Pokemon-Go Data Analysis

Group Members:

Lokesh Kanagala #800957960
Vidyasagar Kanugula #800957543
Virinchi Ande #800970447

Description:

Predicting where a Pokemon can appear in the future, is one of the interesting challenge, on which a Data Scientist can work on. There have been many competitions/challenges on public platforms like Kaggle, where users work on Pokemon-Go datasets to predict Pokemon appearances.

Our project analyzes a Pokemon sightings dataset consisting of roughly 2,93,000 historical appearances of Pokémon's having latitude longitude coordinates, appearedLocalTime, Weather, PokemonType etc.

Source Dataset Size: 402 MB
Source Dataset Link: https://www.kaggle.com/semioniy/predictemall

We have developed a machine learning algorithm to predict where Pokemon can appear in future.

Tasks Involved:

K-Means Clustering:

Each record in the dataset consists of a <latitude-longitude> pair that describes the location where a Pokemon is found. But, many of those latitude-longitude pairs represents same location.

For example, The following latitude-longitude pairs < 47.826955, -117.597654>, <47.826888, -117.596629>, <47.827543, -117.597317> represent the same location "Nine Mile Falls, WA USA".

Row(Zone=5, lat=47.826834, long=-117.5978)
Row(Zone=5, lat=47.826955, long=-117.597654)
Row(Zone=5, lat=47.826888, long=-117.596629)

To tackle this challenge, we used K-means clustering on latitude and longitude pairs and have clustered the locations into 100 zones.

Naïve-Bayes:

Given the climate, day and time, our model predicts the zone in which a Pokemon is likely to appear.

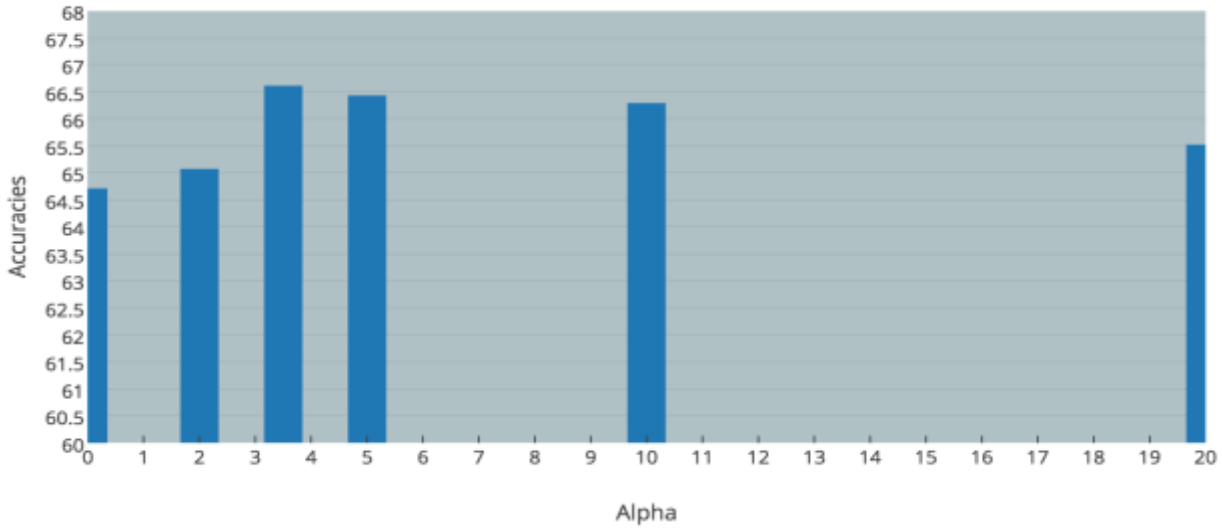
Formulation of the hypothesis:

P(zone | time, day) = P (day | zone) * P(time | zone) * P(zone)

P(day | zone) = [count(zone $\cap$ day)+α] / [count(zone) + α * dayVocabulary]

P(climate | zone) = [count(zone $\cap$ climate)+α] / [count(zone) + α * climateVocabulary]

P(time | zone) = [count(zone $\cap$ time)+α] / [count(zone) + α * timeVocabulary]



After testing the model with different alpha values, we chose α=3.5 since, it gave the highest accuracy of 66.3%.

Output Screenshots:

Input: Filename, Time, Weekday, Climate

Output:

The Zones in which you can find a Pokemon is 44, 67, 74, 83, 85

Member Contributions:

Lokesh Kanagala: K-Means , Naïve-Bayes.
Vidyasagar Kanugula: K-Means, Naïve-Bayes.
Virinchi Ande: Data Preprocessing, K-Means.