**Unit 10 Assignment 1**

Laurence T. Burden

Purdue Global University

IN300: Programming for Data Analysis

John Brooks

June 29, 2021

**Python Code and Results**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**R Code and Results**

**Text, chat or text message

Description automatically generated**

**Java Code and Results**

**Text

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Compare and Contrast Languages**

There are multiple ways to analyze large amounts of data. Java required a good amount of boilerplate code to make this project work. It is not out of the ordinary and might be alleviated with the right library. The other two languages used libraries to make this process easier.

R was another simple and quick program to write. A total of 3 executable lines of code. Using a library that allows for SQL statements also helped me understand what was being accomplished.

Python gave me the biggest issue this week. The suggested method was not working and after many attempts to troubleshoot, I gave up on this method and reworked the problem. This allowed me to learn about working with tuples versus lists in Python. Pandas will return a tuple even when using the “tolist()” method. Python is still enjoyable to write and the documentation was easy to understand.

The resulting output for this week’s project is a good example of real-world data analysis. Automating security checks is necessary when dealing with tens of thousands of data items. A human simply doesn’t have the time and would be prone to error after growing bored reading all of those lines.

**Supervised and Unsupervised Learning Techniques**

The difference between supervised and unsupervised learning is the presence or absence of labels (Delua). Supervised learning uses data that has labels during the training process. This allows the algorithm to place a label on unknown data that fits the same features like the training data. Unsupervised learning techniques are used to show patterns or associations between unlabeled data.

Two types of supervised learning algorithms are random forest and linear regression. A random forest is a large group of decision trees used to classify each piece of data. The output of each tree is then compared and the most likely answer is computed. Linear regression uses a linear equation to separate data into different categories. The category of the data is then used as the output of the model.

Two types of unsupervised learning are k-means clustering and neural networks. K-means clustering creates the requested number of clusters and then separates the supplied data into these clusters based on the features. A neural network is a set of algorithms that are meant to mimic the human brain (Nicholson). The artificial brain then attempts to recognize patterns in the data and break it up into buckets.

**References**

Delua, J. (n.d.). *Supervised vs. Unsupervised Learning: What's the Difference?* IBM. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

Nicholson, C. (n.d.). *A Beginner's Guide to Neural Networks and Deep Learning*. Pathmind. https://wiki.pathmind.com/neural-network.