Data Analytics Capstone Topic Approval Form

**Student Name:** André Davis

**Student ID:** 010630641

**Capstone Project Name:** Machine Learning SPAM Detection Powered by Enron/TREC Public Spam Corpus

**Project Topic**: This initiative aims to develop a proficient model capable of accurately classifying unlabeled content as either SPAM or HAM (Not Spam). To achieve this, the project will utilize the TREC Public Spam Corpus and Enron Emails dataset from 2007, available at <https://www.kaggle.com/datasets/purusinghvi/email-spam-classification-dataset>.

**This project does not involve human subjects research and is exempt from** [**WGU**](https://www.wgu.edu/) **IRB review.**

**Research Question:**   
Is it possible to develop a machine learning model that can accurately classify new content as SPAM or non-SPAM (HAM), using a dataset comprising known SPAM and regular content from the 2007 TREC Public Spam Corpus and Enron emails?

***Hypothesis****:*

**Null hypothesis (H0)**- The machine learning model, developed using the 2007 TREC Public Spam Corpus and Enron emails dataset, cannot accurately classify new content as SPAM or non-SPAM (HAM) beyond what would be expected by chance. This implies that any observed accuracy in classification is due to random variation in the data.  
 **Alternate Hypothesis (H1)**- The machine learning model, developed using the 2007 TREC Public Spam Corpus and Enron emails dataset, can accurately classify new content as SPAM or non-SPAM (HAM) better than what would be expected by chance. This suggests that the model has learned discernible patterns within the dataset that enable it to effectively differentiate between SPAM and non-SPAM content.

**Context:** *Explain why the situation or question would benefit from a data analysis in less than 500 words.*Click here to enter text.

**Data:** *Identify data you will need to collect that is relevant to the situation or question.*Click here to enter text.

*If an existing data set will be used, describe the data set.* Click here to enter text.

*Explain who owns the data and why you are allowed to use this data for your capstone project.* Click here to enter text.

*Note: If you are using restricted information, please have the Third-Party Authorization Form signed by an authorized agent on behalf of the data owner. The data owner’s legal name is required on the form.*

**Data Gathering:** *Describe the data-gathering methodology you will use to collect data.*

**Data Analytics Tools and Techniques**: *Identify the appropriate data-analysis technique you will use to analyze this data.*

**Justification of Tools/Techniques:** *Explain why the data-analysis technique you chose is an appropriate technique to analyze the data collected*.

**Project Outcomes**: *List the key anticipated project outcomes and deliverables in less than 500 words*.

**Projected Project End Date**:

**Sources**:

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.