Chapter 5
Machine Learning Algorithms for Diabetes Classification Within the CRISP-DM Framework

Ismail Lamaakal
https://orcid.org/0009-0009-6532-2573
University Mohamed I, Morocco

Bentaleb Youssef
University Mohamed I, Morocco

Yassine Maleh
https://orcid.org/0000-0003-4704-5364
Sultan Moulay Slimane University, Morocco

Ibrahim Ouahbi
University Mohamed I, Morocco

Khalid El Makkaoui
University Mohamed I, Morocco

ABSTRACT

This research paper explores the application of machine learning for early and accur- ate diabetes diagnosis, utilizing the CRISP-DM framework for structured analysis. We conducted a comprehensive study using the Pima Indians Diabetes database to assess five prominent machine learning algorithms: Support Vector Machine (SVM), Logistic Re- gression, Decision Tree, Naive Bayes, and Random Forest. We enhanced these algorithms' predictive capabilities through GridSearchCV

DOI: 10.4018/979-8-3693-8387-2.ch005

optimization. Our analysis revealed the Decision Tree classifier as the top performer for diabetes diagnosis accuracy within the CRISP-DM framework, achieving impressive performance metrics: 96% recall, 94% accur- acy and F1 score, and 90% precision. Despite exploring ensemble learning, which combines insights from all models, it did not surpass the effectiveness of the standalone Decision Tree model. Given the medical context of our study, we prioritized the recall score, focusing on correctly identifying actual diabetic patients.

INTRODUCTION

Diabetes has recently become a chronic metabolic disease and a big cause of concern in most leading areas of public health, which, now, are attended to by major global forces. In the year 2019, approximately 463 million adults aged between 20 and 79 years were living with diabetes, according to statistics by the International Diabetes Federation (IDF) International Diabetes Federation (2019), and the projection for this number is 700 million by the year 2045, (Cho et al., 2018). The disease not only poses some serious health risks, such as the complications of the heart and kidneys, blindness, and failure of body organs, (Zimmet et al., 2014), but also presents a considerable economic burden on health facilities all over the world.

Timely diagnosis of diabetes is important from the point of view of management and preventing complications. Until now, the diagnosis was dependent on the clinical presentation followed by laboratory tests, but that was not a very timely and accurate process. Machine learning (Lamaakal, El Makkaoui, & Ouahbi et al., 2024; Lamaakal, El Mourabit, & El Makkaoui et al., 2024; Lamaakal, Maleh, & Ouahbi et al., 2024; Lamaakal, Ouahbi, & El Makkaoui et al., 2024; Lamaakal et al., 2025) represents one of the applied artificial intelligence branches, where it comprises systems that could aid in early diagnosis through data mining over big sets of data in search of corresponding patterns and correlations, (Kavakiotis et al., 2017). Using machine learning algorithms, it helps healthcare providers identify the possibility of probable diabetes at an earlier stage, hence an apt opportunity for intervention at the right time with proper and individualistic treatment plan, (Dagliati et al., 2018).

The other framework guiding a systematic process that may be followed to build machine learning models for health applications is the CRoss-Industry Standard Process for Data Mining (CRISP-DM), (Wirth & Hipp, 2000).

These phases are organized in six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Adherence to the CRISP-DM framework, therefore, becomes an assurance of the ability to develop and deploy machine learning models with effectiveness that translates into improved patient outcomes in healthcare. In this study, we apply the CRISP-DM