|  |  |
| --- | --- |
|  | |
| Programming for Big Data –Assignment 2  SPARK | |
|  | |
| Ciaran Finnegan – TU060 (Part Time) – Second Year  PROG9813: 2022/2023  MSc in Computer Science (Data Science)  Student No : D21124026  23/4/2023 |  |
|  |  |

Table of Contents

1 Introduction 3

1.1 Purpose of Report 3

2 Problem Definition – Kaggle Competition Objective 4

2.1 Kaggle Competition – This is the way… 4

2.2 Kaggle Competition – Rationale For Selection 4

3 Data Pre-Processing 5

3.1 Data Exploration 5

3.2 Feature Engineering 6

3.3 Data Preparation 6

4 Model Creation 7

4.1 Choice of Algorithms 7

5 Module Evaluation 8

5.1 Key Performance Metrics 8

5.2 Chosen Model : Why? Explanation of Decision 9

6 Spark : Key Features Used in Assignment 10

6.1 Spark 1 10

6.2 Spark 2 11

7 Kaggle Competition : My Entry 12

7.1 Outcome of Entry 12

7.2 Evaluation of Performance 12

# Introduction

## Purpose of Report

This report collates all the delivery requirements for Assignment 2 in the Programming for Big Data module (PROG9813: 2022-23).

The primary content of the report covers;

* How

# Problem Definition – Kaggle Competition Objective

## Kaggle Competition – This is the way…

*Figure – VM Shared Folder Set Up*

## Kaggle Competition – Rationale For Selection

The way…

# Data Pre-Processing

## Data Exploration

This diagram…

## Feature Engineering

In the…

Why?

## Data Preparation

In the…

Why?

# Model Creation

## Choice of Algorithms

A description…

# Module Evaluation

## Key Performance Metrics

The..

## Chosen Model : Why? Explanation of Decision

The…

# Spark : Key Features Used in Assignment

## Spark 1

The..

## Spark 2

The…

# Kaggle Competition : My Entry

## Outcome of Entry

The..

## Evaluation of Performance

The..