E-ISSN: 2674-5224, DOI: 10.36962/PAHTEI

AZƏRBAYCAN ALİ TEXNİKİ MƏKTƏBLƏRİNİN XƏBƏRLƏRİ

PROCEEDINGS OF AZERBAIJAN HIGH TECHNICAL EDUCATIONAL INSTITUTIONS

ВЕСТНИК ВЫСШИХ ТЕХНИЧЕСКИХ УЧЕБНЫХ ЗАВЕДЕНИЙ АЗЕРБАЙДЖАНА

VOLUME 22 ISSUE 11 2022 CİLD 22 BURAXILIŞ 11 2022

https://bsj.fisdd.org/index.php/pahtei

ISSN: 1609-1620, E-ISSN: 2674-5224
P A H T E I
REFEREED & REVIEWED JOURNAL
E-ISSN: 2674-5224
VOLUME 22 ISSUE 11 2022
116-129

SELECTING A MACHINE LEARNING ALGORITHM FOR CREATING A HEPATOCELLULAR CARCINOMA PREDICTION SYSTEM

Masuma Mammadova[1], Zarifa Jabrayilova[2], Lala Garayeva[3]
[1,2,3]Institute of Information Technology of ANAS, Azerbaijan
[1,2,3]Department number 11, [1]Head of Department, Corresponding Member of Azerbaijan National Academy of Sciences, Doctor of Technical Sciences, Professor, http://orcid.org/0000-0002-2205-1023 , Email: mmg51@mail.ru
[2]Chief researcher, Doctor of Technical Sciences, Associate professor, http://orcid.org/0000-0002-9661-5805, djabrailova_z@mail.ru
[3]Junior researcher, Email: [3]karayevalala.01@gmail.com

ABSTRACT

The formation of e-health has stimulated the development of intelligent systems that provide information support for medical decisions. These systems are used to make a diagnosis, choose a more effective treatment method, predict, search for suitable conditions (precedents), control and schedule therapy, recognize and interpret images, monitor the clinical-pharmacological properties (toxicity) of drugs, etc. The basis of these systems are diseases in the specific subject area of medicine, their possible causes, development period, clinical manifestations, observed signs, symptoms, etc. Successes achieved in the creation of such systems that prevent errors in making medical decisions, have necessitated the creation of a system for diagnosis and prediction of hepatocellular carcinoma (HCC), known as liver cancer, which is the third leading cause of lethal cancer in the world.

HCC is characterized by a set of clinical manifestations of critical conditions, each of which, in turn, is defined by a set of clinical signs and data. The analysis of these data shows that in the conditions of information abundance, the physician has to make a decision by referring to a part of the obtained information. As a result, errors occur in physicians' decisions determined by certain combinations of a vast number of indicators and clinical signs. To predict HCC based on such numerous, diverse and heterogeneous unstructured data, preference is given to the method of artificial intelligence, i.e., machine learning. Machine learning enables data processing, presentation, and making predictions based on the results obtained.

This article explores the possibility of applying machine learning algorithms to create an HCC prediction system and solves the problem of selecting the best algorithm. The study conducted in this regard uses the HCC Dataset database taken from Kaggle platform, and 49 features/attribute data of 165 patients are referenced. The experiment conducted for the prediction of HCC is presented in stages.

In the first stage, pre-processing (or the process of database cleaning) is performed in order to bring the data into a uniform form, while the cleaning of unrelated and scattered data is implemented with the direct involvement of the user. Libraries Scikit-learn, Pandas, NumPy, etc. are used in the Jupiter programming environment to make the database beneficial and simplify data processing. Correlation heatmaps are used to find both linear and non-linear relationships between variables. Min-max scaling is applied to one or more feature columns to normalize the data.

The second step analyzes all features present in the database, determines their types and target class.

The third step presents the criteria (precision, recall, F1-ScoreF, accuracy) for evaluating the performance of machine learning algorithms used for classification in the database and their determination formulas.
Finally, the fourth step determines the performance evaluation criteria with the application of Support Vector Machine, Random Forest, Logistic Regression machine learning algorithms. Jupiter software in Anaconda environment is used to analyze the result. Different classifiers are selected to achieve results accuracy and also to see how the data performed on different classifiers. Referring to the comparative analysis of the obtained results, the method with the best results is selected for the creation of the HCC prediction system. The Random Forest algorithm is estimated to show highest accuracy according to the criteria of the error matrix.
Thus, it is justified to use the Random Forest machine learning algorithm to improve the accuracy of forecasts in the creation of an intelligent system of HCC prediction.
Keywords: hepatocellular carcinoma, intelligent prediction system, machine learning algorithm, confusion matrix, accuracy criterion.

HEPATOSELLULAR KARSİNOMANIN PROQNOZLAŞDIRILMASI SİSTEMİNİN YARADILMASI ÜÇÜN MAŞIN TƏLİMİ ALQORİTMİNİN SEÇİLMƏSİ

Məsumə Məmmədova[1], Zərifə Cəbrayılova[2], Lalə Qarayeva[3]
[1,2,3]AMEA İnformasiya Texnologiyaları İnstitutu, [1,2,3]11 saylı şöbə
[1]Şöbə müdiri, AMEA-nın müxbir üzvü, texnika elmləri doktoru, professor, http://orcid.org/0000-0002-2205-1023, Email: mmg51@mail.ru
[2]Baş elmi işçi, texnika elmləri doktoru, dosent, http://orcid.org/0000-0002-9661-5805, Email: djabrailova_z@mail.ru
[3]Kiçik elmi işçi, Email: karayevalala.01@gmail.com

XÜLASƏ

Məqalədə qaraciyər xərçəngi kimi tanınan, xərçəng səbəbindən ölənlərin sayına görə dünyada üçüncü yeri tutan hepatosellular karsinomanın (HSK) proqnozlaşdırılması sisteminin yaradılması üçün maşın təlimi alqoritmlərinin tətbiqi imlkanları araşdırılmış, ən yaxşı nəticə göstərən alqoritmin seçilməsi məsələsi həll edilmişdir. Elektron tibbin formalaşması, tibbi qərarların qəbul edilməsində həkim səhvlərinin qarşısının alınması üçün intellektual sistemlərin yaradılması və bu istiqamətdəki uğurlar HSK-nın ilkin diaqnozu və proqnozlaşdırılması üçün süni intellektə əsaslanan metodların tətbiqini aktuallaşdırmışdır. Bu məqsədlə aparılmış tədqiqatda Kaggle platformasından götürülmüş HCC Dataset verilənlər bazasından, Jupiter proqramlaşdırma mühitində scikit-learn, Pandas, NumPy və s. kitabxanalardan istifadə edirmiş, 165 pasiyentin 49 xüsusiyyət/atribut verilənlərinə istinad edilmişdir. HSK-nın proqnozlaşdırılması üçün tətbiq edilmiş Logistic Regression, Support Vector Machine, Random Forest kimi maşın təlimi alqoritmlərinin nəticələri təqdim olunmuşdur. Eksperimentlərin nəricələrinin müqayisəli analizinə istinad eməklə HSK-nın proqnozlaşdırılması sisteminin yaradılması üçün ən yaxşı nəticə göstərən metod seçilmişdir.

Açar sözlər: hepatocellular carcinoma, intellektual proqnoz sistemi, maşın təlimi alqoritmləri, xəta matrisi, dəqiqlik meyarı.

Giriş

Hazırda elektron tibbin formalaşması tibbi qərarların qəbul olunmasına informasiya dəstəyi göstərən sistemlərin inkişafına təkan vermişdir. Xəstəliklərin müxtəlif variantlarda təzahür etməsi, diaqnoz və müalicənin birmənalı meyarlarının mövcud olmaması; hər bir xəstəliyin çoxlu sayda giriş verilənlərinə əsaslanması; xəstəliyi xarakterizə edən göstəricilərin keyfiyyət və kəmiyyət xarakterli, əsasən qeyri-dəqiq olması həkim qərarlarının qəbulunda mümkün səhvlərin qarşısının alınması üçün intellektual metodların tətbiqini zərirətə çevirmişdir [1]. Təcrübəli həkim konkret situasiyada analoji vəziyyətləri nəzərə alaraq öz mülahizələrini təsdiq etmək üçün baza məlumatlarını şəxsi təcrübəsi ilə uyğunlaşdırır, xəstəliyin atipik formalarını müəyyən edir, prosesin dinamikasını proqnozlaşdırır. Bu, həkimin bilik və empirik təcrübəsinə əsaslanan müalicə-diaqnostik qərarların qəbul edilməsi üzrə məntiqi mühakimələrin imitasiyasının vacibliyi məsələsini aktuallaşdırır. Məlumdur ki, təcrübəli həkim-ekspertlərin biliklərinin toplanması, saxlanması, manipulyasiyası, eləcə də, hər bir konkret verilənlər toplusu üzrə xəstəliyin müəyyən edilməsi və adekvat qərarların qəbul edilməsi üçün daha səmərəli vasitə biliklərə əsaslanan intellektual sistemlərdir (tibbi qərarların qəbulunu dəstəkləyən sistemlər). Bu sistemlərin əsasını tibbin konkret predmet sahəsində olan xəstəliklər, onların mümkün səbəbləri, inkişaf müddəti, kliniki təzahürləri, müşahidə olunan əlamətləri, simptomları və s. təşkil edir. Bu sistemlər diaqnoz qoyulması, daha effektiv müalicə üsulunun seçilməsi, proqnozlaşdırma, uyğun vəziyyətlərin (presedentlərin) axtarışı, terapiyaya nəzarət və planlaşdırma, təsvirlərin tanınması və interpretasiyası, dərman vasitələrinin kliniki-farmakoloji xüsusiyyətlərinin (toksikliyinin) monitorinqi və s. məsələlərin həllində tətbiq olunur.

Hazırkı məqalədə də qaraciyər xərçəngi kimi tanınan hepatosellular karsinomanın (HSK) ilkin diaqnozu və proqnozlaşdırılması üçün intellektual sistemin yaradılması istiqamətində aparılan tədqiqatda maşın təlimi alqoritmlərinin tətbiqi ilə bağlı araşdırmaların nəticələri verilmiş, daha yaxşı nəticə göstərən alqoritmin müəyyənləşdirilməsi üçün aparılmış eksperimentlərin nəticələri təqdim olunmuşdur.

Məqsəd

Problemin aktuallığı və əlaqəli tədqiqatlar. Ümumdünya Səhiyyə Təşkilatının hesabatına görə, xərçəng xəstəliyi dünya üzrə əsas ölüm səbəbi kimi qiymətləndirilir və 2020-ci ildə bu səbəbdən 10 milyona yaxın ölüm hadisəsi qeydə alınmışdır [2]. Xərçəng səbəbindən ölənlərin sayına görə 2-ci yeri qaraciyər xərçəngi tutur və onun 80%-ni (ABŞ-da bu göstərici 90% təşkil edir) HSK təşkil edir [3, 4]. Bu şiş xəstəliyi ən çox 60-70 yaşlarında və adətən kişilərdə (qadınlardan 2,5 dəfə çox) rast gəlinir, yüksək riskli ölkələrdə isə daha erkən, yəni 30-40 yaşlarında müşahidə edilir. HSK rastgəlmə tezliyinə görə xərçəng xəstəlikləri arasında 5-6-cı yeri, xərçənglə bağlı ölüm səbəbləri arasında isə üçüncü yeri turur [5]. Hər il dünyada təxminən bir milyona yaxın insanda HSK tapılır [3, 5]. O, adətən xroniki qaraciyər xəstəliyi kimi təzahür edir və ya sirozu olan xəstələrdə müşahidə edilir. Son məlumatlara görə, HSK dünyada daha geniş yayılmış ölümcül xərçəng növlərindən biridir və hər il 600.000-dən çox insanın ölümünə səbəb olur [6, 7].

HSK-nın yaranmasına səbəb olan risk faktorları kimi genetika, yaş, cins, kimyəvi maddələr, hormonlar və qidalanma göstəriir. HSK-da şiş hüceyrələri hepatositlərə bənzəyir, lakin dərəcələrinə görə onlar bir-birlərindən fərqlənirlər. HSK-nın şiş hüceyrələri qan damarlarına

sızmağa və böyüməyə meyllidir. Digər xərçənglər kimi, HCC-də də şiş hüceyrələri müxtəlif mərhələlərdə yavaş-yavaş böyüyür və erkən aşkar olunarsa, daha effektiv müalicə oluna bilər. Qaraciyər xərçənginin proqnoz və müalicəsi üsulunun seçilməsi üçün əsasən şişin yayılma dərəcəsinə istinad edilir, bu məqsədlə bir sıra təsnifatlardan istifadə edilir. Lakin təcrübə göstərir ki, digər şişlərdən fərqli olaraq, HSK-da şişin yayılma dərəcəsi ilə yanaşı, həm də qaraciyərin funksional vəziyyəti, orqanizmin ümumi halı da proqnozda və müalicə üsulunun seçilməsində önəmli rol oynayır. Ona görə də son illər şişin yayılmasını, qaraciyər parenximasını və ümumi vəziyyəti nəzərə alan təsnifatlar işlənib hazırlanmışdır [3].

HSK ilə əlaqəli verilənlərin analizi göstərir ki, informasiya bolluğu şəraitində həkim bu informasiyanın bir qisminə istinad etməklə qərar qəbul etməli olur. Nəticədə çoxlu sayda göstəricilərin, kliniki əlamətlərin müəyyən kombinasiyaları ilə təyin edilən həkim qərarlarında nöqsanlar yaranır və bu HSK-nın diaqnostikası və proqnozlaşdırılması üçün süni intellekt metodlarının tətbiqini, kliniki qərarların qəbuluna dəstək sistemlərinin yaranmasını şərtləndirmişdir.

Hazırda elmi ədəbiyyatda qaraciyər xəstəliklərinin aşkarlanması, diaqnostikası və müalicəsi üçün intellektual sistemlərin işlənilməsi istiqamətində tədqiqatlara rast gəlinir [8-10]. [5]-də HSK-nın proqnozlaşdırılması üçün maşın təlimi və dərin təlim metodlarından istifadənin əhəmiyyəti göstərilmişdir. Müəlliflərin təklif etdiyi HSK-nın mərhələsinin təyini üçün qeyri-səlis qaydalara əsaslanan sistemin işlənilməsi metodikası [11]-də, bu sistemin bilik bazasının formalaşdırılması prinsipləri [12]-də təqdim edilmişdir.

Bu istiqamətdə müəlliflərin apardıqları tədqiqatın növbəti mərhələsi qaraciyər xərçənginin proqnozlaşdırılması sisteminin işlənilməsidir və hazırkı məqalədə maşın təlimi alqoritmlərinin tətbiqi ilə HSK-nın proqnozlaşdırılması məsələsinin həllinə baxılmış, ən yüksək nəticə göstərən alqoritmin seçilməsidir üzrə alınan nəticələr təqdim edilmişdir. Bu yanaşma pasiyentlərin xəstəliklə bağlı toplanmış məlumatları əsasında yaradılmış bazalardan istifadə etməklə proqnoz vermək üçün nəzərdə tutulmuşdur, xəstəliklə bağlı dəqiq və vaxtında qərarların qəbul edilməsində, xəstəliyin qarşısının alınmasında həkumlərə dəstək göstərə bilər.

Maşın təlimi, verilənlərin emal edilməsinə, təqdim edliməsinə və alınan nəticələr əsasında proqnozlaşdırmanı yerinə yetirməyə imkan verir. Bir çox tədqiqatçılar maşın təlimini aşağıdakı kimi təsnif edirlər:

1) əvvəlki məlumatlardan istifadə etməklə təlim və ya öyrənmək (Training or Learning from past data);
2) təsnifat, proqnozlaşdırma və s. kimi tapşırıqları yerinə yetirmək;
3) əvvəlki və indiki məlumatlardan əldə edilmiş təcrübə əsasında performansı artırmaq.

Verilənlər bazasına toplanmış məlumatlara əsaslanaraq HSK-nı proqnozlaşdırmaq üçün nəzarət edilən maşın təlimi alqoritmlərindən istifadə edilmişdir.

[13]-də data miningdən istifadə etdəklə diabet xəstəliyinin proqnozlaşdırılması təklif edilmişdir. Pasiyentin xəstəliyini proqnozlaşdırmaq *üçün K-nearest neighbours (KNN)* və *Naive Bayes* kimi iki klassifikator alqoritmindən istifadə edilmişdir. Proqnozlaşdırma üçün 2000 diabet xəstəsi haqqında məlumata malik bazaya istinad edilmiş, KNN və *Naive Bayes* alqoritmlərindən istifadə edərək proqnozlaşdırmada yüksək nəticə almağa nail olunmuşdur.. Qeyd edək ki, istinad edilən bazaların daha çox məlumata malik olması təklif edilən modelin dəqiqliyinin və səmərəliliyinin artırılmasında mühüm məqamdır.

[2, 5]-də təsvir edilən tədqiqatlartda maşın təlimi və genetik alqoritmlərdən istifadə edərək müxtəlif növ xəstəliklərin proqnozunun təkmilləşdirilməsi imkanlarına baxılmışdır. Bu məqsədlə

ISSN: 1609-1620, E-ISSN: 2674-5224
PAHTEI
REFEREED & REVIEWED JOURNAL
E-ISSN: 2674-5224
VOLUME 22 ISSUE 11 2022

müxtəlif növ xəstəliklərin xüsusiyyətlərini xarakterizə edən verilənlərdən ibarət bazalara istinad edirmişdir. Bu bazalar, adətən, həm strukturlaşdırılmış, həm də strukturlaşdırılmamış verilənlərdən ibarət olur.

Verilənlər toplusunda bir-biri ilə əlaqəli olmayan xüsusiyyətləri tapmaq üçün genetik alqoritmlərdən istifadə edilir və Recurrent Neural Network (RNN) vasitəsilə strukturlaşdırılmamış verilənlərdən lazımi xüsusiyyətlər çıxarılır (başqa sözlə desək, baza təmizlənir). Əldə edilən yeni bazaya Support Vector Machine (SVM), Random Forest, Logistic Regression və s. klassifikatorlarına əsaslanan maşın təlimi metodlarını tətbiq etməklə sistemin düzgünlüyü yoxlanılır.

Beləliklə, hazırkı məqalənin məqsədi HSK-nın ilkin diaqnozu və proqnozlaşdırılması məsələsinin həlli üçün maşın təlimi alqoritmlərinin tətbiqi imkanının araşdırılması və ən yaxşı alqoritmin seçilməsidir.

Metodlar

HSK-nın proqnozlaşdırılması üçün maşın təlimi alqoritmlərinin tətbiqi. HCC-nin proqnozlaşdırıılmasında maşın təlimi alqoritmlərinin tətbiqi üçün ilk növbədə müvafiq verilənlər bazası seçilmişdir. Bu məqsədlə Kaggle şirkətinin HCC Dataset [7, 13] adlı açıq verilənlər bazasından istifadə olunmuşdur. Verilənlər bazası Portuqaliya Universiteti Xəstəxanasında HSK-dan əziyyət çəkən 165 klinik xəstənin məlumatları əsasında formalaşdırılmışdır. Verilənlər bazasında Qaraciyərin Tədqiqi üzrə Avropa Assosiasiyası – Xərçəngin Tədqiqi və Müalicəsi üzrə Avropa Təşkilatı (eng. European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer) tərəfindən tövsiyə edilən 49 xüsusiyyət (kliniki əlamət) yer almışdır (şəkil 1).

```
1 liver_df.head()
```

	Gender	Symptoms	Alcohol	HBsAg	HBeAg	HBcAb	HCVAb	Cirrhosis	Endemic	Smoking	...	ALP	TP	Creatinine	Nodule	Major_Dim	Dir_Bil	Iron	Sa
0	1	0	1	0	0	0	0	1	0	1	...	150	7,1	0,7	1	3,5	0,5	52,5	3
1	0	0	0	0	0	0	1	1	0	1	...	120	7	0,58	1	1,8	0,85	32	1
2	1	0	1	1	0	1	0	1	0	1	...	109	7	2,1	5	13	0,1	28	
3	1	1	1	0	0	0	0	1	0	1	...	174	8,1	1,11	2	15,7	0,2	131	7
4	1	1	1	1	0	1	0	1	0	1	...	109	6,9	1,8	1	9	0,1	59	1

5 rows × 50 columns

Şəkil 1. Qaraciyər xərçəngi xəstəliyi üçün verilənlər bazası

Beləliklə, maşın təlimi alqoritmlərinin tətbiqi ilə HSK-nın proqnozlaşdırılması üçün keçirilən eksperiment aşağıdakı addımlardan ibarət olmuşdur:

Birinci addım: verilənlərin ilkin emalı. Bu addımda seçilmiş verilənlər bazasında bir-biri ilə əlaqəli olmayan verilənlərin olması yoxlanılmış və müəyyən hissələrdə verilənlərin səpələnmiş

olduğu müəyyənləşdirilmişdir. Həmin verilənlər şəkil 2-də verilmişdir:

```
1 liver_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 204 entries, 0 to 203
Data columns (total 50 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Gender           204 non-null    int64
 1   Symptoms         204 non-null    int64
 2   Alcohol          204 non-null    int64
 3   HBsAg            204 non-null    int64
 4   HBeAg            204 non-null    int64
 5   HBcAb            204 non-null    int64
 6   HCVAb            204 non-null    int64
 7   Cirrhosis        204 non-null    int64
 8   Endemic          204 non-null    int64
 9   Smoking          204 non-null    int64
 10  Diabetes         204 non-null    int64
 11  Obesity          204 non-null    int64
 12  Hemochro         204 non-null    int64
 13  AHT              204 non-null    int64
 14  CRI              204 non-null    int64
 15  HIV              204 non-null    int64
 16  NASH             204 non-null    int64
 17  Varices          204 non-null    int64
 18  Spleno           204 non-null    int64
 19  PHT              204 non-null    int64
 20  PVT              204 non-null    int64
 21  Metastasis       204 non-null    int64
 22  Hallmark         204 non-null    int64
 23  Age              204 non-null    int64
 24  Grams_day        204 non-null    int64
 25  Packs_year       204 non-null    object
 26  PS               204 non-null    int64
 27  Encephalopathy   204 non-null    int64
 28  Ascites          204 non-null    int64
 29  INR              204 non-null    object
 30  AFP              204 non-null    object
 31  Hemoglobin       204 non-null    object
 32  MCV              204 non-null    object
 33  Leucocytes       204 non-null    object
 34  Platelets        204 non-null    object
 35  Albumin          204 non-null    object
 36  Total_Bil        204 non-null    object
```

Şəkil 2. Qaraciyər xərçəngi xəstəliyindən əldə edilən verilənlərin tipi

Verilənlərin vahid formaya gətirilməsi üçün onların ilkin emalı (və ya verilənlər bazasının təmizlənməsi prosesi) həyata keçirilmiş, əlaqəli olmayan və səpələnmiş verilənlərin təmizlənməsi bilavasitə istifadəçinin müdaxiləsi ilə yerinə yetirilmişdir. Verilənlər bazasını faydalı etmək, verilənin işlənməsini sadələşdirmək üçün Pandas [14] və NumPy [15] kitabxanalarından istifadə edilmişdir. Kodların tətbiqi ilə müxtəlif tip verilənlər eyni tip verilənlərə çevrilmiş və bir-biri ilə əlaqəli xüsusiyyətlər təyin edilmişdir (şəkil 3).

ISSN: 1609-1620, E-ISSN: 2674-5224
PAHTEI
REFEREED & REVIEWED JOURNAL
E-ISSN: 2674-5224
VOLUME 22 ISSUE 11 2022

```
1 liver_df.info()
```

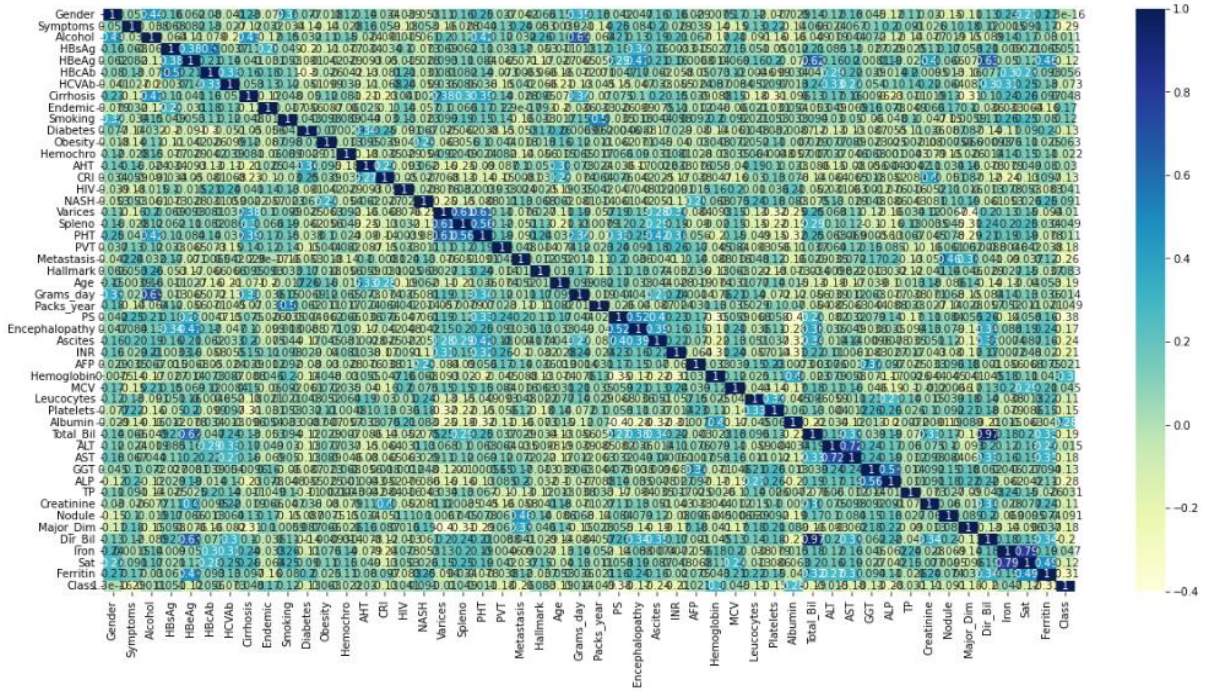
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 204 entries, 0 to 203
Data columns (total 50 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Gender           204 non-null    float64
 1   Symptoms         204 non-null    float64
 2   Alcohol          204 non-null    float64
 3   HBsAg            204 non-null    float64
 4   HBeAg            204 non-null    float64
 5   HBcAb            204 non-null    float64
 6   HCVAb            204 non-null    float64
 7   Cirrhosis        204 non-null    float64
 8   Endemic          204 non-null    float64
 9   Smoking          204 non-null    float64
 10  Diabetes         204 non-null    float64
 11  Obesity          204 non-null    float64
 12  Hemochro         204 non-null    float64
 13  AHT              204 non-null    float64
 14  CRI              204 non-null    float64
 15  HIV              204 non-null    float64
 16  NASH             204 non-null    float64
 17  Varices          204 non-null    float64
 18  Spleno           204 non-null    float64
 19  PHT              204 non-null    float64
 20  PVT              204 non-null    float64
 21  Metastasis       204 non-null    float64
 22  Hallmark         204 non-null    float64
 23  Age              204 non-null    float64
 24  Grams_day        204 non-null    float64
 25  Packs_year       204 non-null    float64
 26  PS               204 non-null    float64
 27  Encephalopathy   204 non-null    float64
 28  Ascites          204 non-null    float64
 29  INR              204 non-null    float64
 30  AFP              204 non-null    float64
 31  Hemoglobin       204 non-null    float64
 32  MCV              204 non-null    float64
 33  Leucocytes       204 non-null    float64
 34  Platelets        204 non-null    float64
 35  Albumin          204 non-null    float64
 36  Total_Bil        204 non-null    float64
 37  ALT              204 non-null    float64
 38  AST              204 non-null    float64
 39  GGT              204 non-null    float64
```

Şəkil 3. Verilənlərin tiplərinin dəyişdirilməsi

Dəyişənlər arasında həm xətti, həm də qeyri-xətti əlaqələri tapmaq üçün Correlation heatmaps istifadə edilmişdir (şəkil 4).

E-ISSN: 2674-5224
VOLUME 22 ISSUE 11 2022
ISSN: 1609-1620, E-ISSN: 2674-5224
PAHTEI
REFEREED & REVIEWED JOURNAL

```
plt.figure(figsize=(20, 10))
sns.heatmap(liver_df.corr(), annot=True, cmap='YlGnBu');
```



Şəkil 4. Qaraciyər xərçəngi xəstəliklərinin verilənlər toplusu üçün *Correlation heatmaps*

Məlumatları normallaşdırmaq üçün minimum-maksimum miqyaslama bir və ya bir neçə xüsusiyyət sütununa tətbiq olunmuşdur. Minimum-maksimum miqyaslama əsasında verilənlərin normallaşdırılması düsturu şəkil 5-də verilmişdir.

```
std = MinMaxScaler()
X = pd.DataFrame(std.fit_transform(X) , columns=X.columns)
X
```

	Gender	Symptoms	Alcohol	HBsAg	HBeAg	HBcAb	HCVAb	Cirrhosis	Endemic	Smoking	...
0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	...
1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	...
2	1.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	...
3	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	...
4	1.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	...
...	...	...	...	...	...	...	...	...	...	...	...
199	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...
200	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...
201	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...
202	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...
203	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

204 rows × 49 columns

Şəkil 5. Qaraciyər xərçəngi məlumat dəstinin MinMaxScaler

ISSN: 1609-1620, E-ISSN: 2674-5224
PAHTEI
REFEREED & REVIEWED JOURNAL
E-ISSN: 2674-5224
VOLUME 22 ISSUE 11 2022

İkinci addım: verilənlərin tədqiqi. HSK-nın verilənlər bazasında mövcud olan bütün xüsusiyyətlər təhlil olunur. Verilənlər bazasında 49 xüsusiyyət/atribut var, onlardan 23 atribut kəmiyyət xarakterli və 26 atribut keyfiyyət xarakterlidir. Birillik hesabata görə, hədəf sinfi (class) olaraq "0" (ölüm) və "1" (Yaşayan) kimi iki dəyişəndən ibarətdir. Cədvəl 1-də *HCC Dataset*-də mövcud olan xüsusiyyətlərin növləri göstərilmişdir.

Cədvəl 1. Verilənlər bazasında mövcud olan xüsusiyyətlərin növləri.

N	Verilənlərin növləri	Xüsusiyyətlər
1.	Formal (Nominal)	Gender, Symptoms, Alcohol, Hepatitis B, Surface Antigen, Hepatitis B e Antigen, Hepattitis B Core Antibody, Hepatitis C Virus, Antibody, Cirrhosis, Endemic, Countries, Smoking, Diabetes, Obesity, Hemochromatosis, Arterial Hypertension, Chronic Renal Insufficiency, Human Immunodeficiency Virus, Non-alcoholic Steatohepatitis, Esophageal Varices, Splenomegaly, Portal Hypertension, Portan Vein Thrombosis, Liver Metastasis, Radiological Hallmark
2.	İnteger	Number of Nodules, Age at Diagnosis
3.	Fasiləsiz olaraq dəyişən (Continuous)	Grams of Alcohol per day, Packs of Cigarettes per year, International Normalised Ratio, Alpha-Fetoprotein (ng/mL), Haemoglobin (g/dL), Mean Corpuscular Volume (fl), Leukocytes (G/L), Platelets (G/L), Albumin (mg/dL), Total Bilirubin (mg/dL), Alanine transaminase (U/L), Aspartate transaminase (U/L), Gamma glutamyl transferase (U/L), Alkaline phosphatase (U/L), Total Proteins (g/dL), Creatinine (mg/dL)
4.	Sıra sayı (Ordinal)	Performance Status, Encephaalograph degree, Ascites degree

Üçüncü addım: təsnifat və xəta matrisinin meyarlarının təyini. İstifadə olunan bazada təsnifləndirməni aparmaq üçün Naive Bayes, SVM, RF və LR maşın təlimi alqoritmləri istifadə olunmuşdur. Maşın təlimində klassifikatorların aşkarlama performansının qiymətləndirilməsi vacib məsələdir. Aşkarlama performansının qiymətləndirilməsində həssaslıq (precision), tamlıq (recall), yanlış pozitiv hallar (false positive rate-FPR), doğru pozitiv hallar (true positive rate-TP), f-ölçü (f-measure), dəqiqlik (accuracy) meyarlarından istifadə olunmuşdur.
Həssaqlıq (P) həqiqi müsbətlərin sayı kimi müəyyən edilir və aşağıdakı kimi təyin edilir:

$$P = \frac{T_p}{T_p + F_p} \qquad (1)$$

Burada: T_p – doğru təsnif edilmiş, proqnozlaşdırma ilə əlaqəli verilənlərin sayı; Fp – səhv təsnif edilmiş, proqnozlaşdırma ilə əlaqəli olmayan verilənlərin sayıdır.
Tamlıq (R) həqiqi müsbətlərin sayı kimi müəyyən edilir və aşağıdakı düsturla hesablanır:

$$R = \frac{T_p}{T_p + F_n} \qquad (2)$$

Burada: Fn – səhv kimi təsnif edilmiş proqnozlaşdırma ilə əlaqəli olmayan verilənlərin sayıdır.

F1-ölçü *(F1-Score)* geri çağırmanın harmonik ortası kimi müəyyən edilir və aşağıdakı düsturla hesablanır:

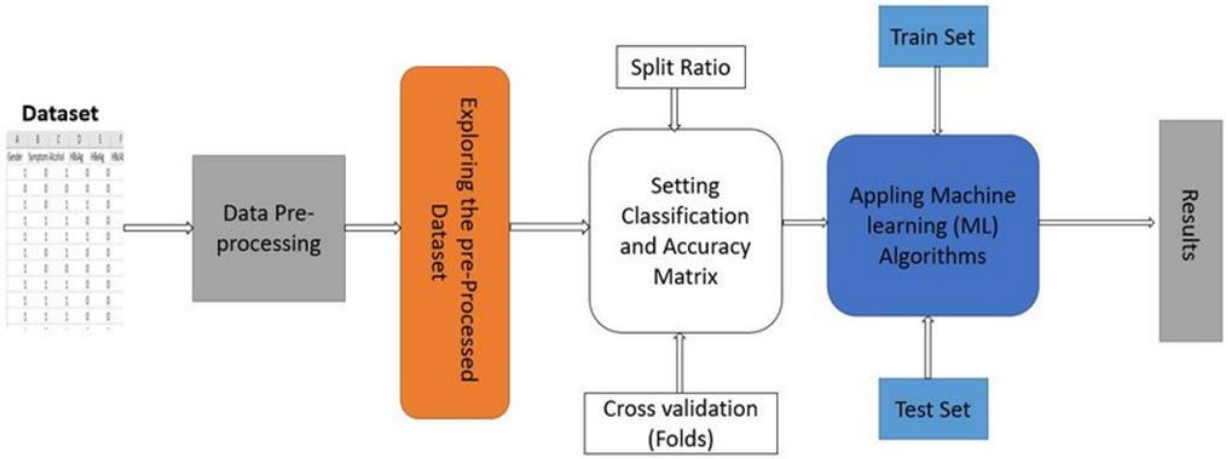
$$F1 = 2 \times \frac{P \times R}{P + R} \qquad (3)$$

Dəqiqlik (Accuracy) aşağıdakı kimi müəyyən edilir:

$$A = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \qquad (4)$$

Yaxşı nəticələr əldə etmək və həddən artıq verilənlərin bir-biri ilə əlaqəsini tapmaq, əlaqəsiz verilənlərdən "qaçmaq" üçün model təliminin keçirilməsi mühüm addımdır. Modelin həddən artıq uyğun və uyğunsuzluğunun qarşısını almaq üçün verilənlərin 90%-i təlim verilənləri, qalan 10%-i isə test verilənləri kimi seçilmişdir. Təlim və test verilənlərinin 90/10 nisbətində seçilməsinin yaxşı klassfiikator dəqiqliyinə nail olmaq üçün kifayət edəcəyi nəzərdə tutulmuşdur. Verilənlər bazasından düzgün istifadə etmək və ən yüksək dəqiqlik nəticələrini əldə etmək üçün [14]-də təqdim edilmiş qiymətləndirmə meyarlarından istifadə edilmişdir.

Dördüncü addım: **klassifikatorlar və ya maşın təlimi alqoritmlərinin tətbiqi və nəticələrin təhlili.** Əldə edilən proqnozu görmək üçün müxtəlif klassifikatorlar və ya maşın təlimi alqoritmləri tətbiq olunur. Bu məqsədlə üç növ maşın təlimi alqoritmindən istifadə edilmişdir: Random Forest (RF), Vektor Maşın (SVM) və Logistik Reqressiya.
HSK-nın proqnozlaşdırılmasında maşın təlimi metodlarının tətbiqi sisteminin arxitekturası şəkil 6-da göstərilmişdir.
Nəticənin təhlili üçün Anaconda mühitində Jupiter proqramından istifadə edilmişdir. Nəticələrdə dəqiqlik əldə etmək və həmçinin verilənlərin müxtəlif təsnifatlandırıcılarda necə işlədiyini görmək üçün fərqli klassifikatorlar seçilmişdir. Cədvəl 2-də dəqiqlik matrisinin meyarlarının qiymətləri və klassifikatorların tətbiqindən alınan dəqiqlik meyarlarının nəticələri verilmişdir. Xətalar matrisi modelin düzgünlüyünü və dəqiqliyini qiymətləndirmək üçün istifadə edilən ən asan və ən sadə yanaşmalardan biridir.



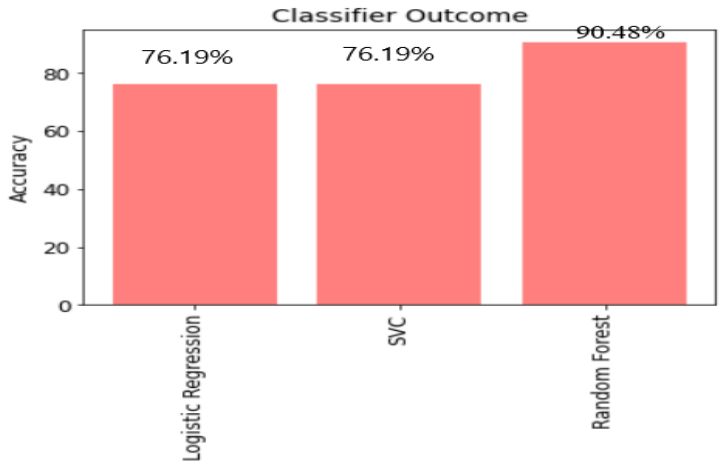
ISSN: 1609-1620, E-ISSN: 2674-5224
PAHTEI
REFEREED & REVIEWED JOURNAL
E-ISSN: 2674-5224
VOLUME 22 ISSUE 11 2022

Şəkil 6. HSK-nın proqnozlaşdırılmasında maşın təlimi metodlarının tətbiqi sisteminin arxitekturası [16]

Cədvəl 2. Xəta Matrisi və performansın ölçülməsi

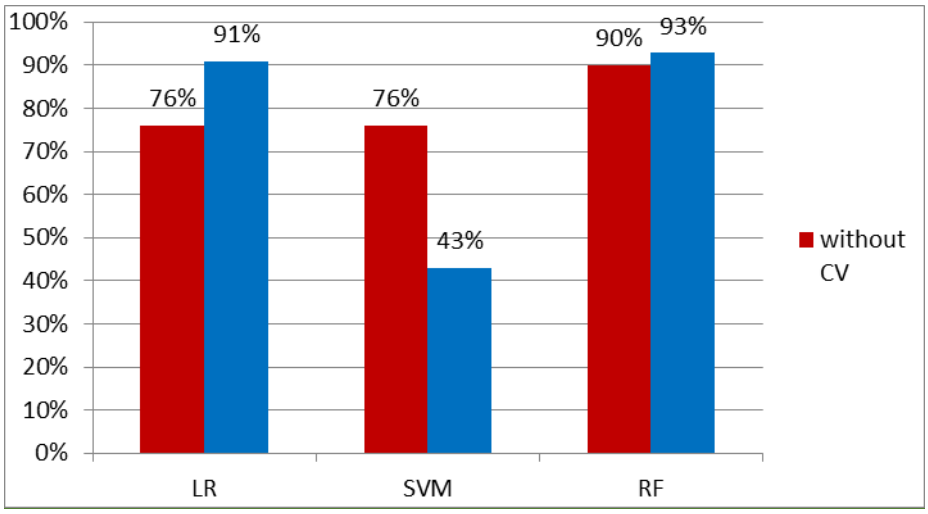
Classifier	TP	FP	FN	TN	P	R	F1
Vektor Maşın (SVM)	9	3	2	7	0.78	0.70	0.74
Random Forest (RF)	10	1	1	9	0.90	0.90	0.90
Logistik Reqressiya(LR)	9	3	2	7	0.78	0.70	0.74

Klassifikatorların tətbiqindən alınan dəqiqlik meyarının qiymətinin qrafik təsviri şəkil 7-də göstərilmişdir.



Şəkil 7. Təsnifləndirmədən əldə edilən dəqiqlik meyarının qiymətinin qrafik təsviri

Sonra *k*-qat çarpaz validasiyadan istifadə edərək çarpaz doğrulama aparılır və müxtəlif klassifikatorlardan əldə edilmiş dəqiqlik meyarlarının qiymətlərinin ədədi ortası götürülərək orta dəqiqlik hesablanır. Təklif edilən eksperimentin səmərəliliyini əsaslandırmaq üçün çarpaz doğrulamadan əvvəl və sonra əldə edilmiş dəqiqlik meyarlarının qiymətlərinin müqayisəsi aparılır və bu müqayisənin qrafik təsviri şəkil 8-də verilmişdir.



Şəkil 8. Çarpaz doğrulamadan əvvəl və sonra əldə edilmiş dəqiqlik meyarlarının müqayisəsinin qrafiki təsviri

Şəkil 8-də şaquli ox üzrə dəqiqlik meyarının qiyməti, üfüqi xətt üzrə fərqli klassifikatorlar göstərilmişdir. Çarpaz doğrulama tətbiq etdikdən sonra klassifikatorlar ilə dəqiq və sabit dəqiqliyə nail olunduğu müşahidə olunur. Random Forest ilə çarpaz doğrulama 93% dəqiqliyə və Logistik Regression ilə isə 91% dəqiqliyə nail olmaq mümkün olmuşdur.

Nəticə

Məqalədə Kaggle platforması HCC Dataset-dən istifadə etməklə HSK-nın proqnozlaşdırılması üçün maşın təlimi metodlarının tətbiqi imkanı araşdırılmışdır. 165 pasiyentin məlumatları əsasında yaradılan verilənlər bazasındakı 49 xüsusiyyət/atribut əsasında HSK-nın proqnozlaşdırılması üçün Random Forest, Vektor Maşın və Logistik Reqressiya maşın təlimi alqoritmlərindən istifadə edilmişdir.

Proqnozlaşdırma alqoritminin ilk mərhələsində bir-biri ilə əlaqəli olan xüsusiyyətlər təyin edilmiş, min-max schale əsasında 49 xüsusiyyət/atribut normallaşdırılmışdır. Baxilan 3 alqortimin tətbiqi ilə xəta matrisi qurulmuş və xəta matrisi meyarları hesablanmışdır. Müəyyən edilmişdir ki, xəta matrisi meyarları üzrə Random Forest ən yüksək dəqiqliklə nəticə göstərir. Alqoritmlərin performasının qiymətləndirilməsi üçün çarpaz doğrulama aparılmış, çarpaz doğrulamadan əvvəl və sonra bu alqortmin yüksək nəticə göstərdiyi müəyyən edilmişdir.

Çarpaz doğrulama tətbiq etmədən Random Forest alqoritmi ilə ən yüksək dəqiqliyə (90,00%) nail olunmuş, çarpaz doğrulamadan sonra isə bu alqoritm 93%-lik ən yüksək dəqiqlik nümayiş etdirmişdir.

Müəlliflərin perspektiv tədqiqatları Random Forest klassifikatoruna istinad ettməklə HSK--nın proqnozlaşdırılması sisteminin işlənilməsidir. Bununla yanaşı digər maşın təlimi alqoritmləri və ANN (Artificial Neural Networks), CNN (Convolution Neural Networks) kimi dərin təlim yanaşmalarından da istifadə etməklə proqnozların dəqiqliyinin yaxşılaşdırılması istiqamətində eksperimentlərin aparılması da nəzərdə tutulur.

ƏDƏBİYYAT

1. Memmedova M.H., Cebrayılova Z.Q. Elektron tibb: formalashması və elmi-nezeri problemleri.- Bakı: "İnformasiya Texnologiyaları" nəshriyyatı, 2019, 350 səh.
2. S. Singh and D. Hanchate, "Improving disease prediction by machine learning," 06 2018.
3. Bayramov, N. Y.: In book: Surgical diseases of the liver.- Baku: Qismet, 2012.
4. Xiaopu S., Fenfang W., Di W., Shan L., Jingyi L., Nan Z., Xiaoni C., Anlong X. Human Hepatic Cancer Stem Cells (HCSCs) Markers Correlated With Immune Infiltrates Reveal Prognostic Significance of Hepatocellular Carcinoma // Frontiers in Genetics., 28 February 2020. https://doi.org/10.3389/fgene.2020.00112
5. Calderaro, J., Seraphin, T. P., Luedde, T., Simon, T. G.: Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma.// Journal of Hepatology.- 2022, 76, 1348-1361
6. D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining,"/ in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1–5.
7. M. Santos, P. Henriques Abreu, P. Garc´ıa-Laencina, A. Simao, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients,"// Journal of biomedical informatics.- 10 2015, vol. 58, pp. 49–59,.
8. Aman S., Babita P. An Efficient Diagnosis System for Detection of Liver Disease Using a Novel Integrated Method Based on Principal Component Analysis and K-Nearest Neighbor (PCA-KNN) // International Journal of Healthcare Information Systems and Informatics.- 2016, vol.11, no.4, pp.56–61.
9. Sartakhti J.S., Zangooei M.H., Mozafari K. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA) // Computer Methods and Programs in Biomedicine.- 2015, vol.108, no.2, pp.570–579.
10. Gorunescu F., Belciug S., Gorunescu M., Badea R. Intelligent decision-making for liver fibrosis stadialization based on tandem feature selection and evolutionary-driven neural network // Expert Systems with Applications.- 2012, vol.39, no.17, pp.12824–12832.
11. Mammadova, M. G., Bayramov N. Y., Jabrayilova Z. G. Development principles of fuzzy rule-based system for hepatocelular carcinoma staging, Eureka:physics and engineering.- 2021, no.3, pp.3-13.
12. Mammadova, M. G., Bayramov N. Y., Jabrayilova Z. G., Manafli M. I., Huseynova M. R. 8th Conference on Control and Optimization with Industrial Applications-COIA'2022 24-26 August.- 2022, Baku, Azerbaijan, pp. 318-320.
13. www.kaggle.com
14. S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over holdout validation on colossal datasets for quality classification,"/ in 2016 IEEE 6th International Conference on Advanced Computing (IACC).- 2016, pp. 78–83.
15. C. R. Harris, K. J. Millman, and et.al."Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2
16. 2020 IEEE 17th India Council International Conference (INDICON) | 978-1-7281-6916-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/INDICON49873.2020.9342443

E-ISSN: 2674-5224
VOLUME 22 ISSUE 11 2022
ISSN: 1609-1620, E-ISSN: 2674-5224
PAHTEI
REFEREED & REVIEWED JOURNAL

ВЫБОР АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАЗРАБОТКИ СИСТЕМЫ ПРОГНОЗИРОВАНИЯ ГЕПАТОЦЕЛЛЮЛЯРНОЙ КАРЦИНОМЫ

Масума Мамедова[1], Зарифа Джабраилова[2], Лала Караева[3]
[1,2,3]Институт информационных технологий НАНА, [1,2,3]Отдел №11
[1]заведующий отделом, член-корреспондент НАНА, доктор технических наук, профессор, http://orcid.org/0000-0002-2205-1023, Email: mmg51@mail.ru
[2]главный научный сотрудник, доктор технических наук, доцент, http://orcid.org/0000-0002-9661-5805, Email: djabrailova_z@mail.ru
[3]младший научный сотрудник, Email: karayevalala.01@gmail.com

РЕЗЮМЕ

В статье рассмотрены возможности использования алгоритмов машинного обучения для разработки системы прогнозирования гепатоцеллюлярной карциномы (ГЦК), являющейся наиболее распространенной среди злокачественных опухолей печени. ГЦК характеризуется набором клинических проявлений критических состояний, каждое из которых, в свою очередь определено множеством клинических признаков, данных. Для прогнозирования ГЦК на основе таких многочисленных, разнотипных и разнородных неструктурированных данных предпочтение дано методу исскусственного интеллекта – машинному обучению. С этой целью использована база данных HCC Dataset из платформы Kaggle, состоящая из 49 признаков/атрибутов 165 пациентов. Для первичной обработки данных использованы библиотеки scikit-learn, Pandas, NumPy с использованием программы Jupiter. Проведены эксперименты на базе алгоритмов Logistic Regression, Support Vector Machine, Random Forest и представлены результаты оценки их эффективности. На основе сравнительного анализа полученных данных выявлен алгоритм с наилучшими показателями эффективности, в данном случае – алгоритм Random Forest.
Ключевые слова: гепатоцеллюлярная карцинома, интеллектуальная система прогнозирования, методы машинного обучения, параметры эффективности.

Publication history

Article received: 22.10.2022
Article accepted: 05.11.2022
Article published online: 16.11.2022
DOI suffix: 10.36962/PAHTEI22112022-116