**DECLARATION**

We declare that

1. the work contained in this report is original and has been done by us under the guidance of our supervisor.
2. the work has not been submitted to any other institute for any degree or diploma.
3. We have followed the guidelines provided by the institute to prepare the report.
4. We have conformed to the norms and guidelines given in the ethical code of conduct of the institute.
5. Wherever we have used materials (data, theoretical analysis, figures and text) from other sources, we have given due credit to them by citing them in the text of the report and giving their details in the references.

Signature of the student

Name: Akansha Jain

Roll number: 1502913007

Signature of the student

Name: Paras Chaudhary

Roll number: 1502913059

Signature of the student

Name: Pooja Agarwal

Roll number: 1502931102

Place: KIET Group of Institutions, Ghaziabad

Date:

**CERTIFICATE**

This is to certify that the project Report entitled, **“CATCH THE FLAMINGO”** submitted by AKANSHA JAIN, PARAS CHAUDHARY and POOJA AGARWAL in the Department of Information Technology of KIET Group of Institutions, Ghaziabad, affiliated to Dr. A. P. J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India, is a record of bonafide project work carried out by them under my supervision and guidance and is worthy of consideration for the award of the degree of Bachelor of Technology in Information Technology of the Institute.

**Signature of Supervisor:**

**Supervisor Name: Mr. Nitin Kumar**

**Date:**

**List of Figures**

|  |  |  |
| --- | --- | --- |
| 1 | How many times each item is purchased | 9 |
| 2 | How much money was made from each item | 10 |
| 3 | Total amount of money spent by the top ten users | 11 |
| 4 | Numeric Binner | 12 |
| 5 | Decision Tree | 13 |
| 6 | Confusion Matrix | 14 |
| 7 | KNIME workflow | 15 |
| 8 | Graph Example | 22 |
| 9 | Longest Conversation Chain | 25 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| 1 | Data Set Overview | 2 |
| 2 | Aggregation | 9 |
| 3 | Filtering | 10 |
| 4 | Sample Selection | 11 |
| 5 | Attribute Selection | 12 |
| 6 | Cluster Centers | 16 |
| 7 | Recommended Actions | 17 |
| 8 | Chat Data Nodes | 18 |
| 9 | Chat Data Relationships | 18 |
| 10 | Schema of the csv files | 19 |
| 11 | Chattiest Users | 27 |
| 12 | Chattiest Teams | 27 |
| 13 | Most Active Users | 28 |

**Abstract**

In this culminating project, we will analyze a data set simulating big data generated from a large number of users who are playing an imaginary game "Catch the Pink Flamingo". During the Project, we will walk through the typical big data science steps for acquiring, exploring, preparing, analyzing, and reporting.

In the initial stages, we will get the data set and get some exploratory analysis using tool such as Splunk. Then we will move into more challenging big data problems requiring the more advanced tools like KNIME, Spark and Neo4j. We will identify the big spenders from the aggregated data from several game data files by using KNIME. Then simulated chat related data to the game to be used in Graph Analytics with Neo4j.

Finally, we will show you how to bring it all together to create engaging and compelling reports. Combining all datasets together, analyzing in different technologies will contribute to our decision in increasing revenue of the game from game players.

**CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | | Page No. |
| Declaration | | i |
| Certificate |  | ii |
| List of Figures | | iii |
| List of Tables | | iv |
| Abstract | | v |
| CHAPTER 1: Introduction  1.1 Catch the Flamingo  1.2 Data Set Overview | | 1  1  2 |
| CHAPTER 2: Motivation | | 6 |
| CHAPTER 3: Literature Review | | 7 |
| CHAPTER 4: Objective | | 8 |
| CHAPTER 5: Implementation Overview  5.1 Exploring Data with Spark  5.2 Data Classification with Knime  5.3 Clustering with Spark  5.4 Graph Analytics of Simulated Chat Data With Neo4j  5.5 Recommended Actions | | 9  9  11  16  17  29 |
| CHAPTER 6: Possible Advantages | | 30 |
| REFERENCES | | 31 |