Final Project

Umulbanin Gulzar

Data Tool

Social media collects all our data based on the posts that we like. Online stores are collecting our data by-products that we view. This can be stored in Text Files, spreadsheets, and databases. After storing data, we can compute statistics and describe the data set. For data analysis, there are two popular tools which are spreadsheets and SQL. The whole point of collecting data is to find patterns in the data. We can find different patterns in the data, such as spotting trends, visualizing with charts, Statistical Fluctuations, making predictions, and Finding correlations.

Big Data

As there is an increase in the number of people connected to the internet, data sets are increasingly larger and larger in size. Most big dates are a collection of dates from lots of little sources. The world's largest particle accelerator is called the Large Hadron Collider. Digital libraries archive vast numbers of historical documents, artifacts, and media. Several healthcare providers store patient data in an electronic health record (EHR). Any application with millions of users also collects big data about user interactions. There are two important considerations when computing a system to store massive amounts of data, which are space and time.

Bias in machine learning

ML or Machine learning is a type of algorithm that improves automatically based on its experience and this is by entering more data. The three general approaches of machine learning are reinforcement learning, unsupervised machine learning, and supervised machine learning. A popular approach to supervised machine learning is the neural network. Training a network is a neural network using a massive amount of labeled data. The accuracy of a neural network is dependent on its training data. A machine learning algorithm can predict future badges on the data that is entered. Facial recognition services use machine learning algorithms to scan a face and detect a person's gender, race, emotions, or even identity.

My progress in Khan Academy

https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101

My dataset

- Link to my dataset
- The data set that I chose is titled " EMPLOYER-BASED HEALTH INSURANCE BY SEX BY AGE". It tells the estimate and the error of margin.

My hypotheses

- One of the hypotheses about this dataset is that the number of females is higher than the number of males, which was very surprising because usually in everything the number of male is higher than the female.