**Project Proposal**

**Clustering Tobacco Users**

**March, 2, 2018**

**Team: The Data Miners**

**Team members: Oscar Garza, Arturo Garza,**

**Ryan Thimons, Taylor Cass**

**Edgar Hernandez.**

**Description:**

**In this project we will be implementing these three clustering algorithms K-Means, K-Mode, and PAM. We will implement them using the same data set and compare the statistic performance of each algorithm. The data set we will use will be from national adult tobacco survey first, however if the dataset is too complex we will generate our own using Python sk-learn. We will be using Python to develop the algorithms.**

**-Algorithms**

**K-Means:**

**K-Means will helps us to find out, new types of groups that are otherwise not labeled ib the data set, helping us realize new things about tobacco users in the United States or confirm assumptions that we have.**

**K-Mode:**

**The k-modes method like the k-means but it define the centroid of a cluster by modes this will give us a more varied result**

**PAM:**

**Pam will give us a more robust in the presence of noise, it will be less influenced by outliers and other values rather than a mean it allows us to minimize the average dissimilarity of objects to their closes selected object by finding its medoids.**

**-References**

**We are using the National Adult Tobacco Survey (NATS) data set from data.org to conduct this project.**

**https://catalog.data.gov/dataset/national-adult-tobacco-survey-nats-f16c9**

**Milestones.**

|  |  |
| --- | --- |
| **Week 1** | **Project proposal** |
| **Week 2** | **Development of K-means Alg.** |
| **Week 3** | **Development of K-mode Alg.** |
| **Week 4** | **Development of Pam Alg.** |
| **Week 5** | **Compare and validate the algorithms.**  **Final Conclusion.** |