**MINISTRY OF EDUCATION AND TRAINING**

**HUNG YEN UNIVERSITY OF TECHNOLOGY AND EDUCATION**

****

**BIG PROJECT**

**BIG DATA ANALYSIS**

**YELP REVIEWS ANALYSIS**

MAJOR: COMPUTER SCIENCE

STUDENT: **VU QUANG PHUC**

CLASS: **124221**

SUPERVISOR: **PhD NGUYEN VAN QUYET**

**HUNG YEN – 2025**

**COMMENTS**

**Comments from supervisor:**

**SUPERVISOR**

**(**Signature and Full Name)

**COMMITMENT**

I solemnly declare that the project for the Big Data Analysis course, titled “Yelp Reviews Analysis” is the result of my independent work.

All references and materials use in this report have been properly cited in the References section.

All data, figures and results presented in this project are truthful and accurate.

If any information is found to be incorrect or plagiarized. I will take full responsibilities and accept my disciplinary actions imposed by the faculty and university

*Hung Yen ................. ......., ...........*

Student

**ACKNOWLEDGEMENT**

Completing this project, “Yelp Reviews Analysis” has been a challenging yet rewarding journey, and it would not have been possible without the invaluable guidance and support I have received along the way

First and foremost, I would like to express my deepest gratitude to the Department of Computer Science, Faculty of Information Technology – Hung Yen University of Technical Education. The resources, academic environment, and opportunities provided by the department have been instrumental in enabling me to carry out this project. I feel incredibly fortunate to have been part of a program that prioritizes academic excellence and hands-on learning.

I owe a special debt of gratitude to my mentor, PhD Nguyen Van Quyet, whose dedication, patience, and expertise have been a cornerstone of my progress. His thoughtful feedback and encouragement throughout the process not only enhanced the quality of this project but also deepened my understanding of advanced Python programming and data analysis. His ability to inspire confidence while challenging me to push my boundaries has left a lasting impact on my learning journey.

I would also like to thank all the faculty members of Hung Yen University of Technical Education. Their dedication to teaching and their tireless efforts to impart knowledge have equipped me with a strong foundation in computer science. The skills and insights I gained during their lectures and practical sessions have proven invaluable in overcoming the obstacles encountered in this project.

I acknowledge that, despite my best efforts, there may still be areas for improvement in this work. I warmly welcome constructive criticism and feedback from my professors, as I believe that every critique is an opportunity for growth.

Thank you for your encouragement and support throughout this journey!

**TABLE OF CONTENTS**

[GLOSSARY OF TERMS 4](#_Toc211384499)

[LIST OF TABLES 5](#_Toc211384500)

[LIST OF FIGURES 6](#_Toc211384501)

[CHAPTER 1: INTRODUCTION 7](#_Toc211384502)

[1.1 Yelp Review dataset 7](#_Toc211384503)

[1.2 Why’s this dataset? 7](#_Toc211384504)

[1.3 Project pipeline 7](#_Toc211384505)

[CHAPTER 2: BACKGROUND 8](#_Toc211384506)

[2.1 Hadoop HDFS 8](#_Toc211384507)

[2.2 MapReduce 8](#_Toc211384508)

[2.3 Data Lake & Data Warehouse 8](#_Toc211384509)

[2.4 Spark & Pyspark 8](#_Toc211384510)

[2.5 Kafka 8](#_Toc211384511)

[CHAPTER 3: IMPLEMENTATION 9](#_Toc211384512)

[3.1 Dataset 9](#_Toc211384513)

[3.2 Exploratory Data Analysis 9](#_Toc211384514)

[3.3 Data Visualization 9](#_Toc211384515)

[3.4 Data Preparation 9](#_Toc211384516)

[3.5 Rules Mining 9](#_Toc211384517)

[3.6 Modeling 9](#_Toc211384518)

[3.7 Actionable Insights 9](#_Toc211384519)

[CONCLUSION 10](#_Toc211384520)

[REFERENCES 11](#_Toc211384521)

# GLOSSARY OF TERMS

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Term | Full form | Meaning |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# LIST OF TABLES

[Table 4.1 13](#_Toc196181312)

# LIST OF FIGURES

[Figure 4.1 : What the fuck 11](#_Toc210817644)

# INTRODUCTION

## Yelp Review dataset

## Why’s this dataset?

## Project pipeline

# BACKGROUND

## Hadoop HDFS

## MapReduce

## Data Lake & Data Warehouse

## Spark & Pyspark

## Kafka

# IMPLEMENTATION

## Dataset

## Exploratory Data Analysis

## Data Visualization

## Data Preparation

## Rules Mining

## Modeling

## Actionable Insights

# CONCLUSION

Saysomething heeerefsdd hqeheoiwjwqfbefuqfqwqw

**Achievements:**

* The project has significantly improved the early detection and prediction of **Weaknesses:**
* Data Limitations: The quality and availability of data remain a challenge,
* populations or regions, leading to concerns about their generalizability.

**Future Enhancements:**

* Data Enhancement: Efforts should be made to acquire and integrate diverse datasets, including genetic data, biomarkers, and longitudinal studies, to improve

# REFERENCES

|  |  |
| --- | --- |
| [1] | Dataset link: [*Open Dataset | Yelp Data Licensing*](https://business.yelp.com/data/resources/open-dataset/) |
| [2] | Documents and slides of mentor Nguyen Van Quyet |

z