A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name:* | Kevin Scully |
| *Student Number* | 24275 |
| *Module Title* | Programming for AI |
| *Assessment Title* | CA2 |
| *Assessment Due Date* | 15th December 2024 |
| *Date of Submission* | 10th December 2024 |
|  |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction 1](#_Toc184842342)

[Objectives 1](#_Toc184842343)

[Problem Description 1](#_Toc184842344)

[Objectives 1](#_Toc184842345)

[Methodology 1](#_Toc184842346)

[A. Dataset 2](#_Toc184842347)

[B. Data Preparation 2](#_Toc184842348)

[i) Nil (0.00) values 2](#_Toc184842349)

[ii) Drop columns 2](#_Toc184842350)

[iii) Outliers 2](#_Toc184842351)

[A. Data Preparation 4](#_Toc184842352)

[Architecture Diagram 4](#_Toc184842353)

[Results & Discussion 4](#_Toc184842354)

[References 6](#_Toc184842355)

[Appendices: 7](#_Toc184842356)

Introduction

Problem Description

Make a classification using your training as well as test data, using your final Neural Network configuration and discuss your findings and rationale for the chosen neural network configuration on the accuracy differential between the training and testing set.[0-15]

Objectives

Methodology

## Dataset

The ‘glass\_data’ set has been provided to be used for this assignment. This dataset provides the breakdown of nine different elements that are used to make a type of glass. There are 214 rows and 11 columns on the dataset. The columns are all numerical, with ‘type’ being categorical. The columns are ID, the nine elements (a column for each element) and the type of glass as a category.

## Data Preparation

The initial analysis of the data showed that there were no null values or duplicated rows.

### Nil (0.00) values

The description of the dataset showed that the columns ‘mg’, ‘k’, ‘ba’ and ‘fe’ had a minimum value of 0.00. There were no minus values within the dataset. It is assumed that the 0.00 values were accurate as it is possible that there was 0.00 of the elements in the formation of the glass type. As these were accurate values, it was decided for the model to only handle the values if they were extreme outliers. (Tukey, 1977) The below graph shows the percentage of the 0.00 values in each column:

A graph of values per column

Description automatically generated

### Drop columns

As the column ‘id’ had no impact on the glass type, this was dropped from the dataset. This would ensure that the column did not impact the model. Additionally, due to the high number of zero values and the low variance in the ‘ba’ column, this was also dropped from the dataset. Low variance columns provide little information and could impact the performance of the model. (Mehta, et al., 2019)

### Outliers

As the information within the dataset was accurate in the creation of the glass type, only extreme outliers were identified and handled. (Tukey, 1977) The outliers were identified using the interquartile range (IQR). The upper bound was calculated using plus 3 and lower bound was calculated using minus 3 in order to identify only extreme outliers to the lower and upper bounds. The results are shown in the below graphs:

A group of white squares with blue dots

Description automatically generated

The extreme outliers were handled by being replaced with either the mean or median to bring them within the bounds. To determine whether the mean or median should be used, the skewness of each column was calculated. If the data was moderately skewed to asymmetric (skewness was between -1 and +1) then the mean was used, and if the data was highly skewed then the median was used. (Hubert & Van der Veeken, 2008) The below graphs show the skewness of each column:

A group of graphs showing different types of data

Description automatically generated

The extreme outliers were replaced using the mean / median for each column. When these values were replaced, the upper and lower bounds would change, pushing values outside of bounds and causing them to become extreme outliers. When this happened, the skewness was re-checked to ensure the correct measurement was being used (mean or median) and the extreme outliers were replaced. This was repeated until there were no extreme outliers within the dataset.

Removing the extreme outliers allows the dense neural network to perform better by reducing noise and overfitting. (Jabbar & Khan, 2015)

### Preparing the Data for the Model

The columns were split into feature and target variables. The target variable was ‘type of glass’ which was a multiclass classification. The data was then split into training and testing sets, with 30% of the data reserved for testing. To ensure that the feature columns were contributing equally to the dense neural network, the features were scaled using standard scaler. This ensured that all features were on the same scale.

The target variable was then checked for distribution. The below graph shows that the class is imbalanced. Synthetic Minority Over Sampling Technique (SMOTE) was applied to rectify this class imbalance in the training set.

A graph of different colored squares

Description automatically generated

SMOTE ensures that each class is equally distributed, as shown by the distribution below after SMOTE was applied.

A chart of different colored bars

Description automatically generated

As the target variable has a multi-class classification, One-Hot Encoding was used to transform the numeric labels into a binary format that the neural network can process the target variable effectively. The data was then ready to be passed through the machine learning model.

## Neural Network

The neural network consisted of three dense layers, two of which were hidden layers that each had 64 neurons. The activation function Relu was used as it has been shown that it converges much more quickly and reliably than other activation functions. (Ahmed & Longo, 2022)

## Neural Network

configurations of neurons, layers, loss functions, and activation functions

Architecture Diagram

A dense neural network was built to

Results & Discussion

* 1200 (plus or minus 10%) words in a report on your topic that includes Introduction, problem description, objectives, methodology, architecture diagram, explanation about configurations of neurons, layers, loss functions, and activation functions, interpretation and discussion of your findings. In addition, also add the screenshots of your working code and its output [0-20].

# References

**There are no sources in the current document.**

***GitHub Link:*** https://github.com/kpscully116/Programmin-for-AI

# Appendices: