Big data analytics in healthcare: tools, challenges and opportunities

EVANGELIA N. PETRAKI
Department of Economics
National and Kapodistrian University of Athens
Sofokleous 1 Str, 105 59
GREECE

Abstract: - The term big data has been used in recent years to describe the enormous volume of data collected by various means and at a very rapid pace in all areas of human activity. Big data differs from traditional data sets and requires special management due to its volume, structure and type. For this reason, specialized technologies for their processing are being developed. The use of machine learning methods to take advantage of the big data collected in the health sector is imperative as it can significantly contribute to making predictions, preventing illnesses, improving patient healthcare, reducing costs and optimizing the use of available resources. This paper deals with Big Data, briefly presenting its characteristics, the sources of its creation, as well as the basic technologies used for its processing and utilization in the health sector. In the second part of this work, medical data on diabetes were used, different software tools were integrated to get information using SQL queries and machine learning methods were applied using the Weka open-source software to evaluate and compare the results for diabetes prediction. The paper concludes with useful concepts arising from the analysis, as well as challenges and suggestions for future research.

Key-Words: - Big Data, Data analytics, Machine Learning, Healthcare, Spark SQL, Weka, Diabetes prediction

Received: August 8, 2025. Revised: October 21, 2025. Accepted: March 5, 2026. Published: June 10, 2026.

1 Introduction

The term big data has been widely used in recent years to describe the enormous volume of data that is collected in multiple ways and devices. Many definitions have been proposed to describe the term big data, which focus on the diversity, enormous size, and complexity of big data, characteristics that make it difficult or impossible to process and manage it with classical methods. Variety, variability, enormous size and complexity are some basic characteristics of big data, and it is necessary to use special methods to process and manage it [1], [2], [3],[4],[5].

Five important characteristics are used to describe big data and are briefly referred to as the 5Vs (Figure 1). These characteristics are: a) Volume, b) Velocity, c) Variety, d) Veracity and e) Value [2],[4], while other scientific reports focus on only the first three characteristics, the 3Vs [3]. Volume is related to the size and scale of the data and refers to the amount of data generated every second. Velocity refers to the speed at which data is generated, and it also concerns the time and delay in handling it. Variety refers to the different formats and the structural heterogeneity of different types and sources of data. Veracity refers to the completeness, accuracy and quality of the data, and it is also related to its reliability and consistency. Value is about how important the data is and how businesses can use it for decision-making.

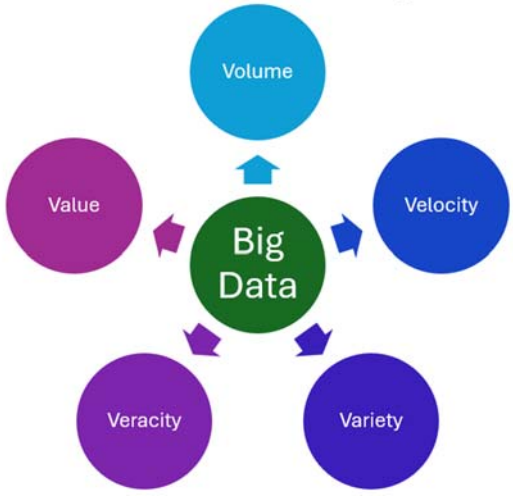

Fig. 1. Big Data 5Vs

Big Data is everywhere, created in every interaction with social media and the internet, in transactions that take place when using smart devices and sensors. In science, Big Data hasn't just changed the way research is done; it's fundamentally changed the scientific method itself. Traditionally, a scientist would make a hypothesis, collect small sets of data, and test it. Today, with so-called "Data-Driven Science," scientists collect vast amounts of data and then use the appropriate Machine Learning

algorithms to discover patterns and trends that the human mind could never see on its own. Big Data has transformed many of the most basic scientific disciplines such as Biology and Medicine, Meteorology, Social Sciences and Humanities, etc. Big data is the “raw material”. With the advanced tools offered by Artificial Intelligence, it is transformed into "predictive power" and helps in better and more effective decision-making.

2 Big Data in Healthcare

In all areas of human activity, the need to use big data is imperative as the knowledge it can provide can improve the decision-making process as well as the prediction of future events. In the health sector, the utilization of big data involves significant challenges and difficulties as it is sensitive personal data, of large volume and of different structures, which needs special processing that requires a combination of different data sources, management of complex data, accuracy, confidentiality and utilization of the appropriate data storage, management and analytical tools.

The use of big data in the healthcare sector can help scientists understand hidden trends and patterns in data that will help them make better predictions, treat patients more effectively, reduce healthcare costs, enhance research, etc. There are many big data sources for the health sector, including medical laboratories, information systems from hospitals and research centers with records and historical data of patients, treatments, drug side effects, research studies about diseases and drugs, etc. Nowadays, medical data is also collected from device applications, such as smart watches. Even discussion forums and social media platforms are a valuable source of medical data. All these data sources form a huge volume of big data that is constantly being enriched with more data and can be used to promote research on the topic of effective disease treatment, improve the performance and conditions of health institutions, as well as upgrade the processes and functions of all those involved in the health sector.

The big data gathered in the health sector can be used for more efficient treatment of patients and better management of health units [9]. The challenges are great and concern dealing with the problems arising from the volume of data, its format and nature, technical issues and tools, etc.

Big data sources in healthcare are divided into the following basic categories [9]:

- Clinical data: from Electronic Medical Records from hospitals, from image centers, pharmacies free text notes from physicians, genomic data etc.
- Biometric data: derived from various types of devices which monitor pressure glucose level, weight etc.
- Financial data: economic data from hospitals, research centers, suppliers etc.
- Data derived from scientific research carried out in research centers, universities, hospitals and concerning drug research, new methods of treatment, medicines etc.
- Data provided from patients
- Data from social media

In the healthcare sector, in addition to the huge volume, the format of the data collected is another important challenge that must be addressed when analyzing the data. Image, video, signal, and molecular data complement the already huge amount of numerical and textual data. The multiple dimensions of interconnected and interdependent heterogeneous multidimensional data create an additional challenge in the analysis process. At the same time, strong requirements for security and anonymization of sensitive medical data require the use of powerful and reliable software tools for data collection, storage, management and analysis [6].

The use of big data in the health sector can contribute in multiple ways by creating applications for Personalized Medicine, Predictive Analysis to predict future events, creation of Electronic Health Records (EHRs) that collect the patient's entire medical history, Telemedicine, Real-Time Patient Monitoring through IoT devices (e.g. smart watches, blood sugar sensors), Drug Research and Development, Epidemiological Surveillance as was done in the COVID-19 pandemic where big data analysis at the population level could identify outbreaks of infection in a timely manner, predict the spread of a disease and help manage public health resources.

Compared to other sectors (e.g. banking or retail), big data in healthcare has unique characteristics and peculiarities. Specifically, big data in healthcare is characterized by Extreme Diversity since the data is not only structured (e.g., blood sugar measurements, age) but also semi-structured or unstructured such as medical notes (free text), imaging (x-rays, MRIs), audio files, or even genomic data. Reliability and accuracy are essential critical features as noise or incorrect data can have consequences for patients' lives. The requirement for continuity in medical data to cover the entire lifespan of an individual, from birth (or even before, through prenatal screening) to death,

creates another layer of complexity that involves connecting fragmented information over decades. Furthermore, the speed (Velocity) at which data is generated in real time, in places such as Intensive Care Units (ICUs) or via mobile devices, is another feature that requires systems that can analyze it immediately. An additional characteristic of medical data is its sensitivity, as it constitutes personal data, which creates high demands for its strict protection. All these special characteristics create special requirements for medical data storage, processing and analysis by software systems.

Big data analytics in healthcare use several software tools and contribute to multiple areas such as [7]:

- Evidence-based medicine
- Genomic analytics
- Analysis to reduce fraud and waste abuse
- Device/remote monitoring
- Patient profile analytics etc.

Technology contributes in multiple ways to big data analytics in health care, while cloud programming, mobile devices, and the Internet of Things, with their respective applications, provide many opportunities and challenges. The application of the 4P model (Preventing, Predictive, Personalized, Participatory) to achieve the 5R goal in Medical (Right patient, Right Drug, Right Dose, Right route, Right time) and the 5R goal in healthcare (Right living, Right care, Right provider, Right value, Right innovate) is at the core of this process to achieve the goal of a unified diagnosis that can explain all of the patient's symptoms.[8]

3 Data Life Cycle

During the data "lifecycle," we can distinguish different phases, each of which can be supported by a set of tools and technologies. Technology of each phase is usually used in combination to create integrated solutions depending on the needs of the organization. There is no specific data life cycle, and the proposed phases that make it up vary in the literature. The consideration of the data life cycle, which in some sources is defined as a “data life cycle” while in others as a "pipeline" or even a "Big Data Value Chain", contributes to the categorization of the various technologies used for big data analytics. [10],[11],[12]

Next, the phases that make up the life cycle of big data are briefly presented. It is important to note that the following sequence is not absolute, and slight variations may be observed in different research proposals.

Data Collection: Data collection includes gathering data from different sources and is a critical part of data analysis, since effective data collection provides the necessary information required for the smooth processing of the remaining phases of the analysis.

Data Processing / Data Cleaning: Data pre-processing and cleaning is a demanding and laborious task, and often very complex and time-consuming. In this phase, data management procedures are carried out to remove noise, unwanted data elements, duplicate entries, fill in missing values, etc. Data cleaning and error detection and correction improve data quality, a requirement that is particularly important for the next steps.

Data Storage: One of the key issues concerning the management of big data is its storage. Data storage must be done in a way that allows scalability, data search, and can flexibly and efficiently support structured, semi-structured, and unstructured data, and meet existing data security requirements.

Data Analysis: is the process of examining and interpreting data to uncover insights, patterns, and trends that inform decision-making, drive innovation, and improve performance. This includes all computational and statistical techniques for analyzing data, such as artificial intelligence (AI) algorithms leading to the acquisition of knowledge, the creation of categorizations and models for predictions, or the extraction of causality.

Data Visualization and reporting: is the representation of data in a visual or graphical form. It is an easy and fast way to express and visualize complex patterns and relationships. Due to the volume, format and complexity of big data, it becomes difficult to visualize it.

4 Software Tools

The Big Data Life Cycle includes all stages from the moment information is created to its final analysis and storage. Big data gathering is done from multiple locations within or outside the organization, while the format of the data varies. Data storage - whether in the traditional table or CSV format, or in other formats - is carried out in data warehouses where data is transformed to obtain the appropriate format for big data analytics platforms and tools. A multitude of software tools and platforms are available for big data which manage the enormous volume, velocity and variety of data that traditional databases cannot process, transforming raw data into useful information. Figure 2 presents the different stages from data collection to data transformation as well as the most important software tools and platforms for

big data which offer functions for performing queries, reporting, data mining etc.

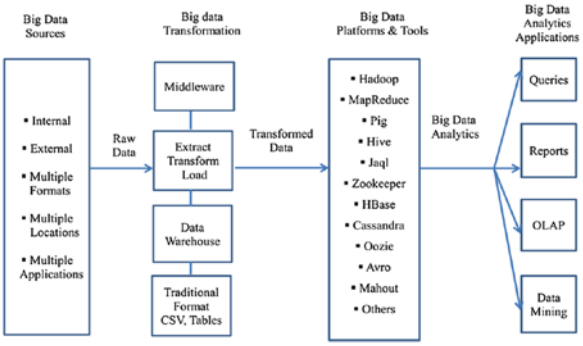

Fig. 2. An applied conceptual architecture of big data analytics [7]

Managing and extracting knowledge from medical data requires a set of specialized tools and technologies that can cope with its volume and complexity. These technologies and tools can be categorized as follows:

• Distributed Processing Platforms: Tools such as Apache Hadoop and Apache Spark are fundamental.

• Cloud Platforms: Major Cloud providers such as Amazon Web Services (AWS), Microsoft Azure and Google Cloud offer specialized platforms for healthcare (e.g. Google Cloud Healthcare API or Azure API for FHIR). They provide the necessary computing power and secure storage with strict compliance with healthcare standards (such as HIPAA/GDPR).

• NoSQL Databases: Because medical data is often unstructured (e.g., diagnostic texts, medical images), traditional relational systems (SQL) are not sufficient. NoSQL databases such as MongoDB or Cassandra are widely used for storage flexibility.

• Artificial Intelligence and Machine Learning (AI & ML): Frameworks like TensorFlow and PyTorch are used to train algorithms. Through them, Deep Learning models are developed that can identify tumors in MRI scans or predict the progression of a disease.

• Data Visualization Tools: Platforms like Tableau, QlikView, and Microsoft Power BI take on the task of transforming complex analytical results into understandable interactive dashboards.

Special mention should also be made of the role of Wearables, such as smartwatches, fitness trackers and advanced biosensors, which are now one of the fastest growing and most important sources of Big Data in health. Wearables bring health monitoring into the user's everyday life and allow continuous monitoring and early warning.

Below is a summary of the most important software tools suitable for storing and analyzing medical data:

Apache Storm: an innovative open-source data processing system for high-scale data streams [13].

Apache Flume: a distributed, reliable, and available tool for the effective collection, aggregation and movement of large amounts of log data. Its architecture is simple and uses a simple extensible data model that allows for online analytic applications. [14]

Apache Kafka: is a global, open-source, high-throughput Apache messaging system, primarily used to perform real-time data pipelines, streaming analytics, and data integration.[15]

OpenRefine: is a powerful and open-source tool used for data cleaning, transformation and exploration, suitable for managing large volumes of data and preparing it for further analysis. [16]

MongoDB: is an open-source document database that provides high performance, high availability, and automatic scalability and is defined as a NoSQL database. [17]

Apache Cassandra: is a NoSQL distributed database, which manages data across thousands of servers, provides high availability as well as high fault tolerance with zero downtime. [18]

Apache Spark: is one of the most popular open-source data processing tools. It is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters. [19]

Tableau for data analysis and visualization [20].

Power BI [21] for data visualization.

Programming languages like Python [22].

Weka: an open-source machine learning and data mining software for data analysis and predictive modeling [23].

Apache Hadoop Distributed File System (HDFS): it is an open-source framework for managing big data as well as supporting big data analytics. It is the primary storage system used by Hadoop applications. [24],[7]

Apache Hadoop MapReduce: is a software framework which provides the interface for applications which process vast amounts of data in-parallel on thousands of nodes hardware in a reliable, fault-tolerant manner. [25],[7]

Hive: a distributed data warehouse that provides SQL-like commands, called HiveQL, to query, manage and analyze massive datasets [26].

Many other software tools for big data like Jaql [27], Zookeeper [28] etc.

In the next section, different tools are used to collect, execute SQL queries and apply machine learning algorithms to datasets concerning diabetes.

5 Using Multiple Data Sources and Software Tools in Data Analysis

Knowledge extraction from data collected in the field of medicine involves many stages and requires special procedures and characteristics from data analysts. The utilization, combination and integration of data derived from multiple sources which have different formats and structure, and the need for using different software tools and platforms increases the complexity of the data analysis process, but at the same time strengthens the data analysis process and provides significant capabilities to extract knowledge by combining many different methods, technologies, and data sources. The difficulties faced by the data analyst are increased and concern the special requirements that must be met for the integration and combination of different data formats, the use of different technologies and applications which are required to cooperate properly and be combined to achieve their interoperability.

The purpose of this section is to briefly present the complexity of this process that a data analyst is called upon to face. Specifically, in the current section, medical datasets from two different data sources are used. To utilize these data sets that come from different sources, it is necessary to structure the appropriate program and build the necessary bridges that will allow seamless communication between the program and the data sources. The programming language used is Python. Through a Python program, data is initially collected from different sources and then fed into the Spark open-source software environment which is used to apply SQL queries to the data. Through this process, the tools that allow the bridging of different software environments are briefly presented, as well as the difficulties and requirements that exist for the interoperability of these environments.

5.1 Using Spark software tool with datasets derived from different locations

For the current research, two different datasets about diabetes disease were used. Diabetes is a serious chronic disease that affects millions of people and places a significant economic burden on the economy.

The purpose of the research is not to perform an extensive analysis of medical data, which would require special knowledge and medical scientists, but to present the technical and programming requirements that exist for data collection and analysis using different data sources and different software tools.

Kaggle dataset

The first dataset is the diabetes health indicators dataset from Kaggle [29]. It is a clean dataset of 253,680 survey responses to the Centers for Disease Control and Prevention. The target variable has three classes, 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. This dataset has twenty-one (21) feature variables and there is class imbalance in this dataset. [29]

Initially, the appropriate libraries, kaggle and pyspark, must be installed. Next, it is necessary to check if the API token for Kaggle, the file kaggle.json, is in the correct folder on your computer. Finally, it is also necessary to install Java. For the current example Eclipse Temurin jdk-17.0.19 [30] was installed because the earlier and later versions were not compatible with the Spark.

The appropriate libraries are imported into the Python program as shown in Figure 3.

```
import os
import sys

# 1. Define path for Java
os.environ["JAVA_HOME"] = r"C:\Program Files\Eclipse Adoptium\jdk-17.0.19.10-hotspot"

# 2. Define System PATH for this notebook
os.environ["PATH"] = os.environ["JAVA_HOME"] + r"\bin;" + os.environ["PATH"]
os.environ["PYSPARK_PYTHON"] = sys.executable
os.environ["PYSPARK_DRIVER_PYTHON"] = sys.executable


from kaggle.api.kaggle_api_extended import KaggleApi
from pyspark.sql import SparkSession
import pandas as pd
import matplotlib.pyplot as plt
```

Fig. 3. Importing kaggle and pyspark in python program

The dataset is then downloaded as shown in Figure 4.

```
print("1. Download data from  Kaggle...")
# Initialize  Kaggle API.
#  kaggle.json file should be in correct path.
api = KaggleApi()
api.authenticate()
dataset_name = 'alexteboul/diabetes-health-indicators-dataset'
print(f"Receiving dataset '{dataset_name}'...")
api.dataset_download_files(dataset_name, path='.', unzip=True)

# When unzip, this is the dataset filename
csv_file_name = 'diabetes_012_health_indicators_BRFSS2015.csv'
```

Fig. 4. Downloading dataset from Kaggle

Next, SparkSession.builder() is used to create a session with Spark. This session will allow the dataset to be imported into Spark using spark.read.csv() method and SQL queries to be executed on the dataset. Before executing queries, it is necessary to create a temporary view of the dataset using the df.createOrReplaceTempView() method. Figure 5 presents the python code.

```
print("\n2. Start  Apache Spark Session...")
# The command creates a Spark Session.

spark = SparkSession.builder \
    .appName("Kaggle_SparkSQL_Analysis") \
    .master("local[*]") \
    .config("spark.sql.shuffle.partitions", "8") \
    .config("spark.driver.memory", "4g") \
    .enableHiveSupport() \
    .getOrCreate()


print("✅  Spark Session initialized succesfully!")

print(f"File is loading  '{csv_file_name}' to Spark...")
# inferSchema=True to automatically recognize  if a column is text, number etc.
df = spark.read.csv(csv_file_name, header=True, inferSchema=True)

# Create a Temporary View to run SQL commands.
# the name is 'my_dataset'.
df.createOrReplaceTempView("my_dataset")

print("\n3. Apply  Spark SQL command...")
```

Fig. 5. Creating Spark session

At this point everything is ready to submit SQL queries to our dataset.

The SQL query is submitted using the spark.sql() method. In the following example, the field Diabets_012, which is the class of the dataset, was used as a grouping field to obtain aggregate results for the number of people belonging to each class, how many of them had a high cholesterol index, consume alcoholic beverages, are smokers, have heart disease, and how many do some physical activity or not. Figure 6 presents the code and Figure 7 the results of the query.

```
# Group data using Aggregation -
query = """
    SELECT Diabetes_012, COUNT(*) as total_count, AVG(BMI), SUM(HighChol), SUM(HvyAlcoholConsump), SUM(Smoker),
    SUM(HeartDiseaseorAttack), SUM(PhysActivity)
    FROM my_dataset
    GROUP BY Diabetes_012
    ORDER BY Diabetes_012
    """

    # spark.sql(query) returns a new  DataFrame with the results.
result = spark.sql(query)

    # Show results
result.show()
```

Fig. 6. Spark SQL query on Kaggle dataset

Diabetes_012	total_count	avg(BMI)	sum(HighChol)	sum(HvyAlcoholConsump)	sum(Smoker)	sum(HeartDiseaseorAttack)	sum(PhysActivity)
0.0	213703	27.742522162548023	81030.0	13216.0	91824.0	15351.0	166491.0
1.0	4631	30.724465558194776	2875.0	208.0	2282.0	664.0	3142.0
2.0	35346	31.944010637695592	23686.0	832.0	18317.0	7878.0	22287.0

Fig. 7. Results of Spark SQL query

Data analyst can apply as many SQL queries as he wants to the dataset, get results and draw the corresponding conclusions. For our data set, we see that 4631 people belong to class 1 (prediabetes). The average body mass index (BMI) for people in class 1 is 30.7, more than half of them (2875 of 4631) have high cholesterol index and 2282 of them are smokers. Class 2 includes 35346 people who have diabetes disease and a higher body mass index (31.94) compared to class 1. From our dataset we observe that most people belong to class 0, that is, they are healthy and do not suffer from diabetes. This information can be visualized (Figure 9) using the appropriate Python libraries and methods, as shown in Figure 8.

```
print("\n4. Create plot  (Matplotlib)...")
import pandas as pd
    # Use .toPandas() to transfer SQL results to Pandas DataFrame
pandas_df = result.toPandas()

    # Create Bar Chart
plt.figure(figsize=(10, 6))

    # Χρησιμοποιούμε τις στήλες 'rating' (στον άξονα X) και 'total_count' (στον άξονα Y)
plt.bar(pandas_df['Diabetes_012'].astype(str), pandas_df['total_count'], color='skyblue', edgecolor='black')

plt.title('Population distribution', fontsize=14, fontweight='bold')
plt.xlabel('Category: 0=no diabetes or only during pregnancy, 1=prediabetes,  2=diabetes ', fontsize=12)
plt.ylabel('Number of rows', fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()
```

Fig. 8. Visualizing SQL results

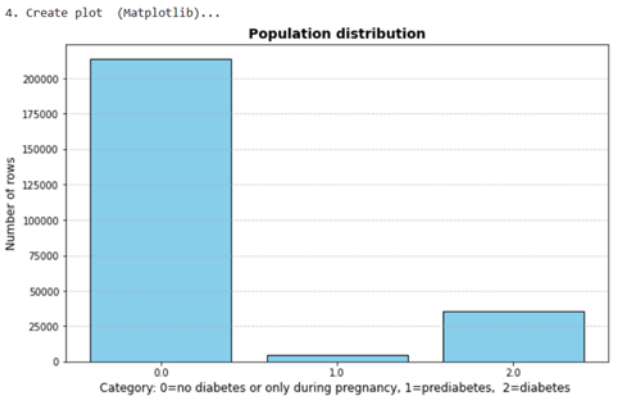

Fig. 9. Population distribution

Mendeley dataset

The second dataset about diabetes disease was derived from Mendeley free and secure cloud-based communal repository. The dataset consists of medical information about patients' Sugar Level Blood, Age, Gender, Creatinine ratio, Body Mass Index (BMI), Urea, Cholesterol (Chol) etc. [31]

To utilize this dataset, the Python programming language was used. Specifically, to download the datafile from Mendeley repository, the request library is necessary, which allows to download the datafile and use it within python program. Code in Figure 10 imports the request library into the program and creates a Spark session.

```
# Use another data source

import requests
spark = SparkSession.builder \
    .appName("Mendeley Diabetes Data Import") \
    .getOrCreate()
```

Fig. 10. Importing requests library and creating a new Spark Session

Then the URL of the dataset is specified as well as the filename (Figure 11) and the dataset is downloaded (Figure 12).

```
# 2. Define URL
MENDELEY_URL = "https://data.mendeley.com/public-files/datasets/wj9rwkp9c2/files/2eb60cac-96b8-46ea-b971-6415e972afc9/file_downloaded"
LOCAL_FILE_PATH = "diabetes_dataset.csv"
```

Fig. 11. Defining the URL of the dataset

```
# 3. Download Dataset
print("Get data from  Mendeley...")
response = requests.get(MENDELEY_URL, stream=True)

if response.status_code == 200:
    with open(LOCAL_FILE_PATH, 'wb') as f:
        for chunk in response.iter_content(chunk_size=1024):
            if chunk:
                f.write(chunk)
    print("File succefully downloaded !")
else:
    raise Exception(f"Download failure. HTTP Status: {response.status_code}")

# 4. Insert  CSV into a  Spark DataFrame
df = spark.read.csv(
    LOCAL_FILE_PATH,
    header=True,
    inferSchema=True,
    sep=","
)
```

Fig. 12. Downloading dataset

At this point we can submit any SQL query we want to our data set.

This dataset contains three different values for the class: the "N" value for those who are not diabetic, the "P" value for those who are at a pre-diabetic level, and the "Y" value for those who suffer from diabetes disease. The following SQL query (Figure 13) presents the total number of people, the average BMI, and the average age for each different class value (Figure 14).

```
# 6. Another Spark SQL example
result = spark.sql("""
    SELECT
        CLASS, COUNT(*) as Total_Patients, AVG(BMI), AVG(AGE)
    FROM diabetes_data
    WHERE CLASS IN ('Y','N','P')
    GROUP BY CLASS
    ORDER BY CLASS
""")

# Show results
result.show()
```

Fig. 13. SQL query

```
+-----+--------------+------------------+------------------+
|CLASS|Total_Patients|          avg(BMI)|          avg(AGE)|
+-----+--------------+------------------+------------------+
|    N|           102|22.352941176470587|44.294117647058826|
|    P|            53|23.933962264150942|43.283018867924525|
|    Y|           840|30.793833333333332| 55.38690476190476|
+-----+--------------+------------------+------------------+
```

Fig. 14. SQL query results

From the results presented in Figure 14, we see that 840 people are patients with an average body mass index (BMI) of 30.79 and an average age of 55.38.

The next SQL query (Figure 15) focuses on patients, that is, those with a value "Y" in class, and presents the number of patients by gender and the average age.

```
# 6. Another Spark SQL query
result = spark.sql("""
    SELECT
        CLASS, Gender, COUNT(*) as Total_Patients, AVG(AGE)
    FROM diabetes_data
    WHERE CLASS='Y'
    GROUP BY CLASS, Gender
    ORDER BY CLASS
""")

# Results
result.show()
```

Fig. 15. SQL query to display number of patients and average age

The results (Figure 16) present that most of the patients are male (486 people) with average age 55.25.

```
+-----+------+--------------+-----------------+
|CLASS|Gender|Total_Patients|         avg(AGE)|
+-----+------+--------------+-----------------+
|    Y|     F|           353|55.56940509915014|
|    Y|     M|           486|55.25514403292181|
```

Fig. 16. Total number of patients and average age

Data analysts can utilize the Spark tool and submit multiple SQL queries to derive useful information from the dataset. The purpose of this section is not to provide an extensive data analysis for the specific dataset but to present the special requirements and technical capabilities provided by the different software tools. The last command in the program is spark.stop() to close the Spark Session and free up system resources.

5.2 Using Weka for Diabetes Prediction

In this section, Weka was used as software tool and three Machine Learning algorithms for creating models for diabetes prediction. Weka (Waikato Environment for Knowledge Analysis) is one of the most popular, free and open-source software for Machine Learning and Data Mining. It was developed at the University of Waikato in New Zealand and is written entirely in Java.

Machine learning is a branch of Artificial Intelligence that, through the development of algorithms and techniques, enables computers to learn and improve from data. Algorithms help identify patterns and create models that can be implemented on new, unknown data to make predictions

Three machine learning algorithms were run using the diabetes dataset derived from Kaggle [29] as training and test file and the corresponding models were created: decision trees, Naïve Bayes classifier and the KNN. The results are presented in the following tables: Table 1 presents the results for decision trees, Table 2 for Naïve Bayes classifier and Table 3 for KNN.

Table 1. Results from decision tree classifier

```
Number of Leaves  :     2916

Size of the tree :      3452


Time taken to build model: 4 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      214629               84.6062 %
Incorrectly Classified Instances     39051               15.3938 %
Kappa statistic                          0.1577
Mean absolute error                      0.1577
Root mean squared error                  0.2849
Relative absolute error                 87.4119 %
Root relative squared error             94.8758 %
Total Number of Instances           253680

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.982    0.870    0.858      0.982    0.916      0.219    0.721     0.910     0
               0.000    0.000    0.000      0.000    0.000      -0.003   0.589     0.027     1
               0.136    0.019    0.537      0.136    0.217      0.220    0.726     0.337     2
Weighted Avg.  0.846    0.736    0.797      0.846    0.802      0.215    0.720     0.814

=== Confusion Matrix ===

     a      b      c   <-- classified as
 209017     50   3836 |      a = 0
   4316      0    315 |      b = 1
  30482     52   4812 |      c = 2
```

Table 2. Results from Naïve Bayes classifier

```
Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      203123               80.0706 %
Incorrectly Classified Instances     50557               19.9294 %
Kappa statistic                          0.3208
Mean absolute error                      0.1526
Root mean squared error                  0.3192
Relative absolute error                 84.5659 %
Root relative squared error            106.284  %
Total Number of Instances           253680

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.865    0.498    0.903      0.865    0.883      0.338    0.806     0.956     0
               0.000    0.000    0.000      0.000    0.000      -0.001   0.692     0.033     1
               0.519    0.140    0.375      0.519    0.435      0.332    0.811     0.388     2
Weighted Avg.  0.801    0.439    0.813      0.801    0.805      0.331    0.805     0.860

=== Confusion Matrix ===

     a      b      c   <-- classified as
 184766     14  28923 |      a = 0
   2915      0   1716 |      b = 1
  16987      2  18357 |      c = 2
```

Table 3. Results from k-Nearest Neighbors algorithm

```
Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      207516               81.8023 %
Incorrectly Classified Instances     46164               18.1977 %
Kappa statistic                          0.1484
Mean absolute error                      0.1498
Root mean squared error                  0.33
Relative absolute error                 83.0212 %
Root relative squared error            109.8655 %
Total Number of Instances           253680

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.941    0.815    0.861      0.941    0.899      0.171    0.687     0.912     0
               0.010    0.006    0.029      0.010    0.015      0.006    0.547     0.021     1
               0.179    0.055    0.346      0.179    0.236      0.166    0.689     0.254     2
Weighted Avg.  0.818    0.694    0.774      0.818    0.791      0.167    0.685     0.804

=== Confusion Matrix ===

     a      b      c   <-- classified as
 201141   1188  11374 |      a = 0
   3975     47    609 |      b = 1
  28609    409   6328 |      c = 2
```

In all three cases the accuracy is greater than 80% and specifically as shown in the corresponding tables, 84.6062% for decision trees, 80.0706% for Naïve Bayes classifier and 81.8023% for KNN. The analyst, in addition to accuracy, also receives from the software a number of metrics such as True Positive Rate (sensitivity), False Positive Rate (specificity), precision, recall, F-Measure etc. that help him evaluate both the overall performance of each model, its reliability and how useful it can be for making predictions, as well as the corresponding metrics for each class separately.

As mentioned earlier, the purpose of this section is not to create the optimal model and perform data analysis using the specific data sets. The purpose of the section is to present the need to use and integrate multiple software tools in data analysis, to present the difficulties that arise in this process and the capabilities that the data analyst has. The following section presents the conclusions from current research as well as issues for further study.

6 Conclusion

The purpose of this paper is not to present the machine learning methods and the evaluation metrics of each model or the functionality of Spark SQL tool, but to give an overall picture of the complexity involved in the analysis of big data in all scientific fields, with an emphasis on the field of healthcare. The challenges and problems for analyzing big data in the healthcare sector are multiple due to the huge volume of data that is collected at a rapid pace, its complexity, its different formats, the interdependencies it contains, as well as the security and accuracy requirements that exist. A multitude of tools have been developed to support all phases of the data lifecycle and help the data scientist discover patterns and trends and create models that will enable reliable predictions.

Health scientists can use the power of data analytics to gain useful knowledge that will help them make decisions. The ongoing studies carried out in the health sector offer useful knowledge and help scientists at a research and therapeutic level, and for this reason it is important to continue research with more reliable and accurate data. At the same time, the role of technology and informatics is crucial for increasingly better support for scientists in the health sector.

In this paper, medical data from two different data sources were used and the method of obtaining them through a Python program was presented, as well as the utilization of different tools and methods. These software tools can be used complementary and help the analyst to address different aspects of a specific problem. The problems that can arise when using multiple tools are many and from the current research it appeared that these mainly concern the bridge that must be made between different applications and software tools. At many stages of this process, the data analyst is forced to delve into the technical specifics and functions of each separate environment to bridge the differences and integrate different tools.

The role of healthcare data analyst is particularly critical, as he is called upon to bridge the gap between technology, mathematics and human health. The data it handles is about patients' lives, clinical trials and improving the healthcare system. To succeed in this field, one needs a combination of hard skills in technology, healthcare domain knowledge and soft skills in communication, critical thinking and problem solving.

This study has presented the specific requirements that exist when integrating different data sources and software tools. Further research can be carried out in this field using more data sources of different degrees of structure, more software tools and machine learning methods.

References:

[1] H. Basim Alwan, K.R. Ku-Mahamud, "Big data: definition, characteristics, life cycle, applications, and challenges IOP", Conference Series: Materials Science and Engineering, 769 (1) (2020), doi:10.1088/1757-899X/769/1/012007

[2] A. Gandomi, M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, Volume 35, Issue 2, 2015, Pages 137-144, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[3] B.Furht, F. Villanustre, "Big Data Technologies and Applications", 2016, Springer International Publishing, ISBN 978-3-319-44548-9, DOI 10.1007/978-3-319-44550-2.

[4] D. Gupta, R. Rani, "A study of big data evolution and research challenges", 2019, Journal of Information Science, 45(3), 322–340. https://doi.org/10.1177/0165551518789880

[5] Adriana Alexandru, Cristina Adriana Alexandru, Dora Coardos, Eleonora Tudora, "Healthcare, Big Data and Cloud Computing," WSEAS Transactions on Computer Research, vol. 4, pp. 123-131, 2016.

[6] Belle A., Thiagarajan R., Soroushmehr SM, Navidi F., Beard DA, Najarian K (2015), «Big data analytics in healthcare». Biomed Research International, Volume 2015, Article ID 370194, http://dx.doi.org/10.1155/2015/370194

[7] Raghupathi W, Raghupathi V., (2014) "Big data analytics in healthcare: promise and potential". Health Information Science and Systems, 7;2:3. doi: 10.1186/2047-2501-2-3. PMID: 25825667; PMCID: PMC4341817.

[8] Guo, C., Chen, J. (2023). Big Data Analytics in Healthcare. In: Nakamori, Y. (eds) Knowledge Technology and Systems. Translational Systems Sciences, vol 34. Springer, Singapore. https://doi.org/10.1007/978-981-99-1075-5_2

[9] Batko, K., Ślęzak, A. (2022) The use of Big Data Analytics in healthcare. J Big Data, 9, 3. https://doi.org/10.1186/s40537-021-00553-4

[10] Geerts, G. L., & O'Leary, D. E. (2022). V-Matrix: A wave theory of value creation for big data. International Journal of Accounting Information Systems, Elsevier, vol. 47(C).

[11] Pouchard, L. (2015). Revisiting the data lifecycle with big data curation. International Journal of Digital Curation 10 (2), 176-192.

[12] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86-94.

[13] Apache storm, https://storm.apache.org/ (accessed 14 of September 2025).

[14] Apache Flume, https://flume.apache.org/ (accessed 14 of September 2025).

[15] Apache Kafka, https://kafka.apache.org/ (accessed 14 of September 2025).

[16] OpenRefine, https://openrefine.org/ (accessed 14 of September 2025)

[17] MongoDB, https://www.mongodb.com/ (accessed 14 of September 2025).

[18] Apache Cassandra, https://cassandra.apache.org/_/index.html (accessed 14 of September 2025).

[19] Apache Spark, https://spark.apache.org/ (accessed 14 of September 2025).

[20] Tableau, https://www.tableau.com/products/desktop (accessed 14 of September 2025).

[21] Power BI https://www.microsoft.com/en-us/power-platform/products/power-bi (accessed 14 of September 2025).

[22] Python, https://www.python.org/ (accessed 14 of September 2025).

[23] Weka, https://ml.cms.waikato.ac.nz/weka/ (accessed 14 of September 2025).

[24] Apache Hadoop Distributed FileSystem (HDFS) https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (accessed 14 of September 2025).

[25] Apache Hadoop Mapreduce https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html (accessed 14 of September 2025).
[26] Hive, https://hive.apache.org/ (accessed 14 of September 2025).
[27] Kevin S. Beyer, Vuk Ercegovac, Rainer Gemulla, Andrey Balmin, Mohamed Eltabakh, Carl-Christian Kanne, Fatma Ozcan, and Eugene J. Shekita. 2011. Jaql: a scripting language for large scale semistructured data analysis. Proc. VLDB Endow. 4, 12 (08/2011), 1272–1283. https://doi.org/10.14778/3402755.3402761
[28] Zookeeper, https://zookeeper.apache.org/ (accessed 14 of September 2025).
[29] Diabetes Health Indicators Dataset https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset File diabetes _ 012 _ health _ indicators _ BRFSS2015.csv, a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. (accessed 31 of May 2026).
[30] Eclipse Temurin jdk-17.0.19 https://adoptium.net/temurin/releases?version=17&os=any&arch=any (accessed 31 of May 2026).
[31] Diabetes Dataset Published: 18 July 2020| Version 1 | DOI: 10.17632/wj9rwkp9c2.1 Contributor: Ahlam Rashid https://data.mendeley.com/datasets/wj9rwkp9c2/1 (accessed 31 of May 2026).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The author has no conflicts of interest to declare that are relevant to the content of this article.