DATA VISUALIZATION FOR VOICE GENDER CLASSIFICATION USING XGBOOST ALGORITHM: SIGNIFICANCE AND APPLICATIONS IN MACHINE LEARNING

LILIT TER-VARDANYAN

International Scientific-Educational Centre of NAS RA, Lecturer
lilit.ter-vardanyan@isec.am

DOI: 10.54503/2579-2903-2024.1-175

Abstract

This article presents data visualization of gender classification analysis based on voice using the XGBoost algorithm **[1].** The main objective of the research is to visualize the effectiveness assessment of this method in predicting gender based on acoustic characteristics of audio files and data visualization. We provide a brief methodology of feature extraction from voice data using the librosa library and their visualization. The XGBoost model is trained on the training dataset and evaluated on the test data using performance evaluation metrics. The analysis results include a detailed examination of the data and feature visualization. We also present the accuracy and error probability of the XGBoost model on test data and analyze its performance, providing visualizations.

Keywords and phrases: gender classification, XGBoost, machine learning, dataset, algorithm, librosa, visualization, Python, Matplotlib, Seaborn.

XGBOOST ԱԼԳՈՐԻԹՄԻ ՄԻՋՈՑՈՎԸՍՏ ՍԵՌԻ ՁԱՅՆԻ ԴԱՍԱԿԱՐԳՄԱՆ ՏՎՅԱԼՆԵՐԻ ՎԻԶՈՒԱԼԻԶԱՑԻԱ՝ՆՇԱՆԱԿՈՒԹՅՈՒՆԸ ԵՎ ԿԻՐԱՌՈՒՄՆԵՐԸ ՄԵՔԵՆԱՅԱԿԱՆ ՈՒՍՈՒՑՄԱՆ ՄԵՋ

ԼԻԼԻԹ ՏԵՐ–ՎԱՐԴԱՆՅԱՆ,

ՀՀ Գիտությունների ազգային ակադեմիայի
գիտակրթական միջազգային կենտրոնի դասախոս
lilit.ter–vardanyan@isec.am

Համառոտագիր

XGBoost ալգորիթմն օգտագործվել է որպես գնահատման մեթոդ ձայնային տվյալներն ըստ սեռի վերլուծելու նպատակով: Հետազոտության հիմնական նպատակն է ալգորիթմի, որպես մոդել, կանխատեսումը ըստ սեռի՝ ձայնային բնութագրերի հիման վրա: Մեր կողմից հակիրճ տրամադրվել է ձայնային տվյալներից գործառույթներ դուրս բերելու մեթոդաբանությունը՝ օգտագործելով librosa գրադարանը: XGBoost մոդելը ուսուցանվել է ուսուցանվող տվյալների հավաքածուի հիման վրա և մոդելի ճշգրտության գնահատում կատարվել էթեստային տվյալների հիման վրա՝ օգտագործելով կատարողականի գնահատման չափանիշները: Վերլուծության արդյունքները ներառում են տվյալների մանրամասն ուսումնասիրություն և առանձնահատկությունների վիզուալացում: Ներկայացվել է նաև XGBoost մոդելի ճշգրտությունն ու սխալի հավանականությունը թեստային տվյալների հիման վրա, կատարվել է տվյալների վերլուծություն և դրաց

վիզիալացումը:

Բանալի բառեր և բառակապակցություններ. XGBoost, ձայնային տվյալներ, librosa, Python, Matplotlib, Seaborn, մեքենայական ուսուցում, վիզուալիցում, dataset, ալգորիթմ:

ВИЗУАЛИЗАЦИЯ ДАННЫХ ДЛЯ КЛАССИФИКАЦИИ ГОЛОСА ПО ПОЛУ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА XGBOOST: ЗНАЧИМОСТЬ И ПРИМЕНЕНИЕ В МАШИННОМ ОБУЧЕНИИ

ЛИЛИТ ТЕР-ВАРДАНЯН

Международный научно-образовательный центр НАН РА, преподаватель

lilit.ter-vardanyan@isec.am

Аннотация

В данной статье представлена визуализация данных анализа классификации пола по голосу с использованием алгоритма XGBoost. Основной целью исследования является визуализация оценки эффективности этого метода в предсказании пола на основе акустических характеристик аудиофайлов и визуализация данных. Мы представляем краткую методологию извлечения признаков из голосовых данных с использованием библиотеки librosa и их визуализацию. Модель XGBoost обучается на обучающем наборе данных и оценивается на тестовых данных с использованием метрик оценки производительности. Результаты анализа включают в себя подробное исследование данных, визуализацию признаков. Мы также представляем точность и вероятность ошибки модели XGBoost на тестовых данных и анализируем ее производительность, предоставляя визуализации.

Ключевые слова и словосочетания: гендерная классификация, XGBoost, машинное обучение, набор данных, dataset, алгоритм, librosa, визуализация, Python, Matplotlib, Seaborn.

Introduction

Gender identification and classification based on voice characteristics is an important task in security systems such as biometric authentication systems, in the field of audio processing, and data analysis. Data visualization is an important tool for analyzing and presenting information in a graphical format, as well as quickly and conveniently detecting anomalies in a database and identifying various patterns, whether related or unrelated. This is important for selecting the right model for training. Additional graphical representations can help illustrate relationships between different variables and make conclusions more understandable for analysts and data specialists. Visualization also contributes to a better understanding of data and its structure, aiding in making more informed decisions based on that data. In this study, we explore types of data visualization in Python for the task of voice-based gender classification, where the main goal is to determine a person's gender based on acoustic characteristics of their voice such as the average frequency in the voice signal spectrum, frequency component dispersion measure, average dominant frequency in the acoustic signal, maximum dominant frequency in the acoustic signal, etc. Extracting acoustic characteristics from voice data plays a key role in gender classification effectiveness. To understand the relationship and differences between acoustic characteristics, data visualization is used. Python offers many libraries for creating various charts, diagrams, and

other visualizations. In this article, we will look at the main libraries like Matplotlib, Seaborn, and Plotly and show how to use them to create data visualizations. Each of these libraries offers its own capabilities and visualization style, thus the choice depends on preferences and specific tasks at hand. For a brief overview of the libraries and their capabilities, see Milovanovic et al [2].

Matplotlib is a cross-platform data visualization library built on NumPy arrays. Matplotlib consists of several types of plots such as line, bar, scatter, histogram, etc. Matplotlib comes with a wide range of charts (Line, Bar, Histograms, Scatter, Pie Charts). These charts help understand trends and patterns, as well as correlations. These are usually tools for analyzing quantitative information. Seaborn is a library primarily used for building statistical graphics in Python. It is built on top of Matplotlib and provides beautiful default styles and color palettes to make statistical graphics more attractive. In machine learning tasks related to audio data, another important library is librosa. It is used for analyzing spectral and temporal characteristics using various methods. To achieve higher model performance before training, we preprocess the data, including feature standardization, outlier removal, and splitting the dataset into training and testing subsets. (fig.1)

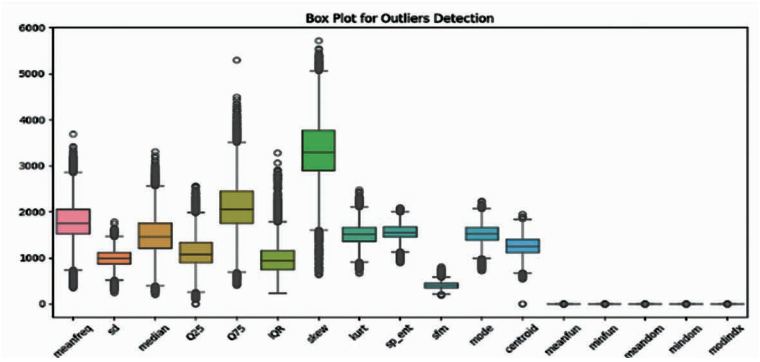
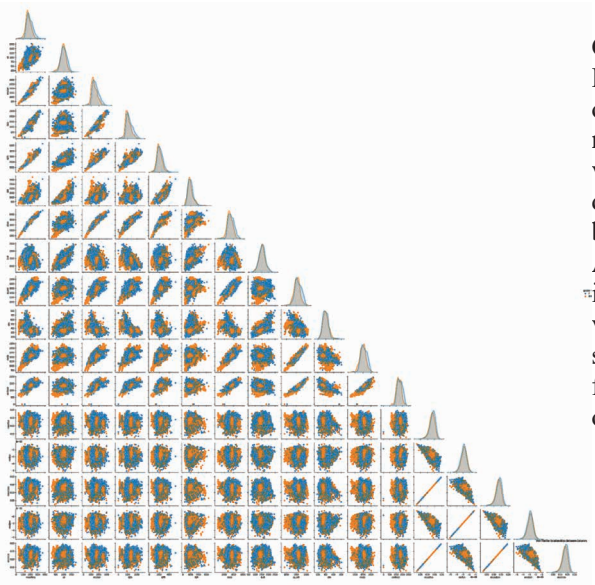


Fig. 1. Detecting anomalies using the visualization method of boxplots



Visual Comparison of Voice Characteristics Between Genders: Pairplot allows us to create a matrix of plots where each feature pair is represented as a scatter plot. This visually compares the distribution of different acoustic voice characteristics between male and female voices. Additionally, this method helps in identifying relationships between voice characteristics and gender, selecting the most informative features, and detecting possible outliers and anomalies (fig 2).

Fig.2. Comparison of voice characteristics between different genders using the pairplot method

Machine Learning Model: To solve this task, we use the XGBoost machine learning method, widely used in classification and regression tasks. For gender classification based on voice characteristics, we extract relevant features from audio files.

```
# Import necessary libraries
from sklearn.preprocessing import LabelEncoder  # 1 ; 0
from sklearn.preprocessing import StandardScaler
import matplotlib.pylab as plt
from sklearn.model_selection import train_test_split  # test and train validation parts
from sklearn import metrics
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns #visualization
import librosa #for voice
import xgboost as xgb #model, boosting algorithm
import joblib

def extract_features(audio_path):

    y, sr = librosa.load(audio_path, duration=3)  # Load audio, limit to 3 seconds
    # Basic statistics
    meanfreq = np.mean(librosa.feature.spectral_centroid(y=y, sr=sr))
    sd = np.std(librosa.feature.spectral_centroid(y=y, sr=sr))
    median = np.median(librosa.feature.spectral_centroid(y=y, sr=sr))
    Q25 = np.percentile(librosa.feature.spectral_centroid(y=y, sr=sr), 25)
    Q75 = np.percentile(librosa.feature.spectral_centroid(y=y, sr=sr), 75)
    IQR = Q75 - Q25
    skew = np.mean(librosa.feature.spectral_rolloff(y=y, sr=sr))
    kurt = np.std(librosa.feature.spectral_rolloff(y=y, sr=sr))
```

Meaning of Acoustic Features:

- Meanfreq: Mean frequency (in kHz) - the average frequency in the voice signal spectrum.
- SD: Standard deviation of frequency - a measure of frequency component dispersion.
- Median: Median frequency (in kHz) - the middle value in the overall range of frequencies in the voice signal spectrum.
- Q25: First quartile (in kHz) - the value below which 25% of frequency components lie.
- Q75: Third quartile (in kHz) - the value below which 75% of frequency components lie.
- IQR: Interquartile range (in kHz) - the difference between the third and first quartile values.
- Skew: Skewness - a measure of the asymmetry of frequency component distribution.
- Kurt: Kurtosis - a measure of the sharpness of the frequency component distribution peak.
- Sp.ent: Spectral entropy - a measure of the diversity of spectral components.
- SFM: Spectral flatness - a measure of the flatness of the voice signal spectrum.
- Mode: Mode frequency - the most frequently occurring frequency.
- Centroid: Frequency centroid - the weighted mean frequency value of the spectrum.
- Meanfun: Mean fundamental frequency measured on the acoustic signal.
- Minfun: Minimum fundamental frequency on the acoustic signal.
- Maxfun: Maximum fundamental frequency on the acoustic signal.

- Meandom: Mean dominant frequency on the acoustic signal.
- Mindom: Minimum dominant frequency on the acoustic signal.
- Maxdom: Maximum dominant frequency on the acoustic signal.
- Dfrange: Dominant frequency range on the acoustic signal.
- Modindx: Modulation index - calculated as the accumulated absolute difference between neighboring fundamental frequency measurements divided by the frequency range.

These features provide information about various spectral and temporal characteristics of the voice signal, necessary for accurate gender classification. We analyzed 15,469 audio files, including 7,972 female voices and 7,497 male voices, with durations ranging from 20.14 seconds to 25.18 seconds.

Visualization of Feature Distributions: We visualize the distribution of individual acoustic features using amplitude-time, spectrogram, and violin plots. These visualizations help understand the data structure and identify potential differences between male and female voices (fig 3,4).

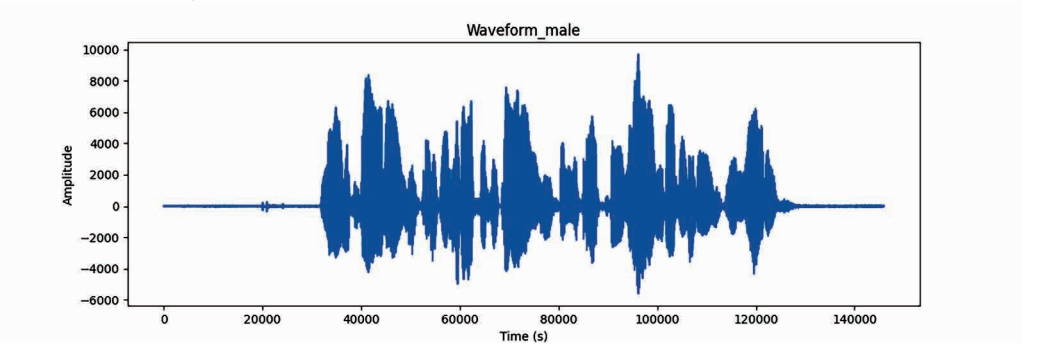


Fig. 3. Amplitude-time characteristics of the female voice

Figures 3 and 4 show the amplitude-time characteristics of male and female sounds (where the X-axis represents time and the Y-axis represents the amplitude of sound oscillations). Male and female voices can differ in several aspects, although they also depend on individual characteristics of each person. Frequency and Peak Height: Male voices typically have a lower fundamental frequency (lower on the Y-axis), leading to higher and wider amplitude peaks on the graph. In contrast, female voices have a higher fundamental frequency (higher on the Y-axis), resulting in narrower and higher amplitude peaks.

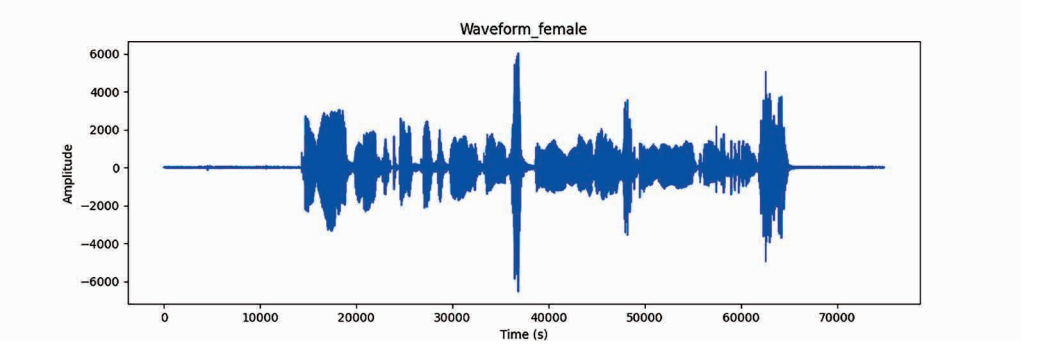


Fig. 4. Amplitude-time characteristics of the male voice

Intensity and Amplitude: Male voices are often characterized by higher amplitude of sound oscillations (higher values on the Y-axis), which is related to larger vocal cords and a deeper voice. Female voices, with smaller vocal cords, tend to have lower amplitude of sound.

Duration of Sound Signals: Differences in the duration of sound signals can also be observed on the graph. Male voices, usually with a lower pitch, may have longer and continuous periods of high amplitude. Female voices, with a higher pitch, may have shorter and less intense periods of amplitude.

Spectrograms are widely used in sound analysis and processing, as well as in speech and music technology. They help analyze sound characteristics, identify frequency features of sound signals, and even recognize speech or sound patterns. A voice spectrogram (or audio spectrogram) is a graphical representation of sound that displays changes in frequency and amplitude of sound oscillations over time (fig. 5,6). It is a three-dimensional image where the X-axis represents time, the Y-axis represents frequency, and the color or brightness indicates the sound amplitude at a particular frequency at a specific time. In a voice spectrogram, you can see which frequencies dominate at a certain time and how they change over time. Typically, vocal sounds are displayed as vertical bars or spots on the spectrogram, where brighter areas correspond to higher amplitudes at specific frequencies.

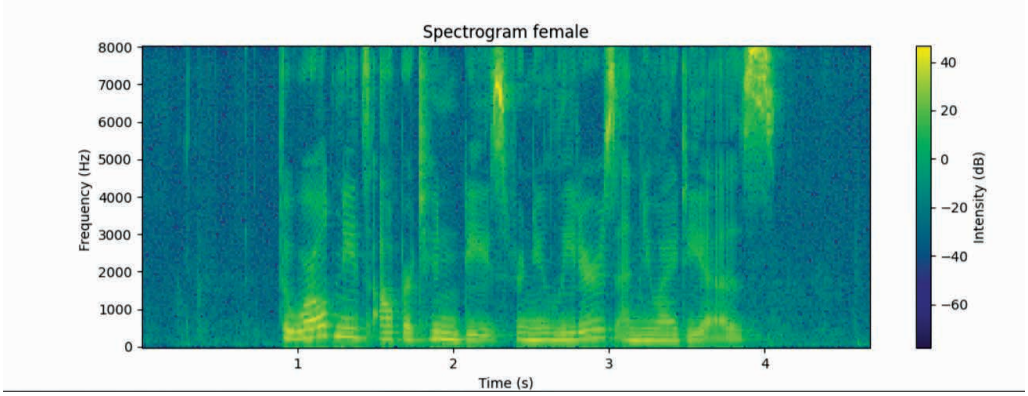


Fig. 5. Spectrogram of the female voice

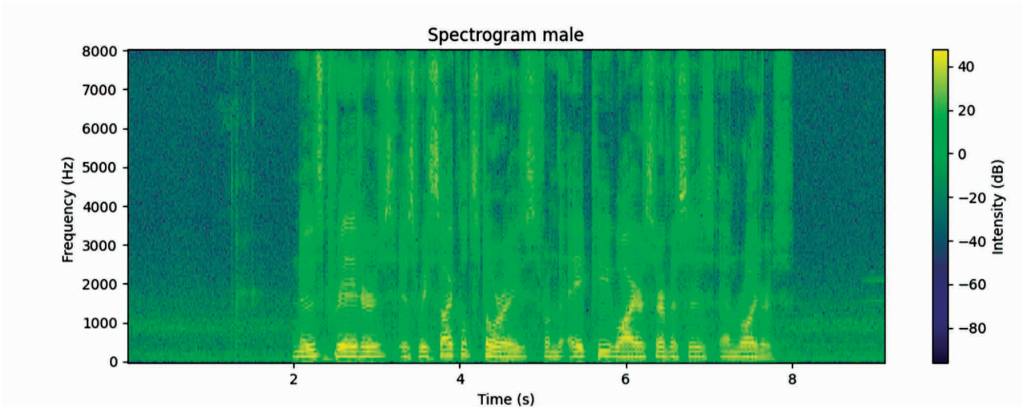


Fig. 6. Spectrogram of the male voice

Violin plots are widely used for a deeper understanding of the data being used. These plots are used to visualize the distribution of numerical data and have several advantages and applications (fig. 7).

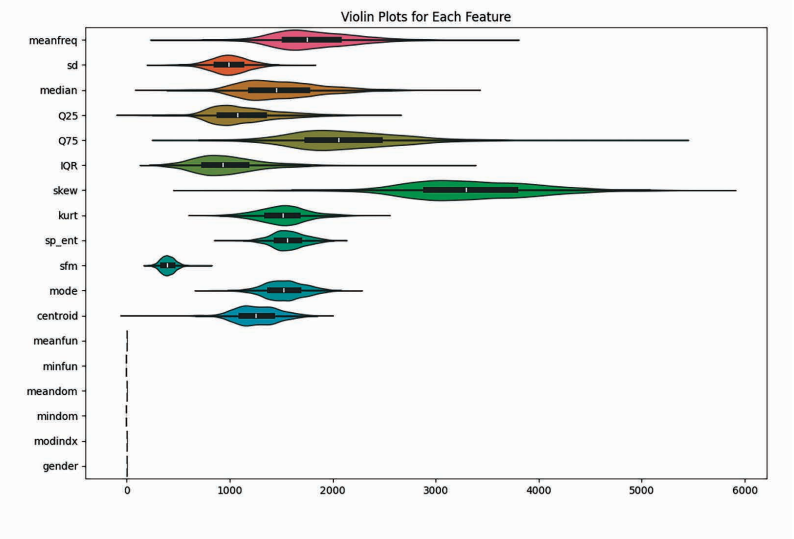


Fig. 7. The violin plot represents the distribution of pitch frequencies of male and female voices by the height of their sound

For data analysis using violin plots related to gender determination from voices, the following methods can be applied:

Comparison of Frequency Characteristics: Violin plots allow comparing the distribution of pitch or other acoustic voice characteristics between different gender groups. Based on this comparison, conclusions can be drawn about differences or similarities between male and female voices.

Detection of Frequency Distribution Density: The width of the violin plot in different sections reflects the density of voice frequency distribution. This helps determine which frequency ranges have the most frequent or rare values, which can be useful in analyzing differences between genders.

Representation of Key Statistical Parameters: Violin plots can visually represent the key statistical parameters of each group, such as median, quartiles, and range of values. This helps better understand the distribution characteristics of voices and identify potential differences between genders.

Using visualization plays a key role in evaluating the results of model work. For example, creating graphs depicting the relationship between predicted values and actual values allows us to assess the model's effectiveness and identify its weaknesses, etc.

In this article, we visualized the results with different percentages of test data to evaluate the model's effectiveness and the probability of error depending on the percentage of test data (see fig. 8, 9).

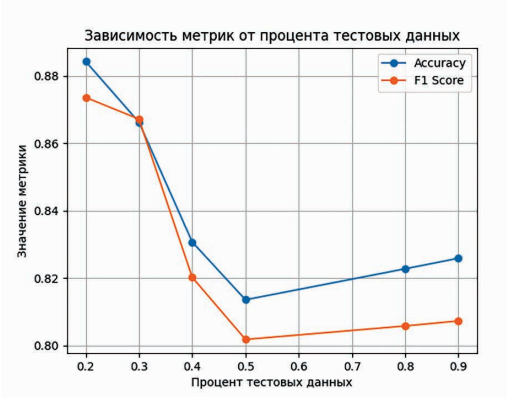


Fig. 8. Accuracy and F1 Score graphs depending on the percentage of test data

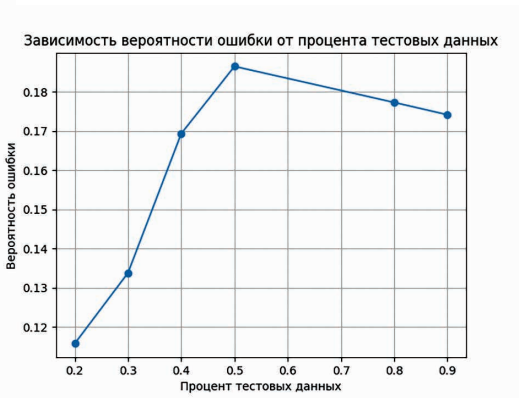


Fig. 9. Error probability depending on the percentage of test data

Conclusion

In this article, we presented the visualization of data analysis for voice-based gender classification using the XGBoost algorithm. Our results demonstrate that the use of visualization plays an important role in evaluating the model's performance and analyzing data before model training. Thus, our work represents a significant contribution to the field of voice-based gender classification and underscores the importance of data visualization for the analysis and interpretation of machine learning results **[3, 4]**.

References

1. Chen T. and Guestrin C., Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785794. ACM, 2016.
2. Milovanovic I., Foures D., Vettigli G. Python Data Visualization Cookbook - Second Edition Published by Packt Publishing, 2015.
3. Raschka S. and Mirjalili V., Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition.
4. Raschka J. S., Liu Y., Mirjalili V. and Dzhulgakov D., Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python.

The article has been submitted for publication: 05.03.2024
Հոդվածը ներկայացվել է տպագրության. 05.03.2024
Статья представлена к публикации: 05.03.2024

The article is sent for review: 09.04.2024
Հոդվածն ուղարկվել է գրախոսության. 09.04.2024
Статья отправлена на рецензию: 09.04.2024

The article is accepted for publication: 18.04.2024
Հոդվածն ընդունվել է տպագրության. 18.04.2024
Статья принята к печати: 18.04.2024