand the factors contributing to absenteeism. By use historical data on student attendance, demographics and academic performance, we aim to build models that highlight key indicators of absenteeism, helping institutions develop strategies to improve student engagement.

The process includes data collection and preprocessing (handling missing values, encoding categorical variables, and standardizing features), followed by implementing algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, and Decision Tree. Models are evaluated using metrics like accuracy, precision, recall, and F1-score to determine the best performer.

Unsupervised Learning This component applies K-Means clustering to weather history data to identify patterns and categorize different weather phenomena. Clustering reveals natural groupings, providing insights into climate behavior.

The process involves data collection and preprocessing (normalizing data and handling missing values), followed by K-Means clustering to partition the data based on similarities. The number of clusters is determined using the Elbow Method or Silhouette Score. Clusters are analyzed to identify distinct weather patterns, using visualizations and centroid interpretation.

By integrating both techniques, this project provides a comprehensive analysis, leveraging each approach's strengths to extract insights and improve decision-making in educational and environmental contexts.

Objectives

- **Supervised Learning:**

Predict student absenteeism use various models.

Identify the most accurate model and key absenteeism factors.

- **Unsupervised Learning:**

Cluster weather data to reveal patterns.

Understand and categorize different weather phenomena.

By integrating both approaches the project aims to extract meaningful insights and enhance decision making in education and environmental contexts.

3. Literature Review

Extensive research has been conducted on predicting student outcomes using machine learning. Techniques such as logistic regression, decision trees, SVM and KNN have been widely used to predict dropout rates and academic performance, These studies highlight the importance of feature selection, data preprocessing and the choice of algorithm in achieving accurate predictions. For example, factors such as attendance, past academic performance, and socio economic background have been identified significant predictors of student success.

In the domain of climate studies clustering techniques like K-Means have been applied to categorize weather patterns and identify trends. These applications range from agricultural planning, where understanding seasonal weather patterns can optimize crop management, to disaster preparedness, where recognizing patterns in historical weather data can enhance forecasting and response strategies. The effectiveness of these techniques often depend on the quality of the data and the appropriateness of the chosen clustering method.

4. Methodology

The project is divided into two main parts supervised learning and unsupervised learning. In the supervised learning part multiple supervised learning algorithms is implemented on the absenteeism dataset to predict student absenteeism. In the unsupervised learning part, K-Means clustering is applied to the weather history dataset to uncover distinct weather patterns.

The absenteeism dataset contains features related to student demographics, academic performance, and attendance records, with the target variable indicating absenteeism. Features may include age, gender, academic scores, attendance rates and socio economic indicators. The weather history dataset contains historical weather data with features such as temperature, humidity, precipitation, wind speed and atmospheric pressure

For the absenteeism dataset, missing values were handled by either dropping incomplete record or using imputation techniques. Features were standardized to ensure consistency during model

training, and categorical variables were encoded using techniques such as one hot encoding. For the weather history dataset missing values were either dropped or imputed. Numeric features were normalized to ensure they contribute equally to the clustering process, and any outliers were identified and addressed to prevent skewing the results.

Sporvisor learning data processing

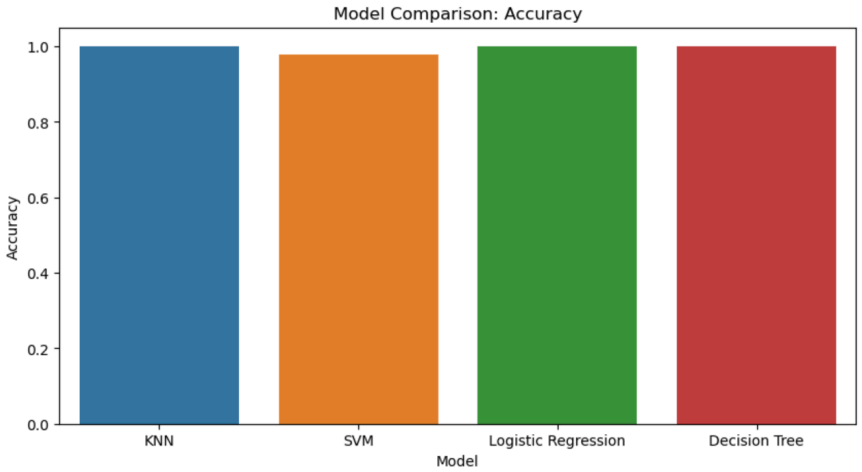
```
  ethnicity sex age learner_status  days_absent
0         A   M  F0              SL            2
1         A   M  F0              SL           11
2         A   M  F0              SL           14
3         A   M  F0              AL            5
4         A   M  F0              AL            5
Index(['ethnicity', 'sex', 'age', 'learner_status', 'days_absent'], dtype='object')
ethnicity         0
sex               0
age               0
learner_status    0
days_absent       0
dtype: int64
```

Unsupervised data processing

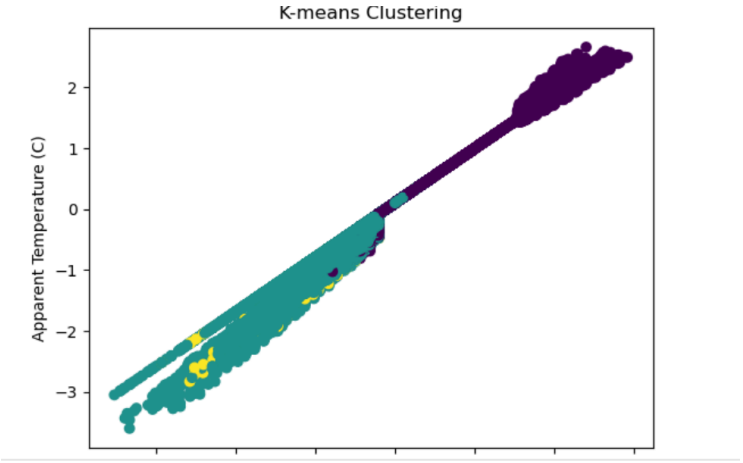
	Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Speed (km/h)	Bearing (degrees)	Visibility (km)	Loud Cover	Pressure (millibars)	Daily Summary
0	2006-04-01 00:00:00.000 +0200	Partly Cloudy	rain	9.472222	7.388889	0.89	14.1197	251	15.8263	0	1015.13	Partly cloudy throughout the day.
1	2006-04-01 01:00:00.000 +0200	Partly Cloudy	rain	9.355556	7.227778	0.86	14.2646	259	15.8263	0	1015.63	Partly cloudy throughout the day.
2	2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204	14.9569	0	1015.94	Partly cloudy throughout the day.
3	2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269	15.8263	0	1016.41	Partly cloudy throughout the day.
4	2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259	15.8263	0	1016.51	Partly cloudy throughout the day.

In the absenteeism dataset relevant features were selected based on their correlation with the target variable to ensure the models is trained on the most informative data. For the weather history dataset, features were selected based on their variance and importance to ensure meaningful clustering.

Supervised feature



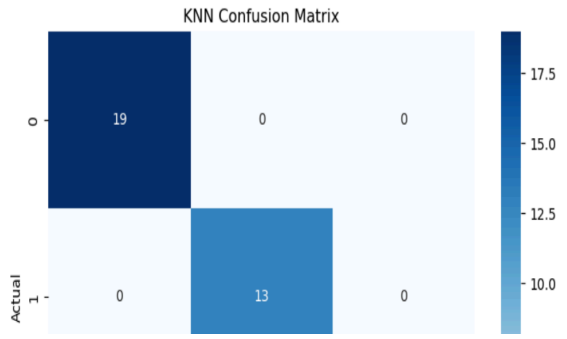
Unsupervised feature



5. Supervised Learning

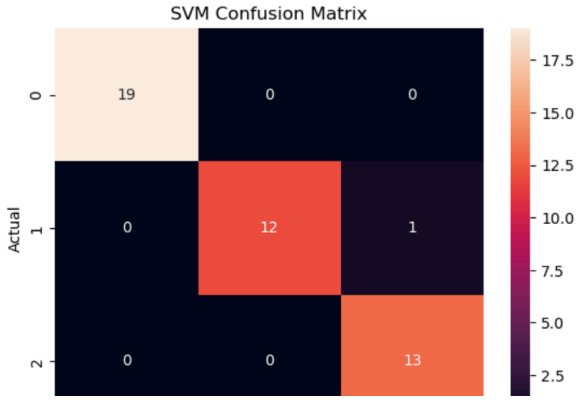
Algorithms Overview

K-Nearest Neighbors (KNN)** is an instance-based learning algorithm that classifies a data point based on its nearest neighbors.

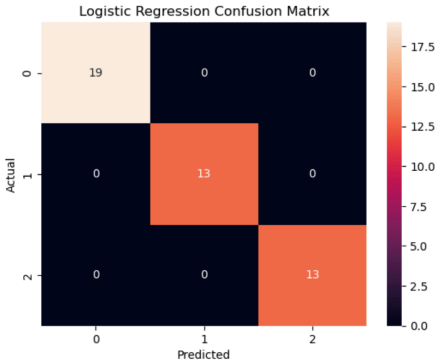


Support Vector Machine (SVM) is classifier that finds the optimal hyperplane to separate data into classes.

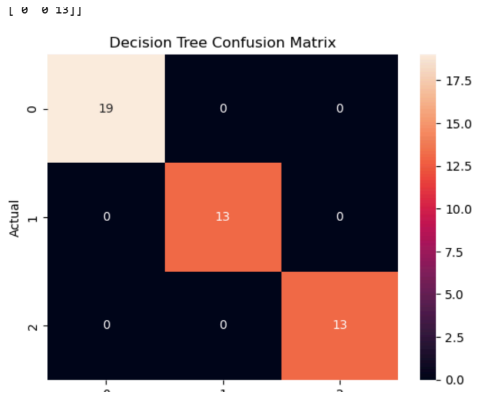
```
SVM Accuracy: 0.9777777777777777
SVM Confusion Matrix:
[[19  0  0]
 [ 0 12  1]
 [ 0  0 13]]
```



Logistic Regressionis linear model that predicts the probability of a binary outcome using a logistic function.



Decision Tree is model that splits data into subsets based on feature values creating tree structure for decision making.



Implementation

The algorithms are implement using the absenteeism dataset. The process involves splitting the datas into training and testing sets training each model on the training data, and making predictions on the test data.

Evaluation Metrics

The models are evaluated using the following metrics

Accuracy The ratio of correctly predicted instances to the total instances.

Precision The ratio of correctly predicted positive observations to the total predicted positives.

Recall The ratio of correctly predicted positive observations to all observations in the actual class.

F1-Score The weighted average of Precision and Recall.

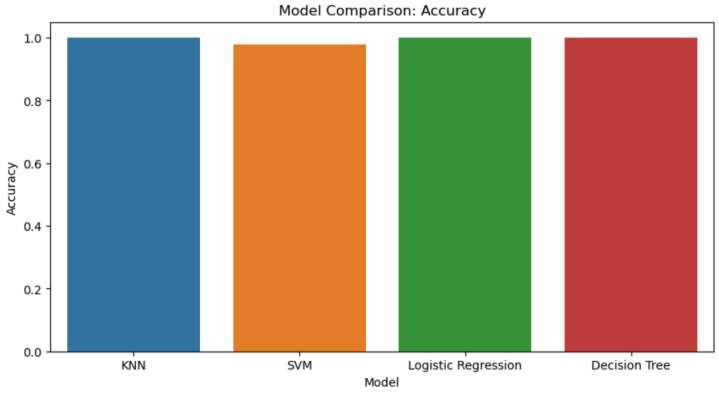
Confusion Matrix A table to describe the performance of classification model by showing the actual versus predicted classifications.

These metrics provide a comprehensive assessment of each models performance, helping to identify the most effective model for predicting students absenteeism.

6. Results and Analysis

The performance of K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, and Decision Tree models was compared based on their accuracy, precision, recall, and F1-score. Accuracy indicates the overall correctness of the model precision measures the correct positive predictions among total predicted positives, recall evaluates the correct positive predictions among actual positives, and F1-score provides the harmonic mean of precision and recall.

Visualizations included confusion matrices to show the distribution of predictions bar plots for comparing model accuracies and ROC curves to evaluate the trade off between true positive and false positive rates for each model. These visualizations help in understanding the strengths and weaknesses of each model in predicting student absenteeism.



7. Unsupervised Learning

Algorithms Overview

K-Means Clustering partitions data into K clusters based on feature similarities. It iteratively assigns data points to clusters and updates the cluster centroids to minimize variance within clusters.

Implementation

The weather history dataset was preprocessed by normalizing numeric features, K-Means clustering was then applied and the optimal numbers of clusters was determined use the Elbow Method and Silhouette Score, Clusters were analyzed to identify distinct weather patterns.

```
Formatted Date               0
Summary                      0
Precip Type                517
Temperature (C)              0
Apparent Temperature (C)     0
Humidity                     0
Wind Speed (km/h)            0
Wind Bearing (degrees)       0
Visibility (km)              0
Loud Cover                   0
Pressure (millibars)         0
Daily Summary                0
dtype: int64
```

Evaluation Metrics

The quality of the clusters was evaluated using the Silhouette Score which measures how similar a data point is to its own cluster compared to other clusters.

	Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibility (km)	Loud Cover	Pressure (millibars)	Daily Summary	Cluster
0	2006-04-01 00:00:00.000 +0200	Partly Cloudy	rain	-0.257951	-0.324102	0.792748	0.478964	0.591157	1.309107	0.0	0.102152	Partly cloudy throughout the day.	1
1	2006-04-01 01:00:00.000 +0200	Partly Cloudy	rain	-0.270141	-0.339134	0.639470	0.499902	0.665655	1.309107	0.0	0.106415	Partly cloudy throughout the day.	1
2	2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	-0.267819	-0.138532	0.792748	-0.993620	0.153478	1.100806	0.0	0.109058	Partly cloudy throughout the day.	1
3	2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	-0.381594	-0.458873	0.486192	0.476638	0.758778	1.309107	0.0	0.113066	Partly cloudy throughout the day.	1
4	2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	-0.332833	-0.362460	0.486192	0.034630	0.665655	1.309107	0.0	0.113919	Partly cloudy throughout the day.	1

8. Project Goals

The main goal were to compares the performance of supervised learning models in predicting student absenteeism and to identify patterns in weather data use unsupervised learning. The project aimed to extract meaningful insights to enhance decision-making in educational and environmental contexts

Comparative Analysis

Supervised learning models were compared based on their accuracy, precision, recall and F1score. K-Means clustering ware used to categorize weather patterns and the results were analyzed to understand the effectiveness of the clustering.

Insights into Effects of Different Parameters

The analysis provided insights into how different features and parameters influenced model performances and clustering quality, For example, the number of neighbors in KNN or the number of clusters in K-Means significantly affected the results.

Comprehensive Analysis of Findings

The comprehensive analysis combined the findings from both supervised and unsupervised learning, supervised models helpe identify key factors contributing to student absenteeism, while clustering revealed meaningful weather patterns.

9. Conclusions

The comparative analysis revealed that Best Model was the most accurate for predicting student absenteeism. K-Means clustering effectively categorized weather patterns providing valuable insights into climate behavior.

The results suggest that Best Model can be using for early identification of at risk students, enabling timely interventions, The clustering analysis can help in agricultural planning and disaster preparedness by understanding weather patterns.

Future research could explore advanced models like neural network and ensemble methods to improving prediction accuracy and clustering quality. Additionally integrating more features and larger datasets could provide deeper insights into student absenteeism and weather patterns.

10. References

1. *Machine Learning for Education*:

- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

- Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-17.

2. *Supervised Learning Algorithms*:

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

- Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (Vol. 398). John Wiley & Sons.

3. *Unsupervised Learning and Clustering*:

- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

4. *Data Preprocessing and Feature Selection*:

- Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier.

- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157-1182.

5. *Evaluation Metrics*:

- Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

- Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

6. *Application of Machine Learning in Education and Climate*:

- Asif, R., Merceron, A., & Pathan, M. K. (2015). Predicting student academic performance at degree level: A case study. International Journal of Intelligent Systems and Applications, 7(1), 49.

- Jiang, Z., & Zheng, C. (2007). Design of a weather forecast system based on data mining. In 2007 International Conference on Wireless Communications, Networking and Mobile Computing (pp. 3224-3227). IEEE.