# Results for Assnignment 1

## Linear regression

The model performs best when the dependent variable is **watch time**, and the independent variable is **confusion points**. Using a linear regression model, the results show that the **degree of fit (R²)** is **0.82**, indicating that the model explains **82%** of the variance in watch time based on confusion points. The **coefficient** for the independent variable (confusion points) is **2.10522396**, meaning that for every additional confusion point, the predicted watch time increases by approximately **2.1 units**. The function of this model is:

Y= 2.10522396x−0.01662473

The **R² value of 0.82** indicates a strong fit between the data and the model, suggesting that confusion points are a significant predictor of watch time. The **positive coefficient** indicates a direct relationship between confusion points and watch time: as confusion points increase, so does the watch time.

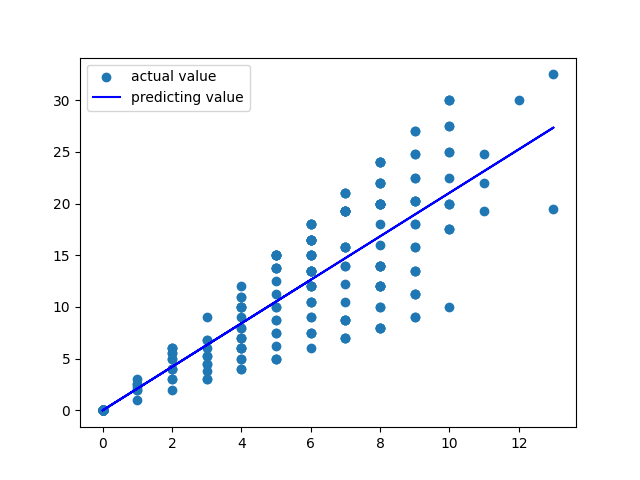


Fig.1 Predicting results of linear regression model

## Multiple regression

The model performs best when the dependent variable is **watch time**, and the independent variables are **participation**, **confusion points**, and **key points**. Using a multiple regression model, the results show that the **degree of fit (R²)** is **0.82**, indicating that the model explains **82%** of the variance in watch time based on these three independent variables.

The **coefficients** for the independent variables are:

* Participation: **0.55757518**
* Confusion points: **2.08731237**
* Key points: **-0.15120358**

The **intercept** of the model is approximately zero (**3.55271368e-15**), meaning that when all independent variables are zero, the predicted watch time is effectively zero due to numerical precision. The function of this model is:

Y=0.55757518⋅x1+2.08731237⋅x2−0.15120358⋅x3

**Model Fit and Interpretation:**

The **R² value of 0.82** indicates a strong fit between the data and the model, suggesting that these three variables—**participation**, **confusion points**, and **key points**—together explain a significant portion of the variation in watch time.

* The **positive coefficient** for **participation** (0.5576) suggests that higher participation levels are associated with increased watch time.
* The **confusion points coefficient** (2.0873) has the strongest influence, showing that as confusion points increase, watch time increases substantially, by approximately 2.1 units per confusion point.
* The **negative coefficient** for **key points** (-0.1512) suggests that higher key points lead to a slight decrease in watch time, holding other variables constant.

# Results for Assnignment 2

## Logistic Regression

Three independent variables: forum.posts, grade, and assignment were used to predict whether a user is certified (yes or no) using a **logistic regression model**.

**Here’re the results of Model Performance: Recall**: 1.000**, Precision**: 1.000**, F1 Score**: 1.000**, Accuracy**: 1.000.The performance metrics below shows that the model achieves perfect scores in recall, precision, F1 score, and accuracy, indicating that the model predicts both positive (certified = 1) and negative (certified = 0) outcomes with 100% accuracy. The confusion matrix confirms this, as it shows **no misclassifications**: 82 instances where the actual value is 0 (not certified) are correctly predicted as 0, and 218 instances where the actual value is 1 (certified) are correctly predicted as 1. The dataset has **725 certified** samples and **275 non-certified** samples, indicating a somewhat imbalanced distribution but not heavily skewed.

图表

描述已自动生成

Fig 2. Confusion Matrix for logistic regression model

The correlation matrix shows a moderate positive correlation between all features:forum.posts and grade: 0.666**,** forum.posts and assignment: 0.682**,** grade and assignment: 0.691**.** These correlations suggest that there is some relationship between the independent variables, but they are not perfectly collinear, meaning the model is not likely to suffer from severe multicollinearity.

While the model shows perfect performance on the test data, which is highly unusual in practical scenarios, it may indicate potential overfitting or a very simple relationship between the input features and the output label.

## Decision Tree

A **Decision Tree Classifier** was used to predict the certification status (certified: yes = 1, no = 0) based on the following independent variables: forum.posts, grade, and assignment.

**Here’re the results of Model Performance: Recall**: 1.000**, Precision**: 1.000**, F1 Score**: 1.000**, Accuracy**: 1.000.These performance metrics indicate that the model is perfectly predicting both classes (certified and non-certified) in the test dataset. This is confirmed by the confusion matrix below, which shows:82 true negatives (correctly predicting non-certified as 0) and 218 true positives (correctly predicting certified as 1).No false positives or false negatives, meaning the model made no mistakes in classification.The data has **725 certified** samples and **275 non-certified** samples, which shows a somewhat imbalanced dataset, though not severely skewed.

**图表

描述已自动生成**

Fig 3. Confusion Matrix for decision tree model

The correlation matrix reveals moderate correlations between the input variables: forum.posts and grade: 0.666, forum.posts and assignment: 0.682, grade and assignment: 0.691.While there is some relationship between the independent variables, it is not strong enough to cause multicollinearity issues.

To sum up, The Decision Tree Classifier is showing perfect performance on the test dataset, which might suggest one of the following: The features (forum.posts, grade, and assignment) are extremely strong predictors of certification. The model could be overfitting to the training data, capturing noise rather than general trends. This is particularly likely because decision trees tend to overfit unless controlled via hyperparameters like max\_depth or min\_samples\_split.

## Naive Bayes

The **Naive Bayes Classifier** (Gaussian Naive Bayes for continuous data) was used to predict the certification status (certified: yes = 1, no = 0) based on the independent variables: forum.posts, grade, and assignment.

The following data tells the result of the model: **Recall**: 1.000, **Precision**: 1.000, **F1 Score**: 1.000, **Accuracy**: 1.000. These results indicate that the Naive Bayes model achieved perfect performance in terms of classification, as it correctly predicted every instance in the test set. Specifically:82 true negatives (correctly predicting non-certified as 0), and 218 true positives (correctly predicting certified as 1), with no false positives or false negatives were observed, indicating a flawless classification.The dataset consists of **725 certified** and **275 non-certified** samples, which indicates a relatively balanced class distribution, though the certified class is somewhat more frequent. The correlation matrix (fig. below) between the independent variables (forum.posts, grade, assignment) shows moderate correlations: forum.posts and grade: 0.666, forum.posts and assignment: 0.682, grade and assignment: 0.691.Although some moderate correlations exist, they are not high enough to suggest multicollinearity issues that could affect the model’s performance.

图表

描述已自动生成

Fig 4. Confusion Matrix for naïve bayes model

To sum up, The Naive Bayes Classifier has achieved perfect scores in this instance. This exceptional performance can suggest a strong relationship between the independent variables (forum.posts, grade, and assignment) and the target variable (certified). However, just like with other models showing perfect scores, this result may suggest potential overfitting, particularly in cases where the dataset could be simpler or less noisy than expected.

# Self-reflections

For this assignment, I would rate myself 4.5/5. I deducted 0.5 points because, despite my best efforts, I still struggled to grasp the mathematical principles underlying each machine learning model. This left me feeling somewhat confused and unsure about choosing the best model for different datasets. However, I am pleased that I wrote all the code independently, utilizing the documentation from the scikit-learn library, online video tutorials, and hints provided in the assignment introduction. One particular aspect that remains perplexing is the classification assignment: regardless of how small or large the C (regularization parameter) value is, every model consistently scored 100/100. This seems implausible to me. In addressing this concern, I consulted ChatGPT, which provided the following insights:

* *Is there an issue with the dataset?*

*The dataset might contain features that lead to overfitting, or the relationship between the labels (the certified column) and the features (forum.posts, grade, assignment) might be too simple, making it easy for the model to separate the data.*

* *Is there collinearity between the features?*

*Your independent variables (forum.posts, grade, assignment) may be highly correlated, which could make the model's decision too straightforward. You can check the correlation between these features using pd.DataFrame.corr().*

* *Class imbalance problem*

*If the certified classes are heavily imbalanced (e.g., most samples are either "yes" or "no"), the model might simply predict the majority class, leading to seemingly "perfect" performance. You can examine the data distribution by using data['certified'].value\_counts().*

* *Randomness and overfitting issues*

*Using a relatively large value of C=100.0 could cause overfitting, as this represents the inverse of regularization strength. You can try reducing the value of C to increase regularization and prevent the model from overfitting the training set.*

Despite these explanations, I am still left confused. This experience has highlighted the importance of understanding the theoretical foundations of machine learning models. I realize that while coding is an essential skill, a deeper comprehension of the underlying mathematical concepts is crucial for effectively selecting and implementing models. Moving forward, I plan to dedicate more time to studying these principles and seeking clarification on complex topics.