PROJECT SYNOPSIS

Project Title:
Body Fitness Prediction using Random Forest Classifier

Year: 2022-23

Group ID: 7

1. Technical Keywords:

1. Data Engineering.
2. PowerApps.
3. Python.
4. Random Forest classifier.
5. Training Data.
6. Test Data.
7. Cloud Platform.

2. Problem Statement:

The body fitness assessment is essential to living a healthy life. Fitness is a state of health and well-being, more particularly the capacity to carry out everyday tasks. Healthy eating, regular physical activity, and rest often attain body fitness. As a result, we become less physically fit, contributing to several chronic problems.

3. Purpose of the Project

We should track our physical fitness utilizing various fitness prediction tools, such as smartwatches, oximeters, B-P machines, etc., to prevent a number of health complications. We can keep track of things such as our B-P, calories burned, and bone weight. Data interchange between the devices occurs using the Bluetooth communication protocol using smart device technology. In this project, we import the data that includes (date, step count, mood, calories burnt, hours of sleep, bool of activity, and weight in kg) and divide the dataset into the training set and testing set. In this project, we're employing a random forest classifier.

4. Objectives:

1. To apply random forest classifier algorithm to predict the fitness level of body among the students.
2. To monitor student's health state and make them aware about their health.
3. To give free health and fitness tips.
4. To support the personalization of individuals in managing their own health.

5. Review of Conference/Journal Papers and Relevant Theory

Fitness is the body's capacity to adapt to life, exercise, and the environment. It is divided into two categories: fitness for a healthy body and fitness for athletic competition. A person's ability to operate their heart, lungs, vascular system, and muscles effectively to complete daily tasks without becoming overly exhausted is closely related to their level of physical fitness, which allows them to have spare time to enjoy their leisure and entertainment activities and prepare for emergencies. Muscle strength, muscle endurance, flexibility, cardiopulmonary endurance, and body composition are the five components of healthy physical fitness that are related to the effectiveness of daily life or physical activity.

The ability to move around physically is essential for daily tasks. To calculate physical fitness level, physical activity must be measured. This study uses multiple linear regression to create novel prediction models for estimating the physical fitness of secondary schools. The datasets include information from a range of subjects about the goal characteristics, such as test results for the 20-meter stage run, 30-meter sprint, balance, and hand grip (right/left). Gender, age, body mass index (BMI), body fat percentage, and the number of curl-ups and push-ups performed in 30 seconds are the predictor variables used to create the prediction models. With the above-mentioned predictor factors, eight physical fitness prediction models have been developed for each objective. It was suggested that the sensor network would serve as the foundation for the Internet of Things as early as 2010. Each component of the access network is connected to the network for information exchange and resource sharing in accordance with the Internet of Things protocol family laid down by the International Organisation for Standardisation. Realize a highly intelligent network system; the existing Internet is being expanded and extended by the Internet of Things. Wireless networks are typically used by the Internet of Things to facilitate communication. Each person is surrounded by thousands of Internets of Things devices, according to the poll. There could be five trillion to one trillion items in the Internet of Things. Real things are individually numbered by the Internet of Things using electronic tags and radio frequency identification technologies. The Internet of Things allows for the precise location and related data of any object, from a book to a car, as long as it is connected to the Internet. Gerpott's research aimed to create a foundation for assessing the efficacy of IoT-enhanced forecasting. Many specialists are currently thinking about using IoT components to improve forecasting, changing or expanding its scope and algorithm portfolio. However, predictive success is not always achieved when sensors and actuators in existing models are automatically connected in an ad hoc manner. Gerpott's strategy is to identify the many functions that IoT components can play in forecasting algorithms. To clarify the development goals of IoT predictive models and give case examples, it is possible by combining models and algorithms with IoT components. These illustrations assist in demonstrating how IoT components playing various roles can help to realize the development objectives of predictive models. According to Gernot, there could be three main uses when IoT components are incorporated into predictive models.
Understanding how IoT adjustments might assist relevant individuals in achieving predictive model development objectives requires an understanding of the functional distinction. This

research's foundation and starting point are sound and adequate, however, it lacks examples to back it up. Burg held that the creation of fitness prediction algorithms for athletes depends heavily on wireless sensors and actuators connected via the Internet of Things. The transmission connection in this intricate heterogeneous system must adhere to strict energy constraints, provide a high level of security, and meet demanding standards for data throughput, delay, and range. Burg began by summarising wireless communication theory in light of the Internet of Things and the anticipated connectivity requirements of athletes. Before concentrating on the main security concerns and features of such systems, the most pertinent wireless communication standards will be analysed based on these principles. Burg specifically emphasized the need for more thorough research on security issues at all protocol layers, including logical layer security and physical layer security, and highlighted the gap between the security functions in the communication standards used in the Internet of Things and athlete fitness prediction models and their actual vulnerabilities through examples. Although this approach offers illustrations for verification, the illustrations are not sufficiently typical.

To assure the accuracy and reliability of athletes' physical fitness prediction data, Yang implemented several audit programs. He then examined the program and discovered some safety problems. Furthermore, analysis reveals that the data in the predictive model are susceptible to integrity forgery attacks, and malicious cloud servers can carry out the forgery attacks even if there is no correct data storage. This means that individual models cannot maintain the privacy of shared data in cloud storage. Any prediction command can be faked by a malicious cloud server using data that appears to be legitimate. Yang then concluded that the lack of enough shielding in the linear combination of randomly collected data has been the primary cause of the insecurity. Last but not least, Yang suggested an enhancement to the audit program that would bring the best prediction outcomes and calculation overhead while maintaining data privacy and ideal index optimization. Although the research data support is comparatively adequate, the cost of use is high, which hinders the rise in popularity.

6. System Architecture

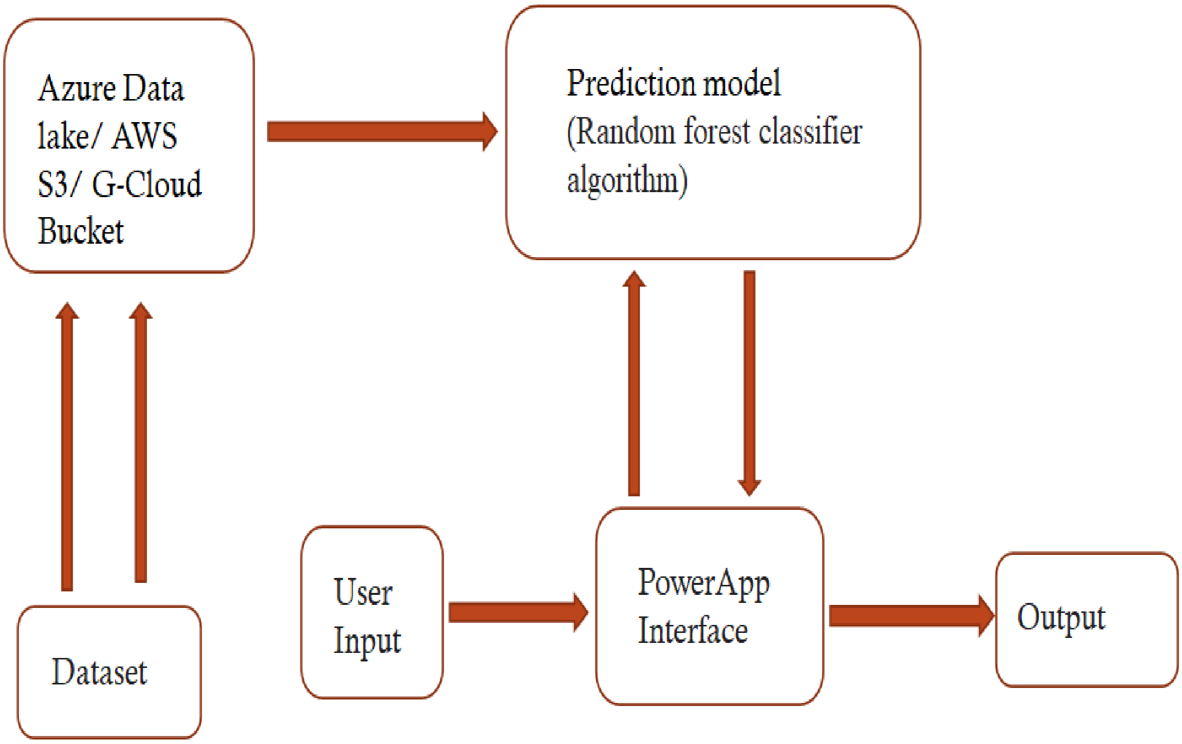

Fig 1. System architecture

7. Modules of Architecture

a. **Dataset:** We will use the body fitness prediction dataset which was retrieved from kaggle.com

1. Examine the relationships between physical activity (measured by counting steps), calorie intake, body weight, number of hours slept, and the perception of being active or inactive.
2. Comparing calorie expenditure based on self-perceived activity and mood (active and inactive)
3. Examine the differences in sleep duration between the mood and self-perceived activity categories (active and inactive)
4. The database comprises 7 columns and 96 observations. "Number of steps' ' (step count), "caloric expenditure" (calories burned), "hours of sleep" (hours of sleep), and "body weight" (weight kg) are its quantitative variables. Additionally, there are qualitative variables for "dates' ' (date), "moods' ' (mood), and "self-perceived activity" (book of active) active or inactive. A value of 300 was given to the variable "humor" to represent "happy," a value of 200 for "neutral," a value of 100 for "sad," and a value of 100 for the variable "self-perceived activity."
5. Contingency tables of categorical variables will be exposed.

b. Azure/AWS/GCloud: We can use different cloud services for VM, storage purposes, and APIs Services

1. **Azure Data Lake:** Developers, data scientists, and analysts may store data of any size, shape, or speed, as well as perform any kind of processing or analytics across platforms and languages, thanks to the comprehensive set of features included in Azure Data Lake.

2. **AWS S3:** An object storage service called Amazon Simple Storage Service (Amazon S3) provides performance, security, and scalability that are unmatched in the market. For a variety of use cases, including data lakes, websites, mobile applications, backup and restore, archives, business applications, IoT devices, and big data analytics, customers of all sizes and sectors may use Amazon S3 to store and preserve any quantity of data.

3. **GCloud bucket:** Any quantity of data can be stored in the cloud and retrieved at any moment from anywhere in the world.You can use cloud storage for a variety of purposes, such as displaying website content, archiving data for disaster recovery, and providing users with direct download access to huge data items.

c. Prediction Model:

Predictive modelling is a commonly used statistical technique to anticipate future results. These are data mining technology solutions that examine both historical and recent data to create a model that predicts behavioural patterns from the data.

d. Training Data:

Training data is the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict. The premise of using training data in machine learning systems is simple, but it is also fundamental to how these technologies operate. The training data is a starting collection of data that programs utilize to learn how to employ cutting-edge technologies like neural networks and produce complex outcomes. Additional data sets referred to as validation and testing sets may be used to complete it.

e. Test Data:

Test data are used to gauge the effectiveness of the algorithm you are using to train the computer, such as its accuracy or efficiency. You need unknown data to test your machine learning model after it has been constructed (using your training data). You can use this data, which is referred to as testing data, to assess the effectiveness and development of your algorithms' training and to modify or improve them for better outcomes.

We use sklearn. ensemble module train_test_split which is used for the training and testing part.

f. Random Forest Classifier Algorithm:

It creates a set of decision trees from randomly selected subsets of the training set. It is basically a set of the decision trees from a randomly selected subset of the training set and then it collects the votes from different decision trees to decide the final prediction

Steps:

1.Select random samples from a given dataset

2.Construct a decision tree for each sample and get a prediction result from each decision tree

3.Perform a vote for each predicted result

4.Select the prediction result with the most votes as the final prediction

g. PowerApp Interface:

Power Apps is a collection of apps, services, connectors, and a data platform that offers a quick development environment for creating unique apps for your company's requirements. You can easily create unique business apps with Power Apps that connect to your data housed in the Microsoft Data-verse underpinning data platform or in a variety of online and on-premises data sources (such as SharePoint, Microsoft 365, Dynamics 365, SQL Server, and so on). Power Apps-created applications offer robust business logic and workflow features to convert your manual business processes into digital, automated ones. Additionally, Power Apps-created apps have a responsive design and can function seamlessly on mobile devices and in browsers (phone or tablet). Power Apps "revolutionizes" the business-app-building process by enabling anyone to build functionality and customized business applications without writing code.

Dependent and Independent variables:

1. Independent variables contain a list of variables on which a book of activity is dependent.
2. The dependent variable is the variable that is dependent on the other variable’s values.
3. Independent variables are mood,step_count, calories burned, hours of sleep, and weight kg.
4. Dependent variables are bool_of_active.

Modules:

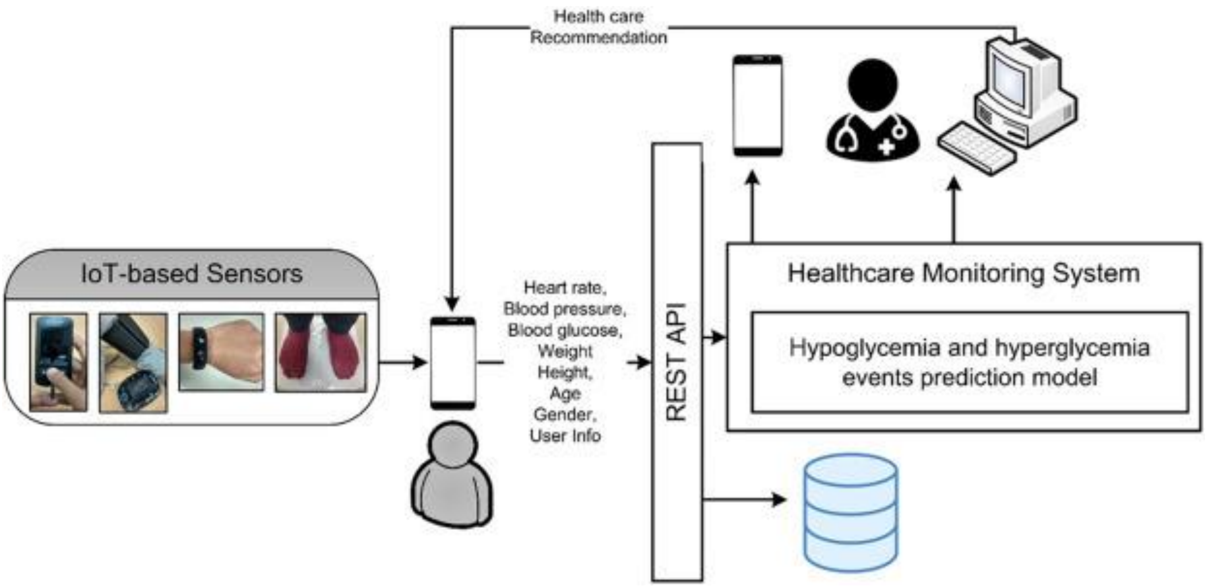

Fig 2. System module

8. System Requirements

8.1 Hardware Requirements

A. Processor: Intel i5 or above, Ryzen 5 or above
B. RAM: 4 GB or above
C. Graphics: Integrated graphics or above

8.2 Software Requirements

A. Operating system: Windows, Linux, or any other
B. Language: Python, SQL, PySpark
C. Notebook (Jupyter)
D. AWS/ Azure Cloud Platform

8.3 Network Requirements

A. APIs
B. Self-Hosted Integration Service

9. Project Plan

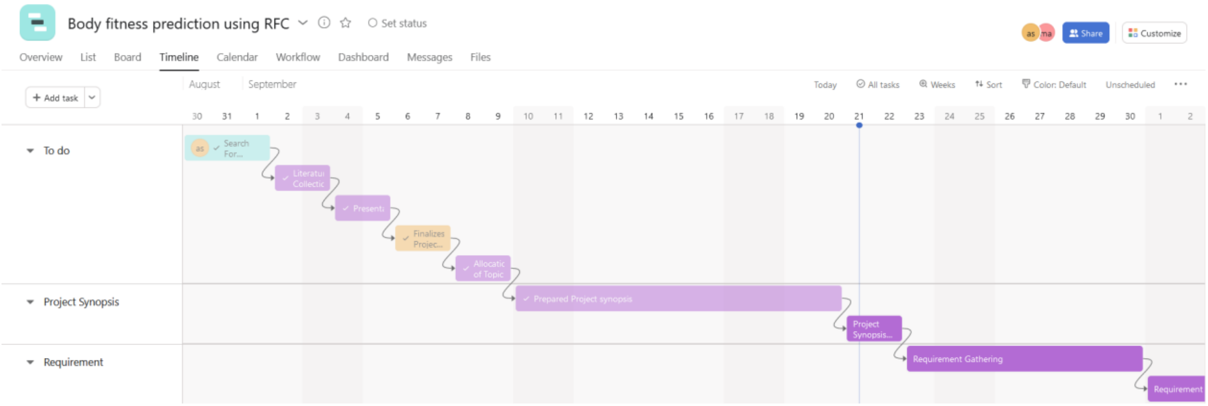

Fig 3. Proposed project plan for sem-I (August-September)

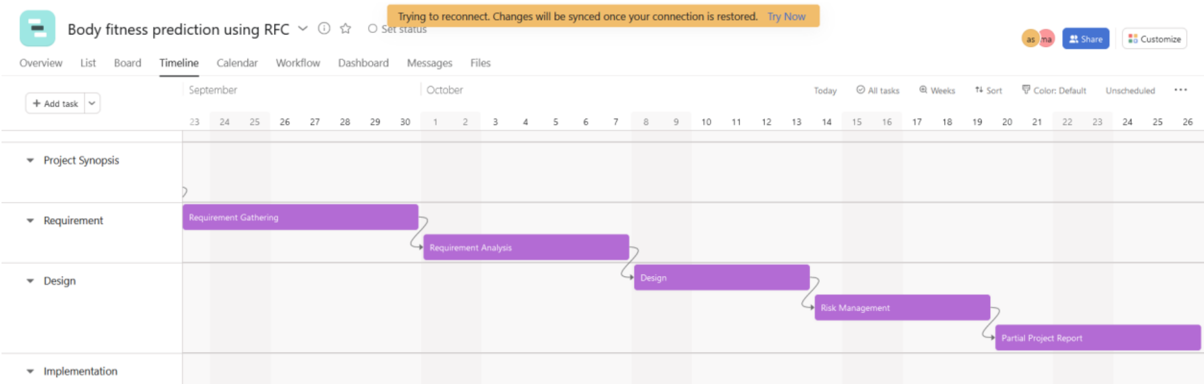

Fig 4. Proposed project plan for sem-I (September - October)

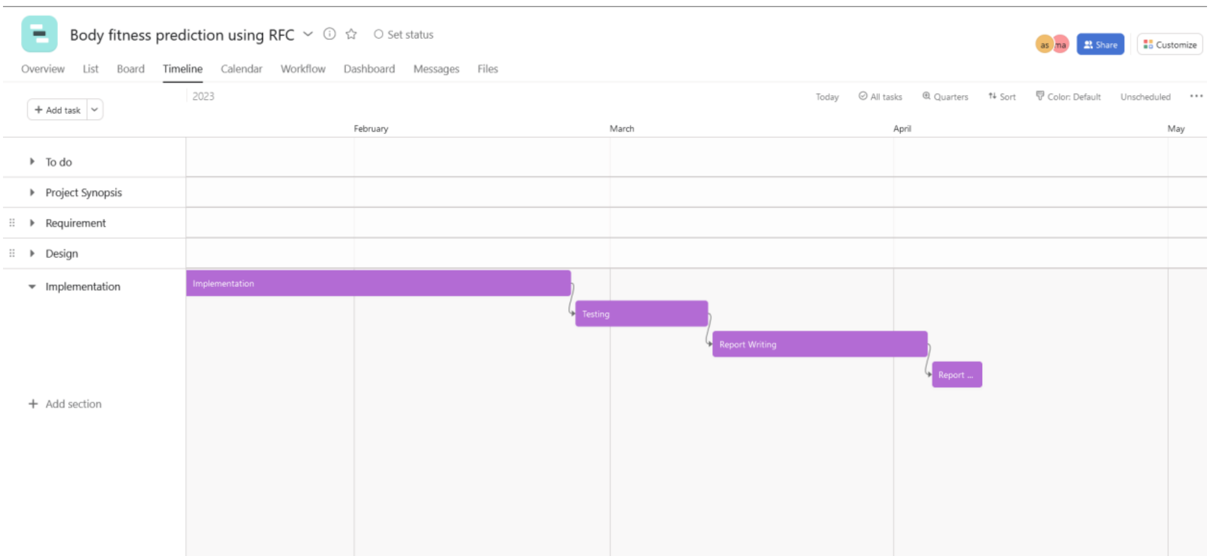

Fig 5. Proposed project plan for sem-II (December - April)

Conclusion:

We will be going to apply Random Forest Classifier algorithm to predict a student's fitness after analysing the data provided by them. We will compare the performance of a Random Forest Classifier and its variants to make predictions. With the help of lowest RMS we will check the better accuracy and an xgBoost Regressor we will help to classify large dataset.

Reference:

1. https://www.hindawi.com/journals/mpe/2021/6680629/
2. https://www.frontiersin.org/articles/10.3389/fpubh.2021.691669/full
3. https://www.researchgate.net/publication/329241584_Multiple_linear_regression-based_physical_fitness_prediction_models_for_Turkish_secondary_school_students
4. https://1000projects.org/body-fitness-prediction-using-random-forest-classifier-project.html
5. https://docs.microsoft.com/en-us/power-apps/powerapps-overview

Sr. No	Roll No	Name of Students	Signature
1	79	Shraddha Dhende	
2	138	Ashutosh Pawar	
3	140	Viraj Rasal	
4	171	Mansi Vadane	

Dr. H. E. Khodke
(Project Guide)

Dr. A. B. Pawar
(Co-ordinator)

Dr. D.B.Kshirsagar
(Head of Comp. dept)