**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CERTIFICATE** | |  |
|  | **INTERNSHIP CERTIFICATE** | |  |
|  | **ABSTRACT** | | **i** |
|  | **ACKNOWLEDGEMENT** | | **ii** |
|  | **TABLE OF CONTENTS** | | **iii** |
|  | **LIST OF FIGURES** | | **vi** |
|  | **LIST OF TABLES** | | **vii** |
| 1 | **INTRODUCTION** | | **01** |
|  | 1.1 | About the Company | 01 |
|  | 1.2 | Introduction to Data Science | 01 |
|  | 1.3 | Overview of Machine Learning | 02 |
|  | 1.4 | About the project | 04 |
| 2 | **LITERATURE REVIEW** | | **05** |
|  | 2.1 | Anaconda | 05 |
|  | 2.2 | Jupyter Notebook | 06 |
|  | 2.3 | Python 3 | 08 |
| 3 | **ANALYSIS** | | **10** |
|  | 3.1 | Purpose | 10 |
|  | 3.2 | Scope | 10 |
|  | 3.3 | Motivation | 11 |
|  | 3.4 | Software Requirements | 11 |
|  | 3.5 | Hardware Requirements | 11 |
| 4 | **SYSTEM DESIGN** | | **12** |
|  | 4.1 | Machine learning system design | 12 |
|  | 4.2 | SciKit-learn | 12 |
|  | 4.3 Numpy | | 12 |
|  | 4.4 | Pandas | 13 |
|  | 4.5 | Matplotlib | 14 |
|  | 4.6 | Training the machine to predict result | 15 |

iii

|  |  |  |
| --- | --- | --- |
|  | 4.6.1 Supervised training | 15 |
|  | 4.6.2 Unsupervised training | 16 |
|  | 4.7 Breast Cancer Wisconsin dataset | 16 |
|  | 4.8 Logistic Regression Algorithm | 18 |
| 5 | **DETAILED DESIGN** | **20** |
|  | 5.1 High level design | 20 |
|  | 5.1.1 Architecture of IPython notebook | 20 |
|  | 5.1.1.1 The IPython kernel | 20 |
|  | 5.1.1.2 The Notebook | 21 |
|  | 5.2 The Notebook design | 21 |
|  | 5.2.1 Notebook dashboard | 21 |
|  | 5.2.2 Structure of a Notebook document | 22 |
|  | 5.2.3 Output format | 23 |
|  | 5.2.3.1 Stream output | 23 |
|  | 5.2.3.2 Display\_data | 23 |
|  | 5.2.3.3 Execute\_result | 24 |
|  | 5.2.3.4 Error | 24 |
|  | 5.2.3.5 Raw NBConvert cells | 25 |
| 6 | **IMPLEMENTATION** | **26** |
|  | 6.1 Overview of system implementation | 26 |
|  | 6.1.1 Selection of programming language – Python | 26 |
|  | 6.1.2 Implementation supports | 26 |
|  | 6.1.2.1 Anaconda | 26 |
|  | 6.1.2.2 Installation of Anaconda on Windows | 27 |
|  | 6.1.3 Implementation using Logistic regression | 28 |
|  | 6.1.3.1 Supervised Learning | 28 |
|  | 6.1.3.2 Supervised Learning Workflow | 29 |
|  | 6.1.3.3 Logistic regression | 30 |
|  | 6.2 Pseudo Code | 30 |
|  | 6.2.1 Reading the CSV file and Cleaning the Data | 31 |
|  | 6.2.2 Finding the Correlation between the Data | 31 |
|  | 6.2.3 Removing the least Dependent Attributes from the Dataset | 32 |
|  | 6.2.4 Training the model and testing the model | 33 |
| **7** | **TESTING** | **34** |

iv

|  |  |  |  |
| --- | --- | --- | --- |
|  | 7.1 | Unit testing | 34 |
|  | 7.2 Integration testing | | 35 |
|  | 7.3 | System testing | 36 |
| **8** | **RESULTS** | | **38** |
| **9** | **CONCLUSION AND FUTURE WORK** | | **42** |
|  | 9.1 | Conclusion | 42 |
|  | 9.2 | Future Work | 42 |
|  | **REFERENCES** | | **43** |

v