PREDICTION OF POLYCYSTIC OVARY SYNDROME (PCOS) DIAGNOSIS USING ARTIFICIAL NEURAL NETWORK ALGORITHM

ZAINAB ALI ALBASHAH

CHAPTER 3

RESEARCH METHODOLOGY

1.1 Introduction

This chapter outlines the methodology employed to investigate the factors influencing Polycystic Ovary Syndrome (PCOS) using Artificial Neural Network (ANN) modeling. The methodology encompasses the Data Science Project Life Cycle which involves data collection, pre-processing, feature engineering, and the implementation of the ANN model. This comprehensive approach ensures the replicability and robustness of the study's findings.

The Data Science Project Life Cycle provides a structured framework for carrying out data-driven projects, ensuring that each phase is systematically addressed to achieve reliable and actionable insights. In the context of this study on PCOS, the data science life cycle is used to analyze and model the data effectively, helping to achieve the research objectives.

1.2 Data Science Project Life Cycle

Problem Definition The first step is to clearly define the problem you are aiming to solve. For this project, the primary goal is to analyze the symptoms of PCOS using a combination of data-driven approaches. This involves understanding the specific questions to be answered and the impact of potential solutions on clinical decision-making and patient outcomes.

1. Data Collection Data collection involves gathering the necessary data for the project. For PCOS, this may include clinical data from medical records, hormone levels, genetic information, and lifestyle factors. The data is sourced from credible databases and research studies, ensuring that it is comprehensive and relevant to the research objectives.
2. Data Preprocessing Once the data is collected, preprocessing is essential to ensure its quality and suitability for analysis. This phase includes cleaning the data to remove errors or inconsistencies, handling missing values, and performing transformations to standardize the data. Feature engineering is also carried out to construct meaningful features that will be used in the modeling process.
3. Exploratory Data Analysis (EDA) EDA is performed to understand the data's underlying structure and patterns. This involves using visualizations and descriptive statistics to identify trends, correlations, and potential anomalies in the dataset. EDA helps to form hypotheses about which factors might be most predictive of PCOS and provides a basis for building more sophisticated models.
4. Modeling The modeling phase involves selecting appropriate machine learning algorithms and training them on the dataset. In this project, Artificial Neural Networks (ANNs) are employed due to their ability to handle complex, nonlinear relationships in data. The model is trained and validated using a subset of the data to ensure it generalizes well to new, unseen data.
5. Evaluation The model's performance is evaluated using relevant metrics such as accuracy, precision, recall, and F1-score. Evaluation helps determine the model's effectiveness in predicting PCOS and identifying key risk factors. This phase may also involve comparing different models to select the best-performing one.
6. Deployment Once the model is validated, it is deployed in a real-world setting where it can be used to predict PCOS in new patients. This phase may involve integrating the model into clinical workflows and ensuring it is accessible to healthcare professionals.
7. Monitoring and Maintenance After deployment, the model is continuously monitored to ensure its performance remains consistent over time. This

involves tracking the model’s predictions and updating it as necessary to accommodate new data or changes in the underlying patterns.

1.3 Data Sources and Collection Methods

The data for this study was sourced from Kaggle and they involve a combination of clinical records, patient surveys, and publicly available datasets on PCOS. The datasets included variables such as age, BMI, insulin levels, androgen levels, menstrual cycle regularity, and the presence of polycystic ovaries. Ethical considerations were adhered to, ensuring patient confidentiality and compliance with relevant data protection regulations.

1.4 Data Pre-processing

Before analysis, the collected two datasets merging into one data underwent several pre-processing steps to ensure quality and consistency:

1. Data Cleaning: Missing values were handled using imputation techniques, while outliers were identified and treated to prevent skewed results.
2. Normalization: Continuous variables were normalized to a common scale to facilitate the ANN model's training process.
3. Categorical Encoding: Categorical variables, such as menstrual cycle regularity, were encoded using one-hot encoding to enable their use in the ANN model.

Feature Engineering: Feature engineering was carried out to enhance the predictive power of the ANN model. This involved creating new variables based on existing ones and selecting the most relevant features for the model:

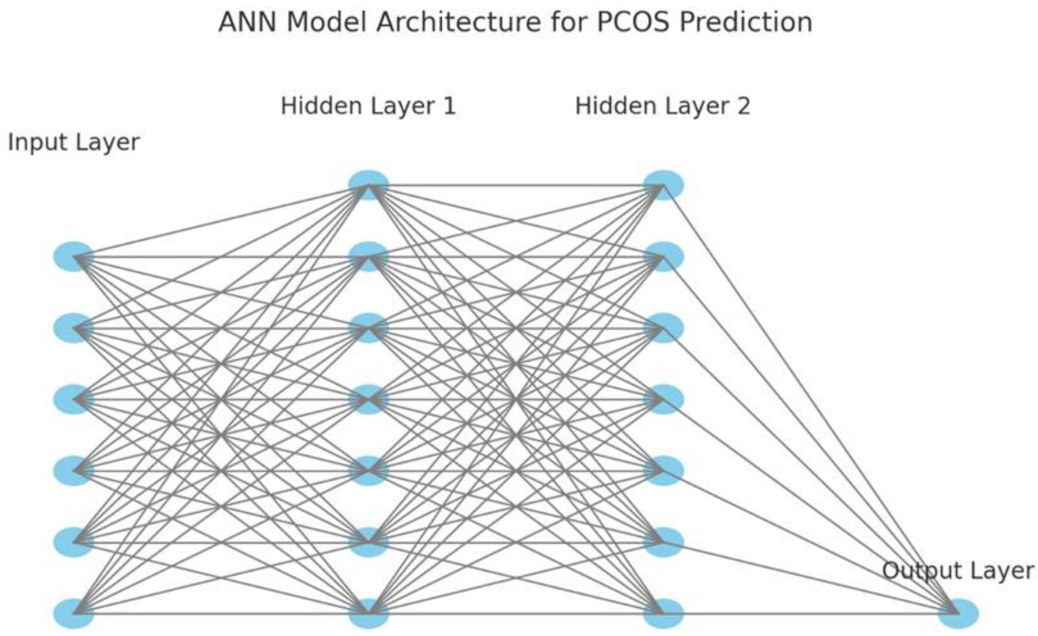
1. Interaction Terms: Interaction terms between variables, such as age and BMI, were created to capture complex relationships.

2. Feature Selection: Statistical techniques, such as correlation analysis and principal component analysis (PCA), were used to identify the most significant features for the ANN model.

1.5 Artificial Neural Network

Artificial Neural Networks (ANNs) have the potential to significantly improve the analysis of PCOS datasets by uncovering complex patterns and relationships that may not be immediately evident through traditional analysis methods. ANNs are inspired by the human brain's structure and functionality, utilizing interconnected layers of nodes to process inputs such as hormone levels, genetic data, and lifestyle factors. To effectively use an ANN for predicting PCOS, researchers first need to gather and clean the data, ensuring there are no errors or missing values that could hinder the learning process. The network is then configured with an input layer that takes in the relevant data, one or more hidden layers that process the information, and an output layer that provides the prediction of whether a patient is likely to have PCOS.

Training the ANN involves feeding it with the PCOS dataset and allowing it to learn from the patterns and correlations within the data. Through this process, the network adjusts its internal parameters, known as weights, to minimize prediction errors. Once the ANN is adequately trained, it can be used to analyze new patient data, offering predictions about the likelihood of PCOS. This capability aids healthcare professionals in making more informed decisions about diagnosis and treatment. Moreover, ANNs can identify which factors, such as specific hormone levels or genetic markers, are most strongly associated with PCOS, providing valuable insights that can drive future research and therapeutic development. By handling the complexity and variability of PCOS data, ANNs offer a powerful tool for advancing our understanding of the condition and improving management strategies for those affected.



ANN Model Architecture for PCOS Prediction