IBM SkillBuild
AI\ML - FINAL PROJECT REPORT
Prepared by : J.Mohana Krishna
mohanakrishnajillella@gmail.com

# Machine Learning for Sustainable Development Goal 6: Insights of virtual reality influence on Education (Quality Education)

## 1. Introduction

Project Objective: To use machine learning to address challenges in education aiming to support SDG 4 by predicting the insights of virtual reality Influence on Educational Engagement Metrics.

Motivation: The motivation behind studying the influence of virtual reality (VR) on educational engagement metrics lies in VR's potential to revolutionize learning experiences. VR can create immersive, interactive environments that captivate students' attention, foster active learning, and promote deeper understanding. The aim is to explore how VR can enhance motivation, participation, and retention in educational settings, leading to more effective learning outcomes compared to traditional or less interactive digital tools. Evaluating these metrics helps educators and institutions determine the value and potential scalability of VR as a transformative educational technology.

## 2. Data Collection

Data Source: Kaggle Dataset

Dataset Description:
- Features: This dataset includes the following features such as Age, Hours_of_VR_Usage_Per_Week, Perceived_Effectiveness_of_VR, Engagement_Level

## 3. Exploratory Data Analysis (EDA)

Summary Statistics: Mean, median, and distribution of each feature.
Visualizations:
- Correlation heatmap to understand relationships between variables.
- Boxplots for outlier detection.
- Histograms to assess the distribution of each variable.

Insights: Key trends or anomalies in pH levels, hardness, or contamination levels.

## 4. Data Preprocessing

### Handling Missing Values: Used median imputation for features with missing values.

Encoding Categorical Variables: One-hot encoding for any categorical features.
Feature Scaling: Standardized features using `StandardScaler` for better performance in machine learning models.

## 5. Machine Learning Model Selection

Model Choices:
- Logistic Regression (for binary classification).
- Random Forest Classifier (for handling non-linear relationships and feature importance).
- Support Vector Machine (SVM) for optimal margin separation.

Why Scikit-Learn: Easy implementation, variety of algorithms, and effective performance metrics.
Evaluation Metric: Accuracy, Precision, Recall, and F1-Score due to the critical nature of accurately identifying contamination.

## 6. Model Implementation

Data Splitting: Split dataset into 80% training and 20% testing sets using `train_test_split` from Scikit-Learn.

***Code Example:***

```
# Import necessary libraries

import pandas as pd

import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_squared_error, r2_score

import joblib

# Load the dataset

data = pd.read_csv('Virtual_Reality_in_Education_Impact.csv')

# Select features and target

features = ['Age', 'Hours_of_VR_Usage_Per_Week', 'Perceived_Effectiveness_of_VR']

target = 'Engagement_Level'

# Drop rows with missing values in the selected columns

data_cleaned = data.dropna(subset=features + [target])
```

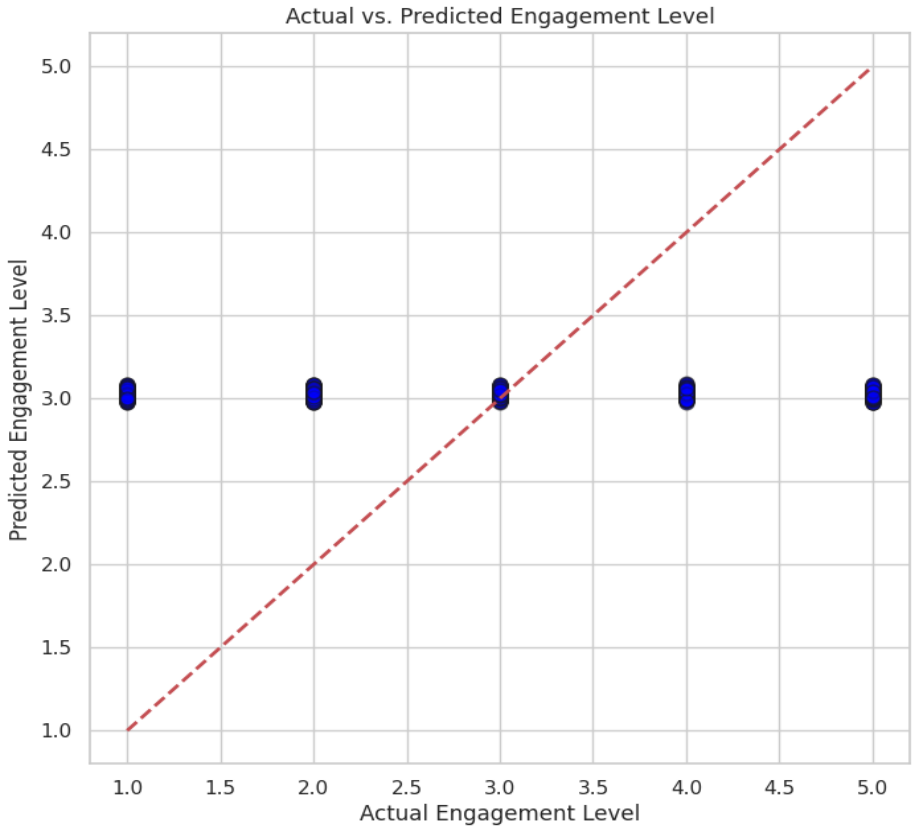
```
# Separate features and target variable
X = data_cleaned[features]
y = data_cleaned[target]

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Initialize the linear regression model
model = LinearRegression()
# Train the model
model.fit(X_train, y_train)
# Save the model
joblib.dump(model, 'vr_engagement_model.pkl')
# Make predictions on the test set
y_pred = model.predict(X_test)
# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)
# Display the performance metrics
print("Model Performance:")
print(f"Root Mean Squared Error (RMSE): {rmse}")
print(f"R-squared (R2): {r2}")
```

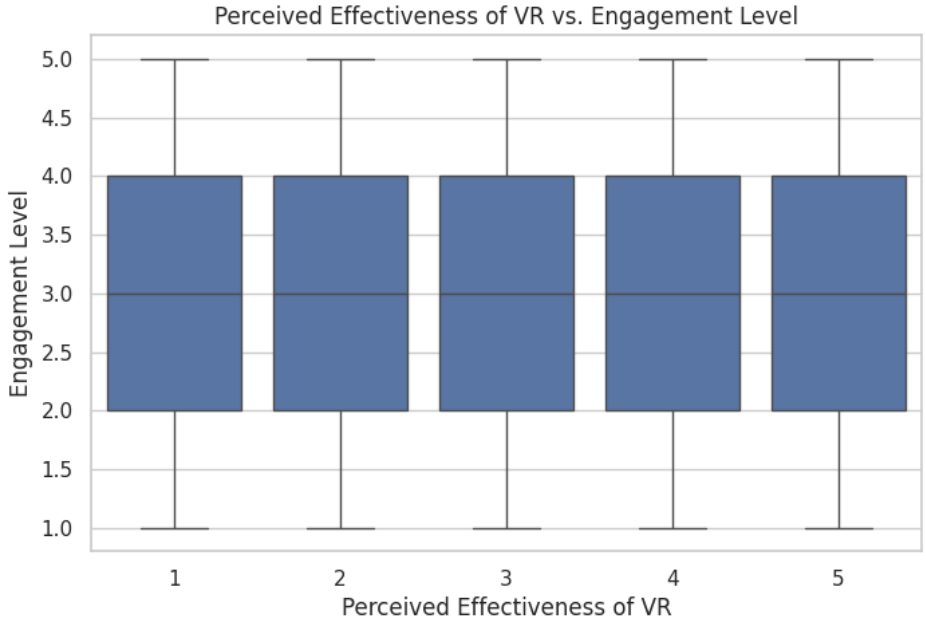
## 7. Results and Evaluation

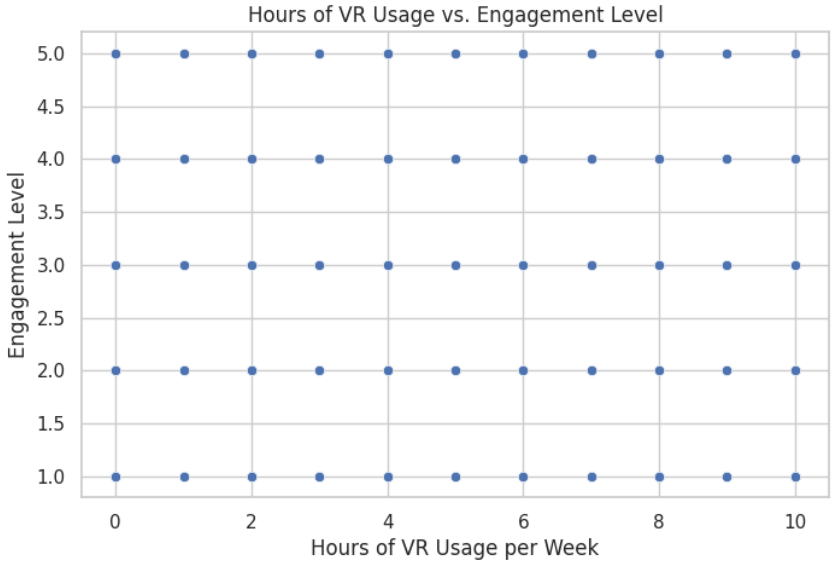
Model Performance:
Root Mean Squared Error (RMSE): 1.4304850911738651

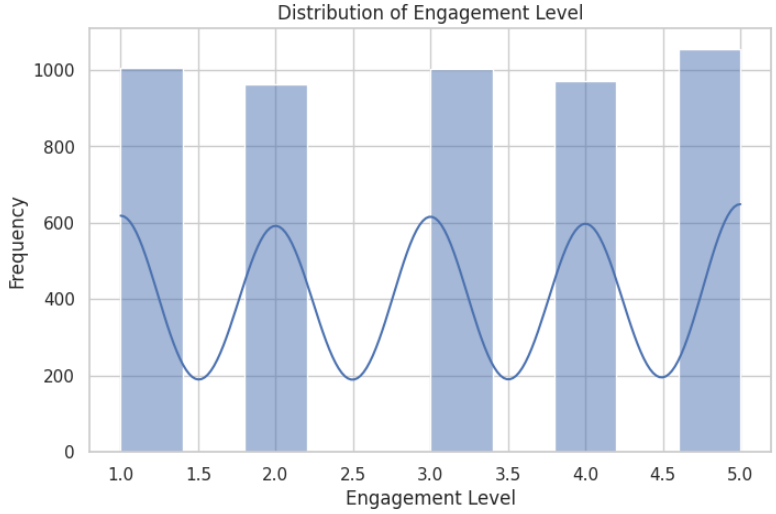
R-squared (R2): -0.001868617018651042

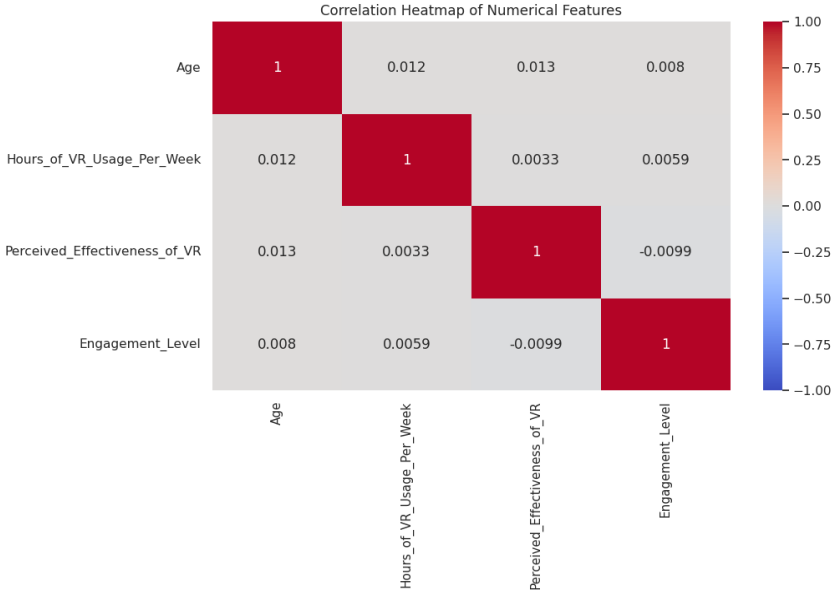
RESULTS:


Actual vs. Predicted Engagement Level
Predicted Engagement Level
Actual Engagement Level
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Perceived Effectiveness of VR vs. Engagement Level
Engagement Level
Perceived Effectiveness of VR
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1
2
3
4
5


Hours of VR Usage vs. Engagement Level
Engagement Level
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
2
4
6
8
10
Hours of VR Usage per Week


Distribution of Engagement Level
Frequency
1000
800
600
400
200
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Engagement Level


Correlation Heatmap of Numerical Features
Age
Hours_of_VR_Usage_Per_Week
Perceived_Effectiveness_of_VR
Engagement_Level
1
0.012
0.013
0.008
0.012
1
0.0033
0.0059
0.013
0.0033
1
-0.0099
0.008
0.0059
-0.0099
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

## 8. Conclusion and Future Work

1.Data Collection Collect data related to impact on education from different sources.

2.Data Analysis Perform exploratory data analysis (EDA) to identify trends, patterns, and anomalies in the dataset.

3.Data Preprocessing Clean the dataset by handling missing values and outliers. Normalize the features to ensure consistent data input for model training.

4.Training Model Select an appropriate machine learning algorithm (e.g., Logistic Regression or Decision Trees). Split the dataset into training and validation sets. Monitor for overfitting and apply regularization techniques (L1 or L2) to prevent it. Train the model on the prepared dataset.

5.Model Evaluation Assess model performance using evaluation metrics like accuracy, precision, recall, and F1 score. Conduct crossvalidation to ensure the model generalizes well to unseen data.

## 9. References

- Kaggle Dataset
- Scikit-Learn Documentation