DIABETES RISK PREDICTION

VALENTINA MAZORRA MEDINA[1]	20201189730
SARA CRISTINA PRADA MEDINA[2]	20201186762
FERNANDO JOSÉ SILVA GUTIÉRREZ[3]	20201186317

Software Engineering

Data Science, Presented to Eng.
Juan Antonio Castro Silva

Universidad Surcolombiana
Neiva, Huila, Colombia
2022

INDEX

1. INTRODUCTION

Every day, a wealth of information is collected in medical databases around the world: diagnoses, prescriptions or test results, which the respective healthcare provider stores in the patient's medical records. This vast amount of information, unprecedented in the history of medicine, fosters a paradigm shift in the way we understand the development of disease, moving from an approach based on expert knowledge and experience to an approach that uses rich information and modernity. Analytical tools.

The benefits of this paradigm shift are particularly evident in the treatment of a chronic and complex disease such as diabetes. The ability to analyze the disease progression of millions of patients will change the way diabetes is diagnosed and treated, improve patients' quality of life and health outcomes, and reduce system costs.

Machine learning (ML) methods have proven to be very useful in exploiting the wide availability of data to generate new knowledge.

Precisely in this context and in order to train future professional software engineers, the Universidad Surcolombiana requires them to acquire the basic tools and experience (even if partial) for the development of data science with the help of artificial intelligence. The Computer Data Science 1 course is responsible for preparing software engineering students for this challenge, where students are evaluated on how they were able to meet such challenge through the development and presentation of their work.

Continuing with the given guidance, this paper proposes to establish a diabetes detector that, by means of previous analysis, detects whether or not the person in question has diabetes. This by means of Python, PostgresSQL, Flask and Numpy to build a page with its back-end, front-end, databases and training model with artificial intelligence (AI) that detects this.

2. JUSTIFICATION

The importance of this study is the contribution to the prediction of the risk of developing diabetes, which is very important because it allows the evaluation of several surrounding factors, which gives positive results for all involved. topic from the economic point of view, early diagnosis is very important, because would not only benefit the patient in terms of costs, such as: transportation, supportive care medications, devices to measure glucose levels, but also to the system in terms of reducing total national costs, as exemplified by serious complications (diabetes).

The cost of the system is also reduced in that it reduces the total national costs in terms of the costs of serious complications (operations, amputations, hospitalizations for seizures), drugs, tolerance tests, injections, etc., which mostly and differently from patients, are not always the same. Which in their majority and different patients, considering the country's population, reach a high economic value for the country.

This is fundamental for the society affected by this disease, because the diagnosis made in number can be great and essential for the patient; however, this issue does not cease to be important for nursing homes, nurses, educational institutions and other members interested in the subject, such as the health system.

3. OBJECTIVES

3.1. GENERAL OBJECTIVE

Develop a model with Artificial Intelligence (AI) to determine whether a person has risk to suffer diabetes or not.

3.1. SPECIFIC OBJECTIVES

- Install libraries, packages and programs necessary for the development.
- Develop the machine learning component to detect diabetes.
- Build the dataset to train the artificial intelligence.
- Generate the entity-relationship model of the database that connects with the artificial intelligence.
- Create the front-end by means of Flask.
- Deploy the program in the cloud.

4. ABOUT THE DATASET

The dataset was collected and circulated by “National Institute of Diabetes and Digestive and Kidney Diseases” which is available at Kaggle in the name of Pima Indians Diabetes Database. The main objective is to predict whether a patient has diabetes or not, based on the diagnostic measurements gathered in the database. All patients belong to the Pima Indian heritage, and are females of ages 21 and above.

The dataset contains 768 observables with eight feature variables. Theses features have been provided to help us predict if a person is at high or low risk of developing diabetes:

- **Pregnancies:** Number of times pregnant
- **Glucose:** Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- **BloodPressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin**: 2-Hour serum insulin (IU/ml)
- **BMI:** Body mass index (weight in kg/(height in m)2)
- **DiabetesPedigreeFunction:** Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
- **Age:** Age (years)

Feature Variables	**Type**	**Unit**
Pregnancies	Integer	—
Glucose	Integer	mg/dL
Blood Pressure	Integer	mmHg
Skin Thickness	Integer	mm
Insulin	Integer	IU/mL
Body Mass Index	Float	kg / m^2
Diabetes Pedigree Function	Float	—
Age	Integer	—
Target Variables (Output)	**Type**	**Unit**
Diabetes Risk	Int	—

5. TRAINING AND EVALUATING THE MODEL WITH RANDOM FOREST CLASSIFIER

To train the model, we used Random Forest Classifier, which is a classification algorithm consisting of many decision trees. As we know, a forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

The observed model’s performance has a classification accuracy of 80.5%

6. DATABASE'S DATA DICTIONARY

A platform was developed to provide to any person the tools to evaluate their diabetes risk through the Diabetes risk prediction model. Also, a database was created to register user's results in order to have a history that can be consulted by the user if they want to.

Data's attribute is provided in the following table:

Entity	Attribute	ID	Type	Size
User	id	id	int	4 bytes
	username		varchar	15
	email		varchar	80
	password		varchar	4 bytes
UserPredictions	id	id	int	4 bytes
	user_id		int	4 bytes
	date		Timestamp without time zone	8 bytes
	result		varchar	25