**DSC 540 Group Project ReadMe File**

**Group 6**

**Members (Student ID/Email):** Abigail Keller (2026918, akelle30@depaul.edu), Kara Faciszewski (1976775, kfacisze@depaul.edu), Joel Fernandez (1394698, jferna26@depaul.edu), Elizabeth Kerrigan (1994142, ekerrig3@depaul.edu). Srinath V.S. (1979936 , SVANAMAM@depaul.edu)

**Introduction:**

The main project notebook is "Final Project.ipynb" which includes a full report with linked table of contents to each relevant section. The code files are linked in this notebook and include:

Jupyter Notebook Code Files:

Data Exploration --> Final Project 01 Data Exploration.ipynb

Pre-Processing --> Final Project 02 Pre-Proccessing.ipynb

K Nearest Neighbors --> Final Project 03.1 K Nearest Neighbors.ipynb

Naive Bayes --> Final Project 03.2 Naive Bayes, Final Project 05.2 PCA Naive Bayes.ipynb

Logistic Regression --> Final Project 03.3 Logistic Regression.ipynb

Decision Tree --> Final Project 03.4 Decision Tree.ipynb

Random Forest --> Final Project 03.5 Random Forest.ipynb

Adaboost --> Final Project 03.6 AdaBoost.ipynb

Gradient Boosting --> Final Project 03.7 Gradient Boosting.ipynb

Perceptron --> Final Project 03.8 Perceptron.ipynb, Final Project 05.8 PCA Perceptron.ipynb

Principal Component Analysis --> Final Project 04.1 Principal Component Analysis.ipynb

Kernel Principal Component Analysis --> Final Project 04.2 Kernel Principal Component Analysis.ipynb

There is additionally a folder including all html versions of the notebooks.

**Requirements:**

The package requirements are included in the requirements.txt file and its contents are also below:

matplotlib == 3.2.2

numpy == 1.19.5

pandas == 1.0.5

scikit\_learn == 0.23.1

seaborn == 0.10.1

Computer requirements include:

8 GB ram with a Quad Core APU should be adequate to run the notebooks with a suitable python environment to run the models locally.

(or)

Google Collab for remote computation with installed packages.

**Installation:**

Install python 3 and use a virtual environment to install the required packages. Anaconda would be preferred for windows and native python with an installed ipython kernel for all other Operating systems.

Install Jupyter Notebook or JupyterLab inside the environment to be able to run the notebooks.

Use pip inside the python environment to install the required packages from the requirements.txt

pip install -r requirements.txt

**Instructions:**

To reproduce the predictions from each notebook, they need to be run with the “Data” file included from the final zipped folder which includes the Attrition.csv (labels) and HR\_Employee.csv (attributes) files which are used in most of our models. The KPCA\_HR\_Employee.csv and PCA\_HR\_Employee.csv files were created to test each model using PCA and KPCA. WA\_Fn-UseC\_-HR-Employee-Attrition.csv was the original data downloaded from Kaggle (<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>). Each model notebook has detailed instructions and comments to reproduce results. The data exploration, pre-processing, PCA, and KPCA notebooks each provide notes on how to execute the functions included. In the "Final Project.ipynb" notebook, we have detailed how the data was processed, explored, and transformed.

Preprocessing steps:

1.  Saved 'Attrition' attribute (Class Label) as Attrition.csv.

2.  Normalized Numeric Attributes: 'Age', 'DailyRate', 'DistanceFromHome', 'HourlyRate', 'JobLevel','MonthlyIncome','MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'

3.  Converted Categorical(Ordinal) attributes to Categorical(Nominal). And then performed Dummy / OneHot Encoding:  'Education', 'EnvironmentSatisfaction', 'JobInvolvement', 'JobSatisfaction', 'PerformanceRating', 'RelationshipSatisfaction', 'WorkLifeBalance'

4.  Performed Dummy / OneHot Encoding: 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime'

5.  Dropped Attributes:'Attrition (Class Label)', 'EmployeeNumber(Index)', 'EmployeeCount', 'Over18', 'StandardHours', <br>'JobLevel(95% Correlation)', 'MonthlyIncome(95% Correlation)'

6.  Saved the remaining attributes as HR\_Employee.csv.