https://doi.org/10.31891/csit-2024-2-4
UDK 004.832.032.6

Nataliia KUNANETS, Orest ZHMURKEVYCH
Lviv Polytechnic National University

INFORMATION SYSTEM FOR DATA PROCESSING IN SPORTS USING THE RANDOM FOREST METHOD

A huge amount of data is collected and generated in modern sports. This data can be used to improve athletes' performance, make more informed coaching and strategic decisions, and increase fan engagement. However, processing, analyzing, and interpreting this data can be challenging. This article is devoted to the development of an information system for data processing in the sports sector using the random forest method. The system aims to ensure efficient collection, processing, and analysis of large amounts of data generated during sports competitions, training, and interaction with fans and other stakeholders.

Research methods. This article proposes an information system (IS) for data processing in the sports industry using the Random Forest (RF) method. As one of the machine learning methods, it is well suited for working with large amounts of data and complex classification and prediction tasks. The proposed IS consists of three main components. The data collection module accumulates data from various sources such as sensors, GPS trackers, websites, and social networks. The data processing module cleans, normalizes, and transforms the data to prepare it for analysis. The data analysis module uses the RF method to analyze data, predict outcomes, identify patterns, and make decisions.

The conducted research has shown that the proposed IS can be an effective tool for predicting the results of sports competitions with high accuracy, identifying patterns in the data that can be useful for coaches and athletes to improve their training and strategy, personalizing training programs and recommendations for athletes, increasing the level of fan engagement by providing them with personalized content and forecasts.

The proposed IS based on the random forest method is a powerful tool for processing and analyzing data in the sports industry. Its use can lead to improved athletes' performance, more informed coaching and strategic decisions, and increased fan engagement.

One of the most powerful and accurate machine learning methods, the random forest method, allows for reliable analysis and forecasting based on various types of data, including player statistics, match results, physiological indicators, and fan behavior data. The article describes the stages of creating an information system: from data collection to data processing, storage, and analysis.

Keywords: information system, sports data, Random Forest method, machine learning, data analysis

Наталія КУНАНЕЦЬ, Орест ЖМУРКЕВИЧ
Національний університет «Львівська політехніка»

ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ ОПРАЦЮВАННЯ ДАНИХ У СПОРТИВНІЙ ГАЛУЗІ ЗА ДОПОМОГОЮ МЕТОДУ ВИПАДКОВОГО ЛІСУ

У сучасному спорті збирається та генерується величезний масив даних. Ці дані можуть бути використані для покращення результатів спортсменів, прийняття більш обґрунтованих тренерських та стратегічних рішень, а також для підвищення рівня залученості вболівальників. Однак обробка, аналіз та інтерпретація цих даних може бути складним завданням. Ця стаття присвячена розробці інформаційної системи для обробки даних у спортивній сфері за допомогою методу випадкового лісу. Система спрямована на забезпечення ефективного збору, обробки та аналізу великих обсягів даних, що генеруються під час спортивних змагань, тренувань та взаємодії з уболівальниками та іншими зацікавленими сторонами.

Методи дослідження. У цій статті пропонується інформаційна система (ІС) для опрацювання даних у спортивній галузі за допомогою методу випадкового лісу (Random Forest, RF). Як один із методів машинного навчання він добре підходить для роботи з великими обсягами даних та складними задачами класифікації та прогнозування. Запропонована ІС складається з трьох основних компонентів. Модуль збору даних накопичує дані з різних джерел, таких як датчики, GPS-трекери, веб-сайти та соціальні мережі. Модуль опрацювання даних очищає, нормалізує та трансформує дані, готуючи їх до аналізу. Модуль аналізу даних, використовуючи метод RF для аналізу даних, прогнозування результатів, виявлення закономірностей та прийняття рішень.

Проведені дослідження продемонстрували, що запропонована ІС може бути ефективним інструментом для прогнозування результатів спортивних змагань з високою точністю., виявлення закономірностей у даних, які можуть бути корисними для тренерів та спортсменів для покращення їхньої підготовки та стратегії, персоналізації тренувальних програм та рекомендацій для спортсменів, підвищення рівня залученості вболівальників шляхом надання їм персоналізованого контенту та прогнозів.

Запропонована ІС на основі методу випадкового лісу є потужним інструментом для опрацювання та аналізу даних у спортивній галузі. Її використання може призвести до покращення результатів спортсменів, прийняття більш обґрунтованих тренерських та стратегічних рішень, а також до підвищення рівня залученості вболівальників.

Один із найпотужніших і найточніших методів машинного навчання – метод випадкового лісу – дозволяє проводити надійний аналіз і прогнозування на основі різних типів даних, включаючи статистику гравців, результати матчів, фізіологічні показники та дані про поведінку вболівальників. Стаття описує етапи створення інформаційної системи: від збору даних до їх обробки, зберігання та аналізу.

Ключові слова: інформаційна система, спортивні дані, метод випадкового лісу, машинне навчання, аналіз даних.

Introduction

Sports have gained significant popularity among both amateurs and professionals in recent years. This has created a need for improved training methods, result analysis, and management of sports events. The rapid

development of technology, particularly in data analysis and artificial intelligence, offers unique opportunities for implementing efficient information systems in the sports sector. The increasing competition in sports generates demands from coaches and athletes to improve strategies, result analysis continuously, and informed decision-making. An information system can become a powerful tool for achieving these goals. Sports organizations and federations are interested in implementing information systems to enhance competition management, monitor the work of coaching staff, and track the physical condition of athletes. An information system can provide sponsors and marketing agencies with valuable analytics on the effectiveness of advertising campaigns, the popularity of sports events, and athletes' success, assisting them in making investment decisions in sports.

These factors highlight the relevance of developing an information system for the sports sector and support the demand for such solutions in this market segment.

State of Research on the Problem

Many professional sports organizations are currently seeking or already utilizing information systems to integrate data from various informational and measurement systems. The challenge lies in the highly heterogeneous requirements for such systems. To determine the functions of information systems, researchers conducted interviews with representatives of different software products. Analysis of the results indicated that there are characteristics common to all information systems proposed for use in the sports sector. However, depending on the type of sport, specific functionalities are required, which led the article's authors to propose a classification of such systems, though they immediately note that a more differentiated classification is necessary. This will help product developers identify missing features in information systems so that they can adapt their products. An information system's important features are modularity, integrated analysis, the availability of content in a user-friendly format, and ease of use [1].

The article's authors [2] justify the feasibility of integrating modern computing and data analysis to create profiles of elite athletes and form sports analytics. A four-component process—descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics—is proposed to structure information and correlate it with athletes' individual responses. According to researchers, this approach will contribute to improving the health preservation system for elite athletes by enhancing methods and means of monitoring well-being and analyzing athletes' results.

The publication [3] noted that analyzing information from various sources in professional football clubs is necessary to make informed decisions regarding team management. For effective performance of this task, it is advisable to use a club's central information management system, which provides football club staff with the information needed to make effective decisions.

The article [4] notes that information systems are increasing in professional sports clubs. Various software products from different suppliers are used, leading to a fragmented situation in sports. Therefore, a standardized and independent overall concept of a club information system is needed. The authors proposed a general concept of an information system for the sports sector, using methods and models from existing fields, especially business analytics. The example of the Liverpool football club illustrates the practical relevance of such an information system. The article formulates the requirements for information systems in the sports sector and develops a concept for its architecture. The research results indicate the complexity of studying this interdisciplinary topic.

The authors [5] note that management, modeling, and comprehensive data analysis have become key factors for innovation in various fields. Two main directions are developing in sports: sports forecasting and sports analytics. Sports forecasting is based on the use of historical data, which allows for the development of models that can accurately predict sports events, including the outcomes of sports competitions, the position of a team in a league, and the score of a match or game. Sports analytics is used to evaluate the performance of professional athletes and teams. The work presents a canonical data analysis model. The authors tested this approach to strategy determination in football.

The article [6] explores the use of cloud computing in the sports sector, focusing on aspects such as tracking athlete performance, fan engagement, event management, and sports marketing. The researchers analyzed the application of cloud computing in the sports industry.

The article [7] provides an overview of big data analysis tools in the sports sector, focusing on methods for analyzing and applications for processing them.

Researchers [8] analyzed a developed sports management information system as an electronic board. The system is designed to analyze and process information regarding usability and accuracy.

The article [9] discusses the use of information technologies to improve the efficiency of sports management and provides an overview of how information technologies are changing the nature of sports management processes in modern society.

Aim and Objectives of the Article

This article analyzes the features of developing an information system for storing and processing data in the sports sector. It focuses on identifying the benefits and opportunities provided by the implementation of such a

system for various participants in the sports environment, from coaches and athletes to sports organizations and fans.

Objectives of the Article:

Analyze the needs and requirements of the sports sector for a data storage and processing information system.

Examine the advantages of implementing an information system for athlete training and development, enhancing the efficiency of coaching work, and managing sports events.

Analyze the potential of using the system to support sports organizations and sponsors in making informed decisions regarding investment and marketing strategies.

Highlight the main aspects of developing, implementing, and maintaining an information system in the sports sector.

Emphasize the importance of such a system's flexibility and adaptability to the changing needs of the sports environment and rapidly evolving technologies.

Problem Statement

In the world of digital technologies, developing an information system for storing and processing data in the sports sector is becoming increasingly relevant. Such an information system will enable coaches and athletes to collect, store, and analyze data on training, physical condition, injuries, and other aspects, aiding in improving the training process, enhancing training efficiency, and achieving better results.

The information system will facilitate the monitoring, storing, and analyzing data on athletes' health, blood pressure indicators, and injuries, allowing for timely identification and prevention of issues. Data on athletes' performance, competition results, and tactical decisions can be collected and analyzed to develop more effective game or training strategies.

Moreover, the information system will allow for a more detailed data analysis using various algorithms and machine learning methods, helping coaches and athletes better understand their performance and refine their strategies. The system can also assist in organizing sports events, managing resources, scheduling, registering participants, and more.

Therefore, developing an information system for storing and processing data in the sports sector will help improve the training process, enhance performance, and achieve new heights in sports.

Requirements for the Information System

To ensure the specified functionality, the information system for storing and processing data in the sports sector must meet several requirements:

- Efficient data storage;
- Convenient and fast access to data;
- Analytical capabilities;
- Integration with other systems;
- Data security;
- Scalability and efficiency;
- Mobility;
- Support for various sports.

The information system should allow users to quickly and conveniently access necessary information through a user-friendly, intuitive interface. It should also offer data analysis tools, including statistical analysis, result forecasting, trend identification, and the detection of correlations between different parameters. The system should integrate with other information systems, such as club or sports facility management systems, electronic broadcasting systems, and external data sources.

The system must ensure a high level of confidentiality, integrity, and availability of data, including measures for encryption, authentication, and access control. It should handle large volumes of data efficiently and provide quick access even under intensive use. Users should be able to use the information system on mobile devices to access information from any location at any time. Due to its flexibility and adaptability, the information system can be used for various sports, including individual and team sports, winter and summer sports, and sports for people with special needs.

The information system is designed to store diverse data types, including athletes' personal data, competition progress and results, player statistics, training and competition schedules, and more. Implementing these requirements will ensure the effective and reliable operation of the information system for the sports sector.

The context diagram illustrates the overall view of the information system, its boundaries, interactions with external entities (actors), and the main data flows. Let's consider an example of an information system that processes data about sports events, teams, players, and fans.

Main components of the diagram:

Information System (IS)

External entities (actors): Administrators, Teams, Players, Fans, Event Organizers

External data sources (other sports databases, social networks)

Main data flows: Input/update of data about teams, players, and events; Generation of analytical reports; Providing data for fans (e.g., match schedules, results); Collecting feedback from fans; Synchronization with external data sources.

Description of the sports information system diagram:

The main data processing center interacts with various external entities.

External Entities:

Administrators. Responsible for entering and updating data in the system, configuring system parameters, and creating reports.

Teams. Send information about their rosters, training, and match results.

Players. Enter and update personal data, view statistics, and receive analytics.

Fans. Access match schedules, results, and news and provide feedback to the information system.

Event Organizers. Provide information about sports events, schedules, and logistics.

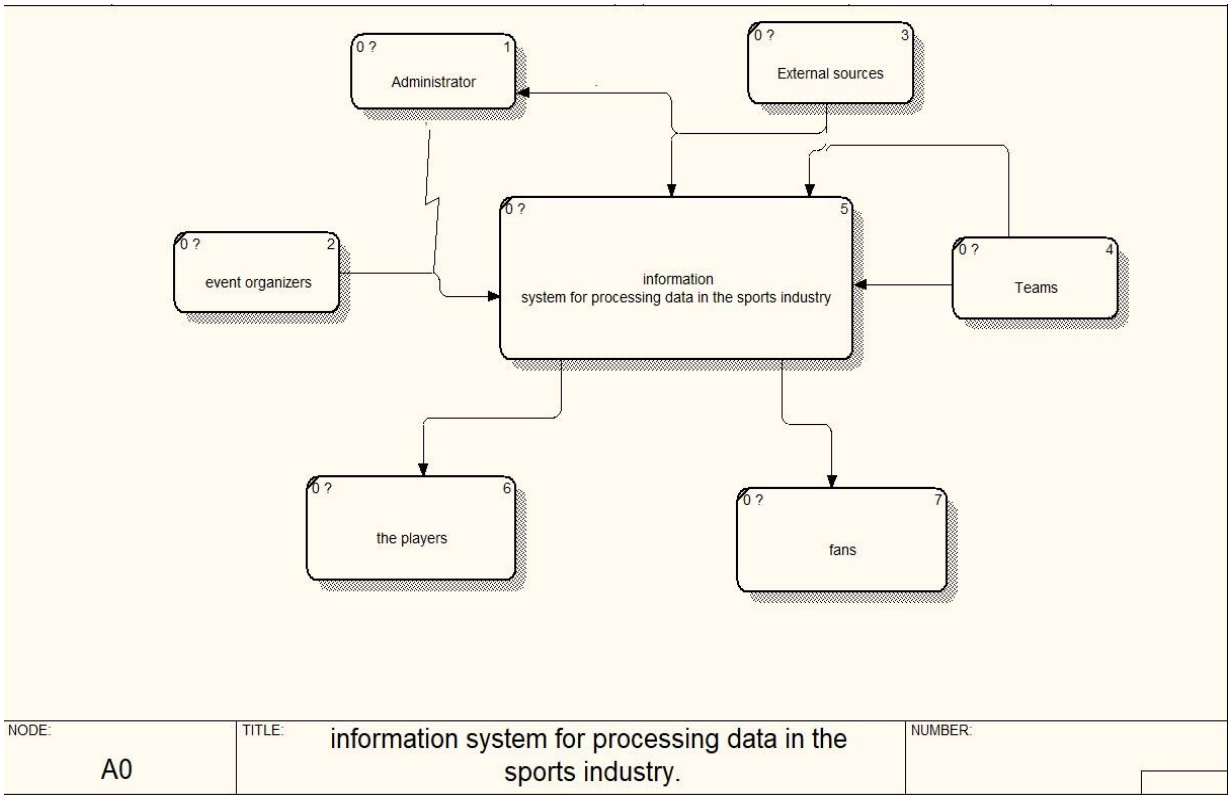

Fig.1 Context Diagram of the Information System for Processing Data in the Sports Sector

External Data Sources:

Integration with other databases and social networks to obtain up-to-date information.

Main Data Flows:

From administrators to IS - Data entry and updates.

From teams to IS - Information about teams and match results.

From players to IS - Personal data updates, statistics.

From event organizers to IS - Information about events and schedules.

From IS to fans - Schedules, results, news.

Feedback from fans to IS - Comments, ratings, requests.

Synchronization with external data sources - Data updates from other sources.

This context diagram overviews the information system, its interactions with key external entities, and main data flows. It serves as a starting point for detailed modeling and system design.

The proposed system includes modules for:

Data collection includes automated interfaces to integrate with various data sources, such as sports databases, social networks, and IoT devices.

Data pre-processing provides cleaning and normalization of data to ensure its quality and compatibility.

Data analysis is based on the random forest method for classifying, regression, predicting sports results, and detecting hidden patterns and trends.

Data visualization, which includes interactive tools for presenting analysis results in the form of graphs, charts, and reports, allows users to interpret the obtained data easily.

User interaction features user-friendly interfaces for administrators, coaches, players, and fans that allow real-time access to the information they need.

Use of the Random Forest Method

A wide range of information technologies can be utilized for data processing in the sports sector, and information systems can be developed based on them. Among these, database technologies, cloud technologies, the Internet of Things (IoT), analytics and machine learning, web technologies, and data visualization are particularly notable. In the sports sector, it is advisable to use various data analysis and machine learning methods to achieve different goals. Each method has advantages and disadvantages, and the choice of a specific method depends on the task and the data characteristics.

This method facilitates classification and regression by constructing many decision trees with random samples of features. In sports analytics, it can be applied to predict match results or identify key factors that influence results.

We will give an example of the scenario of using the Random Forest method in sports analytics: Let's imagine that we analyze the results of football matches and try to predict the results of future games. We have a data set that contains features such as team rank, player age, number of wins and losses in previous matches, average rating of players, number of goals scored and conceded, etc. The analysis takes place according to the following algorithm:

Step 1. Data preparation.
Step 2. Division into training and test sets.
Step 3. Model training.
Step 4. Evaluation of the model.
Step 5. Using the model for forecasting.
Step 6. Updating and improving the model.

First, we collect and clean the data on football matches, ensuring it is structured and accurate. We divide our dataset into training and test sets to train and evaluate the model. We use the Random Forest method to build a model that considers various characteristics of teams and players to predict match outcomes. After training, we assess the model's effectiveness using the test dataset by comparing the model's predictions with the actual match results. With the trained Random Forest model, we can predict the outcomes of future matches, taking into account factors such as team lineups, previous results, player form, and more. After each match, we can update our model by adding new data and adjusting parameters for even more accurate predictions.

Implementing these steps will enable the effective use of the Random Forest method in sports analytics, leading to more accurate predictions and better strategic decisions.

The implementation of the Random Forest method in sports analytics can be presented as follows:

Training the Random Forest Model

Each tree in the forest will select a subset of data for training (with replacement) and randomly select a subset of features for each split.

Let N be the number of trees in the forest, n be the number of data points in the subset, and m be the number of features in the subset.

For each tree k in the forest:

Select a random subset of data D_k of size n.

Select a random subset of features F_k of size m.

Build a decision tree using D_k and F_k.

Predicting with the Forest

For each tree k in the forest:

Make a prediction using the tree for a new observation.

Average (or weight) the predictions or use voting or regression across all trees to obtain the forest's final prediction.

The classification prediction $\hat{y}$ for an object x can be given as:

$$\hat{y}=\frac{1}{N}\sum_{k=1}^{N} N_k\,(x) \quad (1)$$

where $y_k(x)$ is the prediction from the k-the tree.

This approach ensures that our Random Forest model can handle the complexity and variability of sports data, providing reliable predictions and insights.

$$\hat{y} = \text{mode}(\hat{y}_1, \hat{y}_2, \ldots \hat{y}_k), \quad (2)$$

Where $\hat{y}_k$ is the prediction obtained from tree *k*, and mode is a function that returns the most frequently occurring value among them.

This is a general formula for implementing the Random Forest method in sports analytics. Specific parameters, such as the number of trees in the forest, feature selection criteria, etc., may vary depending on the specific implementation and data analysis needs.

Considering that the Random Forest method is a powerful and versatile machine learning algorithm, let's consider an example of applying this method to hypothetical data for classifying types of sports activities.

Suppose we have a dataset containing the following features for each sports event:

Duration of the activity (in minutes).

Number of calories burned during the activity.

Heart rate during the activity (in beats per minute).

Type of activity (e.g., running, swimming, cycling).

Now we want to train a Random Forest model to classify types of sports activities based on these features. Below is an example Python code for this task, using the Scikit-learn library:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Let's assume that sport_data - is a dataset with described features of sports activities

# Splitting the data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(sport_data[['duration', 'calories', 'heart_rate']], sport_data['activity_type'], test_size=0.2, random_state=42)

# Model Initialization and Training Random Forest
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predicting types of activities for the test dataset
y_pred = rf_model.predict(X_test)

# Evaluation of prediction accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

In this example, we initialize and train a Random Forest model on the training dataset and then predict the types of sports activities for the test dataset. The output provides an assessment of the model's prediction accuracy.

This is just an example demonstrating the application of the Random Forest method for classifying sports activities on hypothetical data. Since Random Forest is a machine learning algorithm, its application without using code would be challenging, but we'll still provide a calculation example for classifying types of sports activities using this method.

Suppose we have a dataset containing the following features for each sports event:

Duration of the activity (in minutes).

Number of calories burned during the activity.

Heart rate during the activity (in beats per minute).

Type of activity (e.g., running, swimming, cycling).

We also have a training dataset with these features, consisting of a certain number of examples of sports activities already classified by activity type according to the following algorithm:

Step 1. Initialize the Random Forest with a certain number of trees.

Step 2. We randomly select a subset of the training data for each tree with return.

Step 3. At each node of each tree, we randomly select a subset of features with a certain criterion.

Step 4. We split the data at each node in such a way as to maximize purity (or minimize uncertainty) at each node.

Step 5. Repeat steps 2-4 for each tree.

Step 6. When all trees are trained, we use them to classify new observations.

This is a general description of how the Random Forest method works without a detailed explanation of the calculations performed, as they are quite complex and require the use of algorithms and statistical methods, such as calculating node purity and constructing decision trees.

Forming the training dataset for the Random Forest method in sports analytics involves several steps at the initial stage. It begins with collecting data about sports events, including activity duration, calorie count, heart rate, and activity type. This data can be obtained using specialized sensors, sports watches, mobile applications, or

collected manually. Next, it is advisable to label each data record in the training dataset according to its class or activity type, such as "running," "swimming," "cycling," etc. Here is the algorithm for forming the training dataset for the Random Forest method:

Step 1. Splitting the dataset.
Step 2. Training the model.
Step 3. Model evaluation.
Step 4. Model optimization.
Step 5. Model application.

In the first step, the obtained training dataset is divided into two parts: the training set and the validation set. The training set will be used to train the model, while the validation set will be used to check its effectiveness. We train the Random Forest model using the training dataset based on the provided features and their classes. After training the model, we evaluate its effectiveness using the validation dataset. We analyze how accurately the model classifies new data and determine performance metrics such as accuracy, sensitivity, specificity, etc. We change the model and Random Forest parameters to improve its effectiveness if necessary. After successfully evaluating and optimizing the model, it can be applied to classify new data about sports activities.

The algorithm for collecting data about sports events may look as follows:

Step 1. Defining data collection goals.
Step 2. Choosing data sources.
Step 3. Setting data collection parameters.
Step 4. Initiating data collection.
Step 5. Saving the data.
Step 6. Analyzing and utilizing the data.

Defining the goals and scope of application of the collected data may involve measuring physical activity, monitoring health status, or analyzing training effectiveness. Data sources may include sports watches, fitness trackers, mobile sports applications, or sensors monitoring physical activity. Configuring data collection parameters according to specific needs may include measuring activity duration, calorie count, heart rate, and other parameters based on the available data source capabilities. The data collection process involves setting the necessary parameters and verifying the correct operation of sensors or devices. The collected data is saved in an appropriate format for further analysis and processing. This could be a database, a text file, or any other tool convenient for further work.

Conducting analysis of the collected data to obtain useful information may include identifying trends, establishing relationships between different indicators, and making decisions based on this information.

Here's an example scenario of bootstrapping for each tree in a random forest:

Let's start with a training dataset containing input data intended for training and their corresponding target values. We set parameters and determine the number of trees to create in the random forest (for example, 100 trees) and the size of each random subset (for example, 70% of the total data). When creating the random forest, we randomly select a subset of data with replacement (bootstrapping) from the training dataset for each tree. We build a decision tree on the selected subset of data. When forming the random forest, we store each constructed tree in the random forest. When using the random forest to classify new examples or predict target values, we classify or predict using each tree in the random forest. We store the classification or prediction results from each tree. For classification, we use multi-format voting or average values for predictions. This approach allows each tree in the random forest to "see" only a certain portion of the data, which helps avoid overfitting and ensures greater model robustness.

Below is an example of Python code for creating a random forest using the library.
scikit-learn:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification

# Generating synthetic data for model training.
X, y = make_classification(n_samples=1000, n_features=20, random_state=42)

# Initializing the random forest model with 100 trees.
clf = RandomForestClassifier(n_estimators=100, random_state=42)

# Training the model on the training data
clf.fit(X, y)
```

Now the trained model can be used for classifying new instances. The example demonstrates generating synthetic data for model training, initializing and allowing the training of a random forest model, which is then utilized as the trained model for classifying new data.

Here's the algorithm for implementing the scenario of random feature selection at each node of each tree:

Step 1. Data Preparation.

Step 2. Initializing the random forest model.
Step 3. Training the model.
Step 4. Random feature selection at each node of each tree.
Step 5. Training and evaluating the model.
Step 6. Applying the model to new data.

To implement the algorithm, datasets are loaded and prepared for model training, including features and target variables. Using machine learning libraries such as sci-kit-learn, random forest models are initialized. Prepared data is then passed to the random forest model, and it is trained. At each tree, a random subset of features is chosen for consideration at each node. This process occurs automatically within the random forest algorithm. The model is trained on prepared data and its effectiveness is evaluated using a validation dataset. If the model demonstrates satisfactory performance, it is used to classify new data.

The steps provided offer a general outline of the process for implementing this scenario. Specific details may vary depending on the characteristics of the dataset.

Results & Discussion

We will analyze the data using the example of monitoring an athlete's rehabilitation during a week. We collect data on the number of steps taken by the athlete every day, as well as additional data, such as day of the week, weather, type of training, feeling of fatigue, etc. At the same time, data cleaning, filling of missing values, normalization and coding of categorical variables can occur.To visualize the obtained data, we will apply a diagram that reflects the change in the number of steps throughout the week. Let's imagine we collected data on the number of steps a person took daily throughout the week (Table 1).

We have data on the number of steps throughout the week. Now, let's create a diagram that visualizes this data. The diagram will graphically depict how the number of steps changed daily throughout the week. On the X-axis, the days of the week (Monday, Tuesday, etc.) and the number of steps on the Y-axis will be displayed. For each day of the week, a corresponding marker will indicate the number of steps, for example.

This diagram will allow for easy comparison of the number of steps on different days of the week and identify, for example, which day had the highest number of steps. Such visualization will be a valuable tool for data analysis and identifying dependencies between activity and days of the week. Days with more physical activity will have more steps. For example, on Wednesday and Friday - days with active training - there will be more steps, and on Monday - recovery day - there will be fewer steps.

Table 1.

Number of steps on the corresponding day of the week

Weekday	Number of steps
Monday	5000
Monday	5200
Monday	5100
Tuesday	6000
Tuesday	6100
Tuesday	6050
Wednesday	7000
Wednesday	7100
Wednesday	7200
Thursday	5500
Thursday	5600
Thursday	5650
Friday	8000
Friday	8100
Friday	8200
Saturday	7500
Saturday	7550
Saturday	7600
Sunday	7000
Sunday	7050
Sunday	7100

Fig. 1 Visualization of collected data

To analyze an athlete's activity using a random forest model, a simple model can be created to predict the number of steps based on the day of the week. The given data can be used to train the model and evaluate its performance. First, you need to prepare the data to train the model. Since we have categorical data (day of the week), we need to encode it into a numeric format.

Monday = 1
Tuesday = 2
Wednesday = 3
Thursday = 4
Friday = 5
Saturday = 6
Sunday = 7

According to the data in Table 1, we will train the random forest model. Let's train the random forest model on the input data of Table 1. We receive forecasted data.

Table 1.

Table of projected data

Weekday	Projected data
1	5374.94
1	6030.82
1	6030.82
2	6030.82
2	6962.10
2	6962.10
3	6962.10
3	5610.59
3	5610.59
4	5610.59
4	7929.33
4	7929.33
5	7929.33
5	7719.55
5	7719.55
6	7719.55
6	7085.14
6	7085.14
7	7085.14
7	5374.94
7	6030.82

We will use Python and the scikit-learn library to create a random forest model.

Incoming data

```
extended_data = {
   'День тижня': [1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7],
```

```
    'Кількість кроків': [5000, 5200, 5100, 6000, 6100, 6050, 7000, 7100, 7200, 5500, 5600, 5650, 8000, 8100, 8200,
7500, 7550, 7600, 7000, 7050, 7100]
}

# Створення DataFrame
extended_df = pd.DataFrame(extended_data)

# Вхідні дані (X) та вихідні дані (y)
X_extended = extended_df[['День тижня']]
y_extended = extended_df['Кількість кроків']

# Розділення даних на тренувальну та тестову вибірки
X_train_ext, X_test_ext, y_train_ext, y_test_ext = train_test_split(X_extended, y_extended, test_size=0.2,
random_state=42)

# Створення моделі випадкового лісу
model_ext = RandomForestRegressor(n_estimators=100, random_state=42)

# Навчання моделі
model_ext.fit(X_train_ext, y_train_ext)

# Прогнозування на тестовій вибірці
y_pred_ext = model_ext.predict(X_test_ext)

# Оцінка моделі
mse_ext = mean_squared_error(y_test_ext, y_pred_ext)
r2_ext = r2_score(y_test_ext, y_pred_ext)

# Прогнозування для всіх днів тижня
predictions_ext = model_ext.predict(X_extended)
extended_df['Прогнозовані кроки'] = predictions_ext

mse_ext, r2_ext, extended_df
```

After running this code, we get model performance metrics. Based on the data and training of the random forest model, we obtained the following performance metrics:

- Mean Squared Error (MSE): 58055.77
- R^2 Score: 0.9563

The MSE value of 58055.77 is sufficiently small, which indicates the model's sufficient accuracy. The predicted values' deviation from the actual values is relatively small. The R2 value of 0.9563 indicates that the model explains the variability of the data well. A value close to 1 means the model's high accuracy. The random forest model performed significantly well. A low MSE value and a high R2 Score indicate that the model can accurately predict an athlete's step count based on the day of the week. This shows that the random forest model is adequate for analyzing the number of steps per week, given sufficient data and informative features.

The number of steps an athlete takes daily depends on many factors, such as weather, day of the week, training plan, physical condition, etc. A random forest does a good job of modeling complex relationships between these features. The number of steps can vary greatly from day to day. A random forest is able to handle such irregular data and make stable predictions. A random forest can identify which factors most influence the number of steps taken, which can be useful for training analysis and planning. A random forest can effectively handle missing data by filling it in based on information from other trees in the forest. Using the model to predict the number of steps and analyze the importance of features will allow us to identify the key factors affecting the athlete's activity.

The scientific novelty of the article lies in considering a comprehensive approach to developing and implementing an information system for storing and processing data in the sports industry. The article proposes updated methods and tools for collecting, analyzing, and utilizing data to improve sports event management, enhance training effectiveness and athlete development, and improve interaction between sports organizations and fans. The article's innovation lies in applying advanced data analysis methods, including machine learning techniques and artificial intelligence, to identify patterns and trends in sports activities. Additionally, the article proposes new approaches to using data to support decision-making in the sports industry, which can benefit coaches, athletes, sports organizations, and other stakeholders in the sports environment. Therefore, the scientific novelty of the article lies in proposing advanced solutions and methodologies for developing sports analytics and optimizing management in the sports industry through information technologies.

Conclusions

This paper presents an information system for data processing in the sports industry using the random forest method, which allows efficient analysis of large volumes of data and provides accurate predictions. Research and development of the system made it possible to draw the following conclusions:

The random forest method is highly accurate in classifying and regression of sports data, which makes it particularly useful for predicting match results, evaluating player performance, analyzing fan behavior, and analyzing training results. The model's work is demonstrated by analyzing the results of an athlete's training, and an example of predicting these indicators, which should be achieved to improve his condition, is given. The experiment proved that the input sample's growth improves the model's quality.

The use of advanced machine learning techniques, including random forests, in sports analytics will enable coaches, team managers, and sports analysts to gain deeper insights, facilitating informed decision-making.

The developed information system will provide convenient access to analytical data for various categories of users, including administrators, coaches, players, and fans.

Data preprocessing modules, including cleaning and normalization, significantly improve the output data quality, increasing the accuracy and reliability of predictions. This helps to avoid many errors related to inaccurate or incomplete data.

Based on the results obtained, the proposed information system, which uses the random forest method, effectively processes sports data. This improves the accuracy of data analysis and data prediction. Further research and implementation of additional features may further expand the application of this system in the sports industry.

To improve the results in future studies, the model will be trained on a larger amount of data, which will allow it to better generalize and make more accurate predictions. Additional features such as weather, training time, an athlete's physical condition, etc., can significantly improve the model's performance.

References

1. Blobel Thomas, Rumo Martin, Lames Martin. Sports Information Systems: A systematic review. *International Journal of Computer Science in Sport*. 2021. Vol.20. P.1-22. Doi.10.2478/ijcss-2021-0001.

2. Exel Juliana, Dabnichki Peter. Precision Sports Science: What Is Next for Data Analytics for Athlete Performance and Well-Being Optimization?. *Applied Sciences*. 2024. Vol.14. P.3361. Doi.10.3390/app14083361.

3. Blobel Thomas, Lames Martin. Information Systems for Top-Level Football with Focus on Performance *Analysis and Healthy Reference Patterns*. 2018. P.71-81. Doi 10.1007/978-3-319-67846-7_7.

4.Blobel Thomas, Lames, Martin. A Concept for Club Information Systems (CIS) - An Example for Applied Sports Informatics. *International Journal of Computer Science in Sport*. 2020. Vol.19. P.102-122. Doi.10.2478/ijcss-2020-0006.

5. Marc Garnica Caparrós Cologne Information systems in analytical applications in sports - A data modeling perspective Doctoral thesis accepted for the degree Ph.D. *Natural Science*. 2023.46 p.

6. Xiao L., Cao Y., Gai Y. et al. Review on the application of cloud computing in the sports industry. *J Cloud Comp.* 2023. Vol.12. P.152. doi.org/10.1186/s13677-023-00531-6.

7. Zhongbo Bai, Xiaomei Bai Sports Big Data. *Management, Analysis, Applications, and Challenges, Complexity*. 2021. Vol. 2021, Article ID 6676297, 11 p. doi.org/10.1155/2021/6676297

8. Charmine Rose R. Peñasales, Kristine T. Soberano, Rolan O. Algara E-Board Sports Management Information System with SMS Support. *Usability, Maintainability, Accuracy International journal of multidisciplinary research and analysis.* 2023. Vol. 06, Issue 05. Page 1878. Doi. 10.47191/ijmra/v6-i5-09,

9. Li C., Wang Z. Research on the Applications of Information Technology in Sport Management. In: Qu, X., Yang, Y. (eds) Information and Business Intelligence. IBI 2011. *Communications in Computer and Information Science*. 2012. Vol 268. Berlin, Heidelberg: Springer. Vol.7.-P.9-37. https://doi.org/10.1007/978-3-642-29087-9_37

Nataliia KUNANETS **Наталія КУНАНЕЦЬ**	Doctor of Science in Social Communications, Prof., Lviv Polytechnic National University, Ukraine, 12 S. Bandera Street, Lviv, Ukraine, 79013 E-mail: nataliia.e.kunanets@lpnu.ua https://orcid.org/0000-0003-3007-2462 Contact tel.: 38 098 57 131 88 ResearcherID:R-5222-2017 Scopus Author ID: 57189375884	професор, доктор наук із соціальних комунікацій, професор кафедри інформаційних систем і мереж Національного університету «Львівська політехніка», Львів, Україна
Orest ZHMURKEVYCH **Орест ЖМУРКЕВИЧ**	Master Degree Student Department of information systems and networks Lviv Polytechnic National University Bandera str., 12, Lviv, Ukraine, 79013 E-mail: orest.zhmurkevych@gmail.com https://orcid.org/0000-0002-5227-7138	магістр кафедри інформаційних систем і мереж Національного університету «Львівська політехніка», Львів, Україна