**A PROJECT REPORT**

**ON**

# ANALYSIS OF A DATASET THROUGH DATA MINING ALGORITHMS

SUBMITTED TO UNIVERSITY OF ROME - TOR VERGATA, ROME
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR
THE AWARD OF THE 45 DAYS SUMMER TRAINING PROGRAM
(VISHWANIKETAN UG-FELLOWSHIP)

BY

**AKHIL MUKESH SETHIA**

**UNDER THE GUIDANCE OF**
**Prof. MAURO DE SANCTIS**


**UNIVERSITY OF ROME-TOR VERGATA, ROME**

**VIA DEL POLITECNICO,**

**1, 00133 ROMA RM**

**2017-18**


**UNIVERSITY OF ROME-TOR VERGATA, ROME**

**VIA DEL POLITECNICO,**

**1, 00133 ROMA RM**

# Certificate

This is to certify that Dissertation report entitled,

**ANALYSIS OF A DATASET THROUGH DATA MINING ALGORITHMS**

Submitted By

**AKHIL MUKESH SETHIA**

is a bonafide work carried out by them under the supervision of **Prof. Mauro De Sanctis** and it is submitted towards the partial fulfilment of the requirement of **University of Rome-Tor Vergata, Rome** for the award of the 45 days Summer Training Program (Vishwaniketan UG-Fellowship).

| | |
|---|---|
| Dr. Aparna Bhirangi | Prof. Ernestina Cianca |
| **Indian Supervisor** | **Vice CTIF, Italy Coordinator** |

**Place: Rome, Italy**
**Date: 25/07/2017**


**UNIVERSITY OF ROME-TOR VERGATA**

**VIA DEL POLITECNICO,**

**1, 00173 ROMA RM**

# Certificate by Guide

This is to certify that**, Mr. Akhil Mukesh Sethia** have completed the dissertation work under my guidance and supervision and that, I have verified the work for its originality in documentation, problem statement, implementation and results presented in the dissertation. Any reproduction of other necessary work is with the prior permission and has given due ownership and included in the references.

**.....**

**Prof. Mauro De Sanctis**

**University of Rome-Tor Vergata**

# Acknowledgment

I am grateful to University of Rome-Tor Vergata and Vishwaniketan iMEET for presenting me with an opportunity to be able to come here and collaborate with my mentor **Prof. Mauro De Sanctis.**

Without his guidance and insight I would not have been able to complete my project. His advice and support have helped me shape my project and my perspective toward approaching this project. I would like to express deepest gratitude towards **Prof. Mauro De Sanctis** and sincerely thank him for his time and his inputs.

I sincerely thank **Dr. Simone Di Domenico** for his supervision and his guidance. His continuous suggestions regarding the scope of the project were invaluable. I would also like to thank **Dr. Aparna Bhirangi** whose support throughout the summer training program and guidance on documentation and structuring of the report help me collate my findings.

Without the help and support of the aforementioned individuals, this project would have not have reached a conclusion.

This has been an excellent opportunity for me to get exposure to the technology and quality of research at University of Rome- Tor Vergata. I have personally gained a lot from this summer training. This UG-fellowship has been instrumental for my development, both as an individual and a professional and I am grateful to Vishwaniketan iMEET for the same.

**Date: 25/07/2017** **AKHIL MUKESH SETHIA**

**Place: Rome, Italy**

# List of Tables



# List of Figures

# Abstract

*Data science lays the foundation to future technologies like artificial intelligence, automation and IoT. To make these technologies scalable and cost feasible it is fundamental for data mining algorithms to have low running time. Through maximisation of computational speed, we can implement real time artificial intelligence and IoT solutions from data mining algorithms. This report analyses the performance of data mining algorithms with respect to their accuracy and computational speed. Classification and clustering algorithms namely, naive Baye's, k-nearest neighbours, decision tree classification, agglomerative hierarchical clustering and k-means clustering algorithm have been analysed. Eleven classifier and cluster models have been generated each trained and tested on different combination of attributes of the Iris dataset for each algorithm. A performance metric is formulated, where ratio between normalised accuracy and testing time has been taken. The time and accuracy have been normalised to eliminate the results being skewed by one parameter. Based on these results, we observe that better performance can be achieved if algorithms operate on a subset of all attributes instead of the whole Iris dataset. A reduction in dimensionality of the dataset to optimise the speed of processing has been hypothesised.*

# Contents