**Tittle-AI based Diabetes prediction system**

**Abstract:**

Data mining and Machine learning have become a vital part of different disease detection and helps in prevention of various diseases such as diabetes. The purpose of this paper is to evaluate data mining methods and their performances that can be used for analyzing the collected data about the Diabetes. We found out the most appropriate data mining methods to analyze the data by comparing them theoretically and practically. Some attributes of this dataset are: Age, Body Mass Index, Insulin, Glucose, etc. Methods are applied on these data to determine their effectiveness in analyzing and preventing diabetes.Data wrangling is essentially the process of cleaning unstructured data sets so that they can be explored and analyzed more effectively. It can involve selecting the relevant data from a large set, merging data sets, fixing/removing any corrupt data, identifying anomalies or outliers, standardizing data formats, checking for inconsistencies, etc. Ultimately, the goal is in using data wrangling to give analysts data in a user-friendly format.A neural network is a data processing system consisting of a large number of simple, highly interconnected processing elements in an architecture inspired by the structure of the cerebral cortex portion of the brain. Hence, neural networks are often capable of doing things which humans or animals do well but which conventional computers often do poorly. Neural networks have emerged in the past few years as an area of unusual opportunity for research, development and application to a variety of real world problems.Evaluations on the data showed that the Method with a higher performance is the “Decision Tree”. This was achieved by some performance Measures, such as the number of instances correctly classified, accuracy, precision, recall and F-measure, that has brought better results compared to other methods.Therefore we come to the conclusion that the data mining methods and machine learning contribute to the predictions on the possibility of occurrence of the diabetes.

**Introduction**:

Diabetes is a disease that is increasingly affecting the world even to the most developed countries. Diabetes by the nature of its development as a globally problematic disease requires maximum commitment from medical staff, patients, family and society. Diabetes is a disease with high social, health and economic costs diabetes is a chronic disease characterized by an increase in glucose or blood sugar levels because the body cannot produce insulin or its production is insufficient, or insulin is not able to act on the cells of the organism. Medics still do not know Exactly why such a thing is happening and they have called the cause and still there is no medicines for diabetes and it is caused by x syndrome according to doctors. Historically diabetes treatment has been done by fighting the symptoms and not the cause. According to the World Health Organization, Diabetes affects about 5% of the world’s population and the number of patients is constantly increasing In developed countries, diabetes and the largest number of diabetics are found in people over 65 years of age. Whereas in developing countries where our country is part of the largest number of diabetics is found in the age of 45-64 years, but in recent years type 2 diabetes is more commonly encountered also in The age of 30-40 years. The availability of historical data naturally leads to the application of data mining techniques for pattern discovery. The goal is to find rules that help understand diabetes and make it easier to diagnose it sooner. Prevention of diabetes is of great interest in the field of medicine. The use of data mining accelerates data analysis, and analysts can examine existing data to identify patterns and trends of diabetes.

**Data Wrangling:**

Data scientists run complex models to analyze and interpret data so that it generates actionable insights for an organization’s stakeholders. But before they can do this, the data needs to be complete, consistent, structured, and free from bugs.

**Data Wrangling with python**:

Python is generally considered to be a data scientist’s best friend. According to a 87% of data scientists said they regularly used Python, far more than the next most used languages, SQL (44%) and R (31%).Python is popular because of its simplicity and flexibility, but also because of the huge number of libraries and frameworks that data scientists can use. Here are some of the most useful and popular for data wrangling which includes:

Pandas: One of the tools for data wrangling, Pandas uses data structures called DataFrames, with built-in methods for grouping, filtering, and combining data.

NumPy: This is primarily used for scientific computing and performing basic and advanced array operations.

SciPy: This can execute advanced numerical routines, including numerical integration, interpolation, optimization, linear algebra, and statistics.

Matplotlib: A powerful data visualization library that can convert data into graphs and charts that are more user-friendly when it comes to modeling and analysis.

Scikit-Learn: A popular machine learning and data modeling tool that is really useful for building regression, clustering, and classification models.

**Keys steps of data wrangling:**

Cleaning data

. Deal with Missing data

Transformer data

Explore Data

. Correlation

. Grouping

In this project we are going to use numpy module to predict diabetes using Artificial intelligence;

. Numpy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices.Numpy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python…In Python we have lists that serve the purpose of arrays, but they are slow to process.

NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.

The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.

**Importance of Data wrangling:**

Data wrangling software has become such an indispensable part of data processing. The primary importance of using data wrangling tools can be described as:

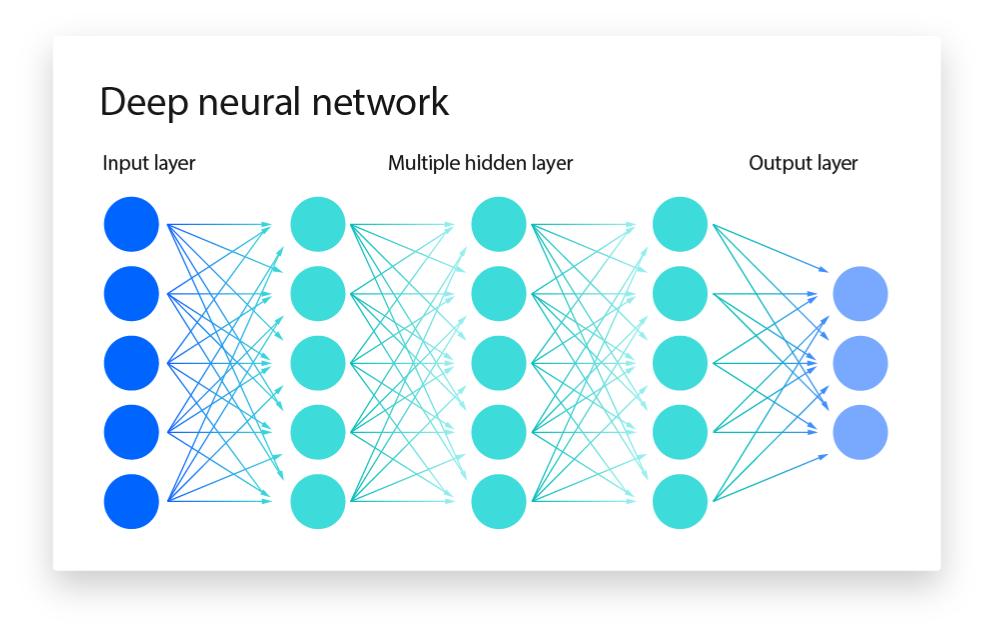
* Making raw data usable. Accurately wrangled data guarantees that quality data is entered into the downstream analysis.
* Getting all data from various sources into a centralized location so it can be used.
* Piecing together raw data according to the required format and understanding the business context of data
* Automated data integration tools are used as data wrangling techniques that clean and convert source data into a standard format that can be used repeatedly according to end requirements. Businesses use this standardized data to perform crucial, cross-data set analytics.
* Cleansing the data from the noise or flawed, missing elements
* Data wrangling acts as a preparation stage for the data mining process, which involves gathering data and making sense of it.
* Helping business users make concrete, timely decisions

**Neural Network**:

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts.



Input Layer– First is the input layer. This layer will accept the data and pass it to the rest of the network.

Hidden Layer– The second type of layer is called the hidden layer. Hidden layers are either one or more in number for a neural network. In the above case, the number is 1. Hidden layers are the ones that are actually responsible for the excellent performance and complexity of neural networks. They perform multiple functions at the same time such as data transformation, automatic feature creation, etc.

Output layer– The last type of layer is the output layer. The output layer holds the result or the output of the problem. Raw images get passed to the input layer and we receive output in the output layer.

**Conclusion :**

We have covered a lot of ground in this workbook. Hopefully, you now have a better understanding, both conceptually and practically, of the steps required to clean your data. You probably also understand that data wrangling is the most time-consuming part of the data science project. It is also indispensable. Very few datasets are tidy from the start. And no modeling, visualization, nor presentation of your data can happen unless it has been tidied and reshaped in order to be ready for the other steps in the process.In these chapters, you have also been introduced to a set of tools to help us wrangle our data. There are many more out there in the R community. As you practice more with R, you will acquire more skills and tools to work with data. Neural network is a vast subject. Many data scientists solely focus only on Neural network techniques.In this session, we practiced the introductory concepts only. Neural Networks has much more advanced techniques. There are many algorithms other than backpropagation.Neural networks particularly work well on some particular class of problems like image recognition.