Joshua Cervantes

CS 1030

Blanche

May 5, 2023

Final Project

**Data Tools**

For the data tools section of this project, I was introduced to different methods and ways where I could store specific data sets. There are text files, spreadsheets, and databases where we can store any collected data and organize them in ways that are easy to read. Specifically, text files contain the data itself, spreadsheets organize and analyze that data, and databases do all of that while allowing us to interact with the data as a programmer. Within the spreadsheets and databases, we can use functions, such as “Average” or “Sum” to help us find a specific number within all of our data. We can also find patterns or trends in our data and using graphs can help us determine the correlation of the data and see where the numbers increase or decrease. In this section I learned how to tell what a table of data is trying to find or compare to.

**Big Data**

In this section, I am introduced to big data sets. The digital world continues to collect massive amounts of data from the internet and how people use any online service. Large amounts of data are usually collected through particle accelerators, digital libraries, medical records, and user-facing applications. Data is collected, but the issue is how can someone store trillions of pieces of data. An example of how large data, like the millions of tweets on twitter, are stored is in the form of disk arrays which connect dozens of hard drives together. When even more space is needed, data and hard drives are stored in data centers, which are buildings that house all the hard drives. In terms of processing large amounts of data, engineers have to come up with faster algorithms or use parallel computing, which processes data in parallel. This reduces the time in processing.

**Bias in Machine Learning**

In this section, Machine learning is defined as an algorithm that improves itself based on experience. There are three types of machine learning techniques. They are reinforcement, unsupervised machine, and supervised machine learning. An important concept is neural networks, which falls under supervised machine learning. These networks work like a brain. The more input eventually leads to an output. The way it works is that a person can insert an image into a neural network, and it will familiarize itself with that image and place it under a specific category. The more images inserted, the more advanced the network becomes. Accuracy of a neural network is not the best since it is learning on its own. Machine learning is used to make predictions about a future outcome based on historical data. Hence, why machine learning is commonly used in the criminal justice system to determine who can commit a crime and when. Facial recognition services also use machine learning algorithms. They can determine the gender or race of a person. Ultimately, there is a risk. The accuracy of the algorithms has not been perfected, so this can lead to problems in the real world if someone is misidentified.

**Unit Test**

The unit test mostly had some similar questions to the exercises I took while reading the previous sections. I got 7/9 questions correct and felt like after taking it I understood most of the material pretty well.

DATA SET – CHESS GAMES

Hypothesis- Do turns affect whether or not the white or black player wins?

The process was me finding the average turns between the players which came out to around 2. The more turns someone had, meant the more they won. Ultimately the white players won 5% more of the time than the black players.