x

VIETNAM GENERAL CONFEDERATION OF LABOUR
TON DUC THANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

PHAN NGỌC HOANG ANH - 520H0511
DANG NHAT KHANG - 520H0371

REPORT OF FINAL PROJECT

FINAL GROUP WORK ASSIGNMENT INTRODUCTION TO MACHINE LEARNING

HO CHI MINH CITY, YEAR 2023

VIETNAM GENERAL CONFEDERATION OF LABOUR
TON DUC THANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

PHAN NGỌC HOANG ANH - 520H0511
DANG NHAT KHANG - 520H0371

REPORT OF FINAL PROJECT

FINAL GROUP WORK ASSIGNMENT
INTRODUCTION TO MACHINE LEARNING

Supervising Lecturer
MR. Le Anh Cuong

HO CHI MINH CITY, YEAR 2023

ACKNOWLEDGEMENTS

This report has been successfully completed by our team with many advantages coming from various aspects, but most importantly, from the guidance of Mr. Le Anh Cuong. Additionally, we are deeply grateful for the enthusiastic teaching in both theoretical and practical classes by our instructor. The factors mentioned above have significantly contributed to the completion of this report. Therefore, we sincerely appreciate the help, both direct and indirect, from our instructor. The report and the product have been carefully and strategically finalized by our team. However, every project has its strengths and weaknesses. We look forward to receiving feedback from our instructor to improve in future assignments, particularly in upcoming courses. Once again, we would like to express our gratitude!

Ho Chi Minh City, date month year 2023
Author
(Sign and clearly state full name)

EVALUATION FORM BY SUPERVISING LECTURER

Name of supervising lecturer:

Comments and Feedback:

Total score according to rubric evaluation form:

Ho Chi Minh City, date month year 2023
Supervising Lecturer

(Sign and clearly state full name)

THE PROJECT WAS COMPLETED
AT TON DUC THANG UNIVERSITY

We, the group members, hereby declare that this is our own research work and has been conducted under the scientific guidance of Mr. Le Anh Cuong. The research content and results presented in this topic are genuine and have not been published in any form previously. The data used in the tables for analysis, comments, and evaluations were collected by the authors from various sources, duly cited in the reference section.

Furthermore, this report incorporates some comments, evaluations, and data from other authors and organizations, all of which are properly cited and referenced.

Should there be any discovery of fraud, we take full responsibility for the content of our Introduction to Machine Learning final report. Ton Duc Thang University is not associated with any copyright or intellectual property violations that might arise from our actions during the project (if any).

Ho Chi Minh City, date month year 2023
Author
(Sign and clearly state full name)

REPORT OF FINAL PROJECT

ABSTRACT

Requirement 1: Perform Statistical Analysis and Data Exploration

In this section, we performed a thorough statistical analysis of the Pima Indians Diabetes dataset. We loaded the data, calculated descriptive statistics, and visualized the data distribution using histograms. Additionally, we examined the relationships between features by creating a correlation matrix heatmap. These analyses provide valuable insights into the dataset's characteristics and feature significance.

Requirement 2: Apply Basic Machine Learning Models

We implemented basic machine learning models, including Logistic Regression, Support Vector Machine (SVM), and Random Forest, to predict diabetes onset. These models were trained and evaluated, with a focus on metrics such as accuracy, precision, and recall. Model comparison allowed us to identify the most effective approach.

Requirement 3: Utilize Feed Forward Neural Networks and Recurrent Neural Networks

In this section, we explored the use of Neural Networks, specifically Feed Forward Neural Networks (FFNN) and Recurrent Neural Networks (RNN), to address the diabetes prediction problem. Detailed guidelines for implementing these models in TensorFlow were provided, including model architecture and training processes.

Requirement 4: Implement Techniques to Prevent Overfitting

To prevent overfitting in the machine learning and neural network models, we employed dropout layers and early stopping. Dropout layers were added to neural network architectures to randomly deactivate neurons during training, and early stopping was used to halt training when the model's performance on a validation set stopped improving. These techniques ensure that the models generalize well to new data.

Requirement 5: Improve Model Accuracy and Analyze Misclassifications

After training the models, we analyzed cases of misclassification to understand model weaknesses. Proposed solutions were implemented, and their impact on model

accuracy was evaluated. This iterative process aimed to enhance the models' predictive capabilities and address areas of improvement.

These abstracts provide a concise overview of each requirement, highlighting the key aspects and outcomes of the analysis and modeling process. If you need further details or have specific questions about any of these requirements, please let me know.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION TO THE TOPIC

1.1 Introduction

Diabetes, a chronic health condition affecting millions worldwide, is often detected too late, leading to severe complications. The Pima Indians Diabetes Database provides a unique opportunity to leverage machine learning for early detection, potentially saving lives and reducing healthcare costs.

1.2 Problem Statement

This report focuses on predicting the likelihood of diabetes onset in individuals based on various medical predictors. Accurate predictions can lead to timely interventions, making this study vital for healthcare advancements.

1.3 Reasons for Selection

The dataset was selected for its real-world relevance and the significant impact machine learning can have in healthcare. It represents a practical case where data science can directly contribute to improving health outcomes.

CHAPTER 2: THEORETICAL FOUNDATION

2.1 Machine Learning Overview

Machine learning, a subset of artificial intelligence, involves training algorithms to make predictions or decisions based on data. It's particularly powerful in finding patterns and insights in large datasets.

2.2 Basic Machine Learning Models

Logistic Regression: A linear model used for binary classification tasks, predicting the probability of an event occurrence.

Support Vector Machine (SVM): A robust classifier that finds the optimal hyperplane to separate different classes.

Random Forest: An ensemble of decision trees, effective in reducing overfitting and improving accuracy.

2.3 Neural Networks

Feed Forward Neural Networks (FFNN): Consists of layers of neurons, where each neuron in a layer connects to all neurons in the next layer.

Recurrent Neural Networks (RNN): Suitable for sequential data, these networks have connections that form loops, allowing information persistence.

2.4 Overfitting Prevention Techniques

Overfitting occurs when a model learns the training data too well, including its noise and outliers. Techniques like dropout, regularization, and early stopping are crucial to prevent this, ensuring the model generalizes well to new data.

CHAPTER 3: ANALYSIS AND DESIGN

3.1 Statistical Analysis and Graphs

3.1.1 Loading the Data

The dataset is loaded using Pandas, which is a powerful Python library for data analysis. Pandas' read_csv function is used to read the CSV file containing the dataset.

Code Snippet:import pandas as pddata = pd.read_csv('path_to_diabetes.csv')

3.1.2 Descriptive Statistics

Descriptive statistics are calculated using the describe() method of Pandas. This provides a summary of the central tendency, dispersion, and shape of the dataset’s distribution.

3.1.3 Histograms

Histograms for each feature are generated using Matplotlib, a plotting library for Python. This visualizes the distribution of each feature.

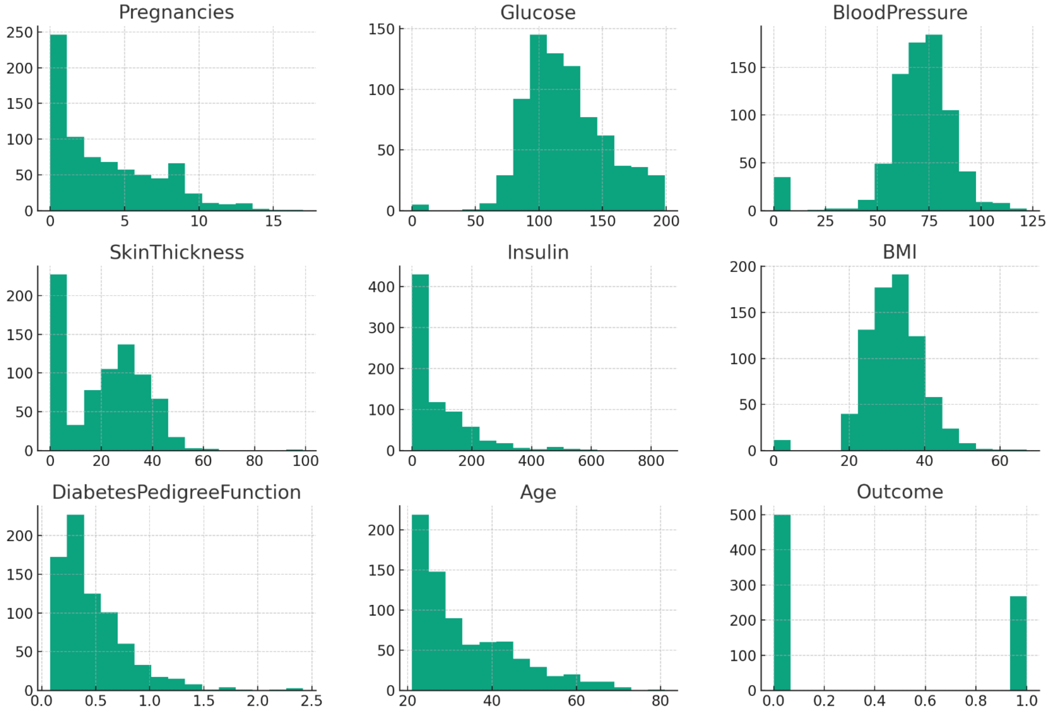

Figure 3.1: Histograms of Features

3.1.4 Correlation Matrix

The correlation matrix is calculated and visualized using Seaborn’s heatmap function. This illustrates the relationship between different features.

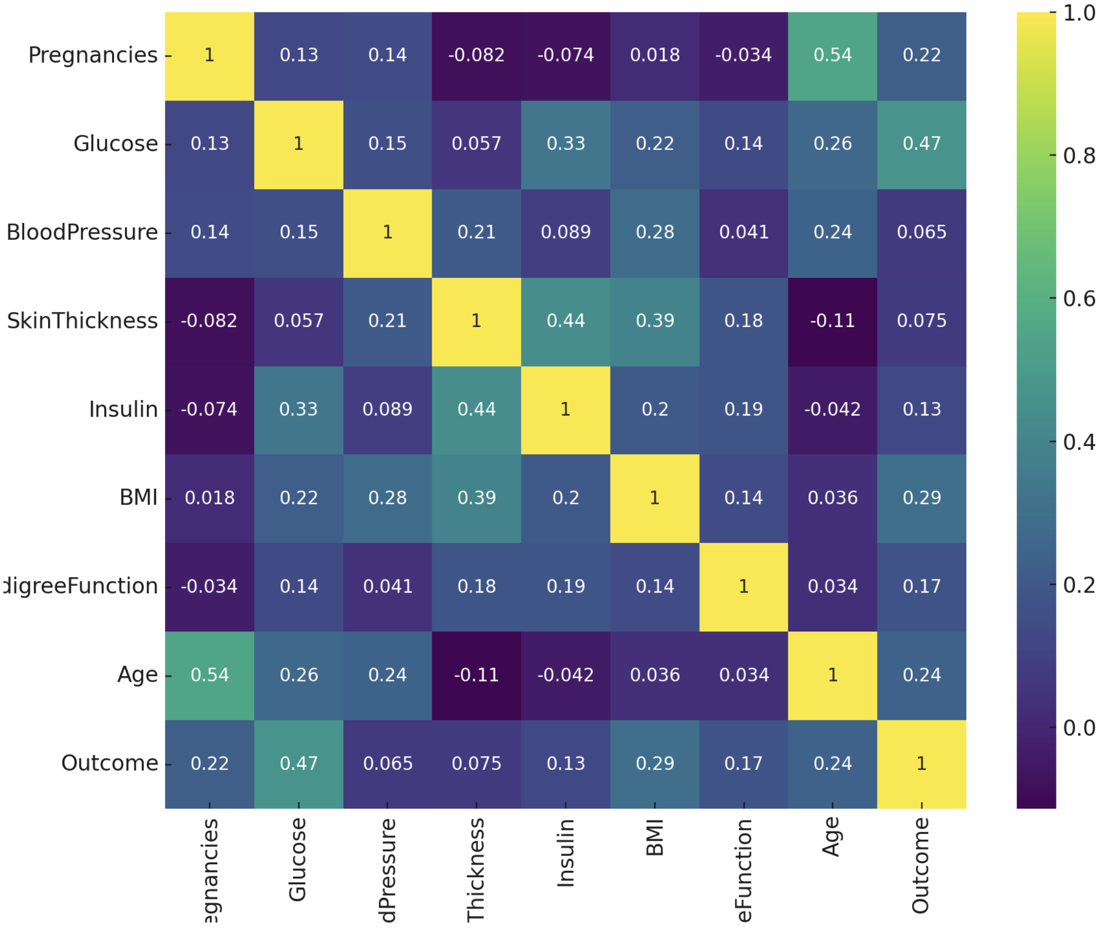

Figure 3.2: Correlation Matrix of Features

3.2 Machine Learning Models

3.2.1 Data Preprocessing

The dataset is split into features and the target variable. It is further divided into training and testing sets. The features are then standardized using StandardScaler.

3.2.2 Model Training and Evaluation

Various machine learning models like Logistic Regression, SVM, and Random Forest are trained using the training set. Their performance is evaluated using metrics like accuracy, precision, and recall on the test set.

3.3 Overfitting Prevention

Dropout and Early Stopping:Dropout layers are added in neural network architectures to prevent overfitting by randomly deactivating neurons during training. Early stopping is used to halt the training when the validation loss stops improving, preventing overfitting.