NAME- Udit Sharma

24BCE11124

CSE-1021

C11-C12-C13

REGRESSION

Regression is a statistical technique used to examine the relationship between a dependent variable (also known as the outcome or target variable) and one or more independent variables (predictors). The goal is to model this relationship to predict the dependent variable's value based on the independent variables.

**Key Types of Regression:**

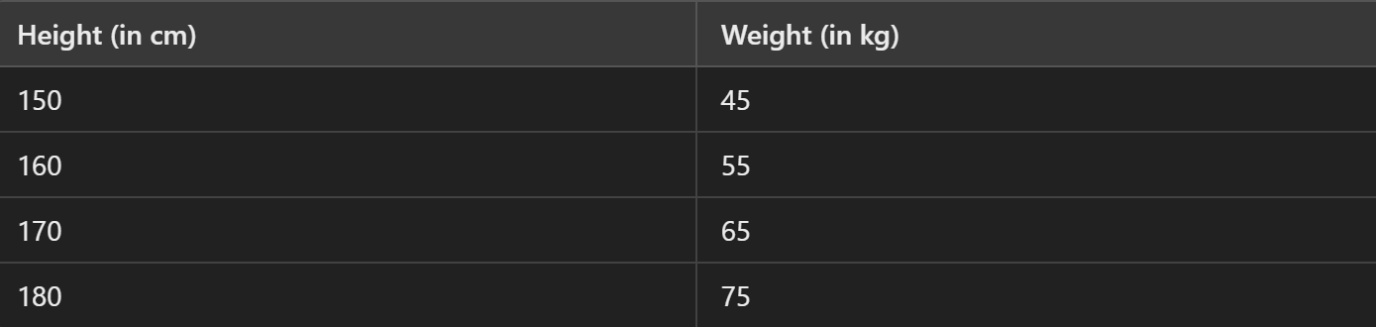
1. **Linear Regression**: Models a straight-line relationship between variables.
2. **Logistic Regression**: Used when the dependent variable is binary (e.g., yes/no).
3. **Polynomial Regression**: Fits a polynomial equation to the data for non-linear relationships.
4. **Multiple Regression**: Involves two or more independent variables to predict the dependent variable.

**Applications:**

* Forecasting (e.g., sales predictions)
* Risk assessment (e.g., credit scoring)
* Understanding relationships (e.g., impact of advertising on sales)

It helps answer questions like which factors significantly impact the outcome and how they interact

Linear Regression

Linear regression is a method used to model the relationship between a dependent variable (output) and one or more independent variables (inputs). The goal is to find the best-fit line that minimizes the difference between predicted and actual values.

Linear regression will calculate the best-fit line, which could be represented by the equation:  
**Weight = 0.5 × Height - 30**

Using this equation:

* For a person 175 cm tall, the predicted weight = **0.5 × 175 - 30 = 57.5 kg**

Linear regression in computers refers to using algorithms to establish a linear relationship between a dependent variable (output) and one or more independent variables (inputs). It is a fundamental machine learning technique, particularly in supervised learning, where the goal is predictive analysis.

**Key Features:**

1. **Predictive Modeling**: Linear regression helps predict an outcome by learning a mapping function based on input features [[2](https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/)].
2. **Simple Interpretation**: The model results in a linear equation, making it easy to interpret and use for forecasting [[5](https://www.ibm.com/topics/linear-regression)].
3. **Training**: The algorithm minimizes the error between actual and predicted values by adjusting the coefficients (weights) through methods like ordinary least squares [[6](https://www.mastersindatascience.org/learning/machine-learning-algorithms/linear-regression/)].

**Example:**

Consider an application predicting house prices based on size. The computer learns from a dataset where both size (input) and price (output) are provided, forming a linear relationship to predict new house prices accurately

Logistic regression

logistic regression is a statistical method used in machine learning to model the probability of a variables. Unlike linear regression, which predicts a continuous value, logistic regression predicts the likelihood of a categorical dependent variable [[1](https://aws.amazon.com/what-is/logistic-regression/)].

**Key Features:**

1. **Classification Tasks**: Logistic regression is primarily used for binary classification problems. For example, determining whether an email is spam or not spam [[4](https://www.geeksforgeeks.org/understanding-logistic-regression/)].
2. **Logistic Function**: It uses the logistic (sigmoid) function to map predicted values to probabilities between 0 and 1 [[3](https://www.ibm.com/topics/logistic-regression)].
3. **Probabilistic Interpretation**: It provides probabilities for each class, which can be used for decision-making or further analysis [[6](https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis)].

**Example:**

In predicting whether a customer will purchase a product (1) or not (0) based on their browsing history, logistic regression assigns probabilities and decides the outcome based on a threshold (commonly 0.5)

Logistic regression is widely used in various computer applications for binary classification tasks. Here's a common example:

**Email Spam Detection:**

In spam detection systems, logistic regression can be used to classify emails as either "spam" or "not spam." The model analyzes features such as the presence of certain keywords, the frequency of links, or the length of the email. Each feature is given a weight, and the logistic regression model calculates the probability of an email being spam. If this probability exceeds a predefined threshold (e.g., 0.5), the email is classified as spam [[2](https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/)].

Another example is **predicting customer behavior in an online store**. Logistic regression can determine factors influencing a customer's likelihood to make a purchase based on past behavior, such as time spent on the website, number of items viewed, and previous purchases

The K-Nearest Neighbors (KNN) algorithm is a simple, supervised machine learning method used for classification and regression tasks. It works by storing all available data and classifying new data points based on their similarity to the existing data. The "k" in KNN represents the number of nearest neighbors considered to make a prediction. For example, in a classification task, the algorithm assigns a class to a new data point by looking at the most common class among its k nearest neighbors [[1](https://www.geeksforgeeks.org/k-nearest-neighbours/)].

KNN is non-parametric, meaning it makes no assumptions about the underlying data distribution. It uses distance metrics (like Euclidean distance) to identify neighbors. Despite its simplicity, KNN is powerful for tasks like image recognition and recommendation systems but can be computationally expensive with large datasets [

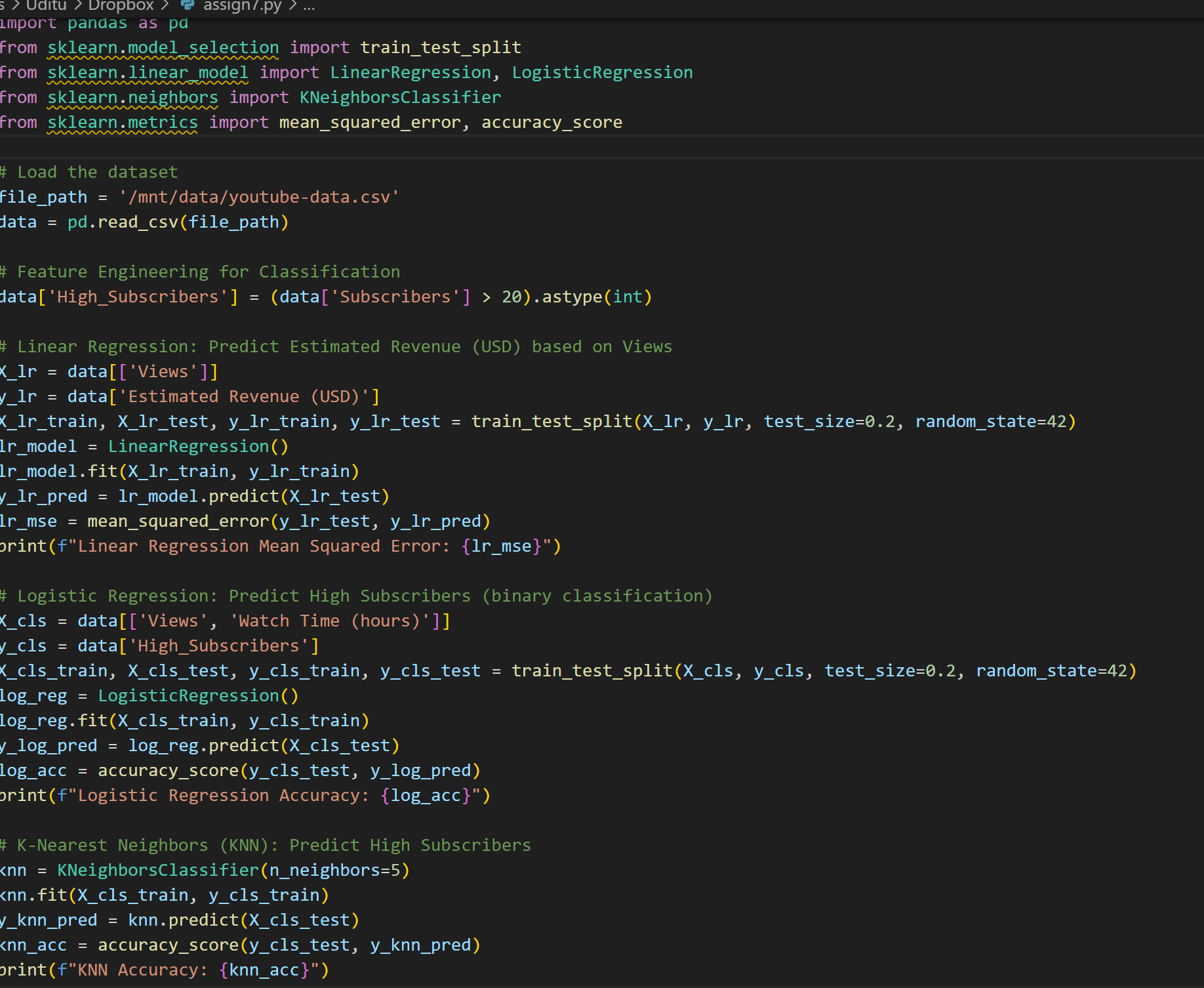
The dataset has multiple features and targets. Let's focus on a simplified subset for the implementations:

1. **Linear Regression**: Predicting Estimated Revenue (USD) based on Views.
2. **Logistic Regression**: Creating a binary classification target by thresholding Subscribers (e.g., Subscribers > 20 = 1, else 0).
3. **K-Nearest Neighbors (KNN)**: Using the same binary target (Subscribers > 20) with Views and Watch Time (hours) as features.

I'll prepare the data and write the Python implementations. ​​

The models have been trained and evaluated on the given dataset. Here are the results:

1. **Linear Regression**: Mean Squared Error (MSE) = **80.67**
2.  **Logistic Regression**: Accuracy = **76.71%**
3.  **K-Nearest Neighbors (KNN)**: Accuracy = **93.15%**



OUTPUT-

