**Data Management Plan**

**Data Collection**

* **What data will you collect or create?**

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984.

For this project, a csv of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 253680 individuals and has 21 features and a target variable. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

* **How will the data be collected or created?**

After downloading the dataset from Kaggle, the dataset will be named as follows:

DM dataset.csv

The dataset also will be placed in a folder named: Diabetes Project

**Documentation and Metadata**

* **What documentation and metadata will accompany the data?**

The data will be accompanied by a data dictionary saved as a .docx (MS Word) format. The data dictionary will be named as follows:

DM\_Dictionary.docx

**Ethics and Legal Compliance**

* **How will you manage any ethical issues?**

The data is publicly available on the National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health website and does not contain any identifying information on participants or information that could be linked to identify them and therefore does not involve additional risk of harm for the participants who took part in the original survey but the new analysis could generate new insights.

* **How will you manage copyright and Intellectual Property Rights (IPR) issues?**

The original dataset was distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

**The availability of the data, including information about ways in which the data will be accessed, and whether there are any issues related to privacy and/or intellectual property**

The project uses dataset from kaggle, more data will be collected from the customers when the application starts working. Customers data will be cyphered as we will use only customer-ID or special ID instead of explicit names or personal information.

When we target insurance companies we will use general statistics that points out the general trends for these customers. initially as we use data from an open source data which is kaggle so we will not face any intellectual property rights issue.