CHAPTER I

INTRODUCTION

RATIONALE OF THE STUDY

Classifying is a process of assigning something or someone into different classes or categories based on shared quality or characteristics. Humans have done this process even before the invention of modern computers, Medieval army commanders sorting their formation based on the roles of the unit, pikes up-front, archers at the back, or a renaissance doctor categorizing medicines on a shelf and labeling them. The goal is simple: to easily manage and analyze the information at hand. A simple task no doubt, but doing it manually, with a large amount of information, let's say documents, well that's where the fun stops. Managing documents manually, or in this study's case, research projects, is not ideal, especially if we're manually categorizing hundreds of research from the repository. To improve this process, we enlist the help of machines in the form of Artificial Intelligence (AI).

The advancement of modern computing gave birth to AI and eventually its underlying fields, Machine Learning (ML) and Natural Language Processing (NLP). AI technology has been around since the 1940s [1]. It's been fine-tuned throughout the past decades. There are a lot of real-world applications that use ML and NLP. For instance: chatbots, language translators, email classification and filtering. With the help of AI and its related fields, document classification can now be done automatically [2].

In 2015, all United Nation member states adopted a resolution that calls for a shared blueprint for peace and prosperity, they call this, 'The 2030 Agenda for Sustainable Development. At its heart are the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership. With this plan in mind, classifying research with the UN SDGs makes the research presently relevant.

This study aims to develop a system that an educational institution can benefit from by creating a web and mobile application that serves as a repository of all research. In addition, research after uploading to the repository is automatically classified according to its SDGs and its research topics (Artificial Intelligence, Machine Learning, Computer Vision, etc.).

THEORETICAL BACKGROUND

Machine Learning

ML is defined as a type of AI that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so [3]. This can be done by using algorithms or models to analyze and draw inferences from patterns in the data. There are three types of Machine learning: *supervised*, *unsupervised*, and *reinforcement learning*.

Supervised Learning is an approach to creating artificial intelligence (AI), where a computer algorithm is trained on input data that has been labeled for a particular output. The model is trained until it can detect the underlying patterns and relationships between the input data and the output labels, enabling it to yield accurate labeling results when presented with never-before-seen data [4]. It can be further down into two categories, ***classification*** and ***regression*** algorithms.

A **classification** algorithm is defined as a technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observations into a number of classes or groups [5]. Examples of classification are SVM and Naive Bayes.

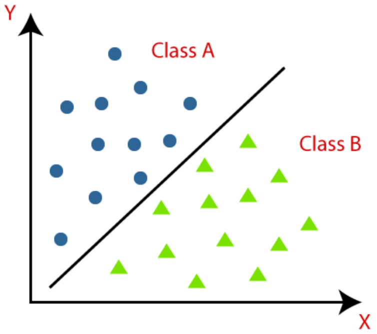

Figure 1. Sample Visualization of a Classification Algorithm

A **regression** algorithm is defined as a model that predicts the output values based on input features from the data fed into the system. Linear and Logistical regression are the most popular regression algorithms.

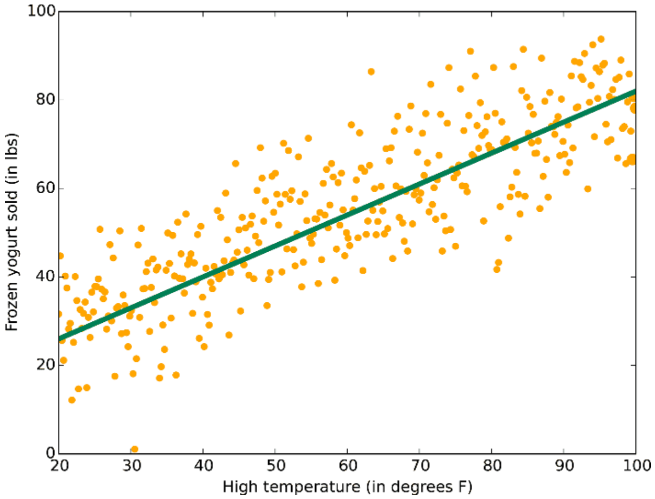

Figure 2. Sample Visualization of a Regression Algorithm

Support Vector Machines is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

Naive Bayes is a simple learning algorithm that utilizes the Bayes rule together with a strong assumption that the attributes are conditionally independent, given the class [7]. Naive Bayes is simple and easy to implement as it does not require a large amount of data for it to accurately predict. One of the disadvantages of this algorithm is, that it treats all predictors as independent variables thus limiting the algorithm's usability in real-world scenarios.

K-Nearest Neighbor Algorithm also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

Unsupervised Learning refers to the use of AI algorithms to identify patterns in data sets containing data points that are neither classified nor labeled [6].

K-means Clustering is one of the simplest and most popular unsupervised machine learning algorithms. The objective of K-means is simple: group similar data points together and discover underlying patterns [8].