https://doi.org/10.59849/aidd.2024.13

Solving the Prognostic Problem of Hepatocellular Carcinoma Using Machine Learning Algorithms in a National Database

Masuma Mammadova[1], Nuru Bayramov[2], Zarifa Jabrayilova[1], Lala Karayeva[1], Mehriban Huseynova[2]

[1]Institute of Information Technologies, Baku, Azerbaijan,
E-mail: karayevalala.01@gmail.com
[2]I Department of Surgical Diseases, Azerbaijan Medical University, Baku, Azerbaijan

Abstract: The emergence of digital medicine has paved the way for expanding research aimed at providing physicians with information support in medical decision-making and preventing medical errors. This trend has made the application of artificial intelligence methods highly relevant in preventing the spread, early diagnosis, and prognosis of hepatocellular carcinoma (HCC), which ranks third globally in cancer-related deaths. In this study, machine learning algorithms such as Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) are applied to predict HCC using the HCC Dataset obtained from the Kaggle platform. A prediction algorithm based on the national HCC dataset is proposed by utilizing the RF method, which yielded the most significant results. The MICE algorithm is applied to impute the missing values in the national HCC database, and data from this database are gradually transferred to the Kaggle HCC Dataset (where prediction outcomes are known) to obtain prognostic results based on clinical patient data.

Keywords: Prediction of hepatocellular carcinoma, machine learning algorithms, filling in missing data in the database, accuracy criteria.

Hepatosellular karsinoma üzrə milli bazada maşın təlimi alqoritmlərinin tətbiqi ilə proqnozlaşdırma məsələsinin həlli

Məsumə Məmmədova[1], Nuru Bayramov[2], Zərifə Cəbrayılova[1], Lalə Qarayeva[1], Mehriban Hüseynova[2]

Annotasiya. Rəqəmsal tibbin formalaşması tibbi qərar qəbuletmədə həkimlərə informasiya dəstəyinin göstərilməsi, həkim səhvlərinin qarşısının alınması istiqamətində tədqiqatların genişlənməsinə zəmin yaratmışdır. Bu tendensiya xərçəng səbəbindən ölənlərin sayına görə dünyada üçüncü yeri tutan hepatosellular karsinomanın (HSK) yayılmasının qarşısının alınması, onun ilkin diaqnozu və proqnozlaşdırılması üçün süni intellekt metodlarının tətbiqini aktuallaşdırmışdır. Tədqiqatda Kaggle platformasından götürülmüş *HCC Dataset* verilənlər bazasından istifad etməklə HSK-nın proqnozlaşdırılması üçün *Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF)* kimi maşın təlimi alqoritmləri tətbiq edilmiş və daha əhəmiyyətli nəticə göstərən RF metodundan istifadə etməklə mövcud milli HSK verilənlər bazasındakı informasiya əsasında proqnozlaşdırma məsələsinin həll alqoritmi təklif edilmişdir. HSK üzrə milli bazadakı boşluqların doldurulması üçün MICE alqoritmi tətbiq edilmiş, kliniki xəstələrin verilənlərinə görə proqnoz nəticələr almaq üçün bu bazadakı məlumatların *Kaggle saytından götürülmüş HCC Datasetinə* (proqnoz nəticələri məlum olan) mərhələ-mərhələ köçürülməsi yerinə yetirilmişdir.

Açar sözlər: hepatosellular karsinomanın proqnozlaşdırılması, maşın təlimi alqoritmləri, verilənlər bazasındakı boşluqların doldurulması, dəqiqlik meyarları.

Giriş

Təcrübəli həkim-ekspertlərin biliklərinin toplanması, saxlanması, manipulyasiyası, eləcə də, hər bir konkret verilənlər toplusu üzrə xəstəliyin müəyyən edilməsi və adekvat qərarların qəbul olunmasıı

üçün daha səmərəli vasitə biliklərə əsaslanan intellektual sistemlərdir. 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, təsvirlərin tanınması və interpretasiyası, dərman vasitələrinin kliniki-farmakoloji xüsusiyyətlərinin (toksikliyinin) monitorinqi və s. məsələlərin həllində tətbiq olunur.

Xərçəng növlərindən biri olan HSK çoxlu sayda kliniki göstəricilər, əlamətlərlə təyin olunan kritiki vəziyyətlərlə özünü biruzə verir. Onun diaqnozu və proqnozlaşdırılması üçün dəqiq, birmənalı meyarlar mövcud deyil [1, 2]. HSK-nın kritiki vəziyyətlərinin müxtəlif tipli, strukturlaşdırılmamış, çoxsaylı göstəricilərlə xarakterizə olunması HSK-nın diaqnozu, proqnozlaşdırılması ilə bağlı qərarların qəbulunda həkim səhvlərinə səbəb olur [3, 4]. Nəticədə göstəriciləri xarakterizə edən kliniki əlamətlərin müəyyən kombinasiyaları ilə təyin edilən həkim qərarlarının qəbulunda xətalar qaçılmaz olur. Bu HSK-nın diaqnostikası və proqnozlaşdırılması üçün süni intellekt metodlarının tətbiqini, kliniki qərarların qəbuluna dəstək sistemlərinin yaranmasını şərtləndirir. Hazırda elmi ədəbiyyatda qaraciyər xəstəliklərinin aşkarlanması, diaqnostikası və müalicəsi üçün intellektual sistemlərin işlənilməsi istiqamətində tədqiqatlara rast gəlinir [5, 6]. Lakin HSK-nın diaqnostikası və proqnozlaşdırılması sistemlərinin işlənilməsinə çox az diqqət ayrılmışdır. [7]-da HSK-nın mərhələsinin təyini üçün qeyri-səlis qaydalara əsaslanan sistemin işlənilməsi metodikası təqdim edilmiş, [8]-də HSK-nın diaqnostikası sisteminin komponentləri və biliklər bazasının formalaşması mexanizmləri verilmiş, [9]-da HSK-nın proqnozlaşdırılması üçün maşın təlimi və dərin təlim metodlarından istifadənin əhəmiyyəti və imkanları araşdırılmışdır. Bu yanaşma pasiyentlərin xəstəlik ilə bağlı toplanmış məlumatları əsasında yaradılmış bazalardan istifadə etməklə proqnoz vermək üçün nəzərdə tutulmuş və bu da xəstəliklə bağlı dəqiq və vaxtında qərar qəbul edilməsi üçün həkumlərə dəstək göstərə bilər.

Hazırkı məqalədə maşın təlimi alqoritmlərinə istinad etməklə qaraciyər xərçəngi kimi tanınan HSK-nın ilkin diaqnozu və proqnozlaşdırılması məsələsinin həlli imkanları araşdırılmış, HSK üzrə milli bazaya daxil olan kliniki xəstələrin vəziyyətinə uyğun proqnoz nəticələrin alınması üçün proqnozlaşdırma məsələsinin həlli alqoritmi təklif olunmuşdur.

2. Problemin qoyuluşu

HSK üzrə milli baza əsasında maşın təlimi alqoritmlərinin tətbiqi ilə proqnozlaşdırma məsələsinin həlli bu bazadakı çoxsaylı boşluqların olması, proqnoz nəticələrin göstərilməməsi səbəbindən problemlər yaradır, daha mükəmməl bazalarda aparılan tədqiqatların nəticələrinə istinad edilməsini tələb edir. Odur ki, maşın təlimi alqoritmlərinin tətbiqilə milli verilənlər bazası əsasında HSK-nın proqnozlaşdırılması üçün aşağıdakı məsələlərin həlli nəzərdə tutulur:

- HSK üzrə mükəmməl bazanın seçilməsi və verilənlərin ilkin emalı;
- HSK-nın proqnozlaşdırılması üçün maşın təlimi alqoritmlərinin tətbiqi və dəqiqlik meyarlarına görə daha yaxşı nəticə göstərən metodun seçilməsi;
- HSK üzrə toplanmış milli verilənlər bazasında boşluqların doldurulması;
- HSK üzrə milli verilənlər bazasına daxil olan kliniki xəstələrin vəziyyətinə uyğun proqnoz nəticələrin alınması üçün seçilmiş maşın təlimi alqoritminin tətbiqi.

3. Problemin həlli

3.1. HSK üzrə mükəmməl bazanın seçilməsi və verilənlərin ilkin emalı

HSK-nın proqnozlaşdırıılmasında maşın təlimi alqoritmlərinin tətbiqi üçün *Kaggle* şirkətinin *HCC Dataset* [10] adlı açıq verilənlər bazasından istifadə olunmuşdur. Verilənlər bazası Portuqaliya Universiteti Xəstəxanasında HCC-dən 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.

Bu baza əsasında maşın təlimi alqoritmlərinin tətbiqi ilə HSK-nın proqnozlaşdırıılması üçün verilənlərin ilkin emalı yerinə yetirilir. Bu mərhələdə bazada bir-biri ilə əlaqəli olmayan verilənlərin

olub-olmaması araşdırılır, səpələnmiş verilənlər müəyyənləşdirilir, onların vahid formaya gətirilməsi həyata keçirilir. Bu mərhələ verilənlər bazasının təmizlənməsi prosesi də adlandırılır və bilavasitə istifadəçinin müdaxiləsi ilə yerinə yetirilir. Verilənlər bazasını faydalı etmək, verilənin işlənməsini sadələşdirmək üçün *Pandas* [11] və *NumPy* [12] kitabxanalarından istifadə edilir. Kodların tətbiqi ilə müxtəlif tip verilənlər eyni tip verilənlərə çevrilir və bir-biri ilə əlaqəli xüsusiyyətlər təyin edilir. Verilənlər arasında həm xətti, həm də qeyri-xətti əlaqələri tapmaq üçün *Correlation heatmaps* istifadə edilir. Məlumatları normallaşdırmaq üçün minimum-maksimum miqyaslama bir və ya bir neçə xüsusiyyət sütununa tətbiq olunur. Minimum-maksimum miqyaslama əsasında verilənlərin normallaşdırılmaq üçün şəkil 1-dəki kodlaşdırmadan istifadə olunur. Daha sonra, *HCC Dataset*-də olan verilənlərin bütün xüsusiyyətləri təhlil olunmuş, buradakı 49 kliniki əlamət/atributdan 23-ünün kəmiyyət, 26-sının keyfiyyət xarakterli olduğu müəyyənləşdirilmişdir. Hədəf sinfi (class) – birillik hesabata görə, 0 (ölən) və 1 (yaşayan) kimi ikili dəyişəndən ibarət olan proqnoz nəticədir.

3.2. HSK-nın proqnozlaşdırılması üçün maşın təlimi alqoritmlərinin tətbiqi və dəqiqlik meyarlarına görə daha yaxşı nəticə göstərən metodun müəyyənləşdirilməsi

HCC Dataset-də təsnifləndirməni aparmaq üçün *LR, SVM* və *RF* maşın təlimi alqoritmləri istifadə olunmuşdur. Maşın təlimində klassifikatorların aşkarlama performansının qiymətləndirilməsi üçün həssaslıq, tamlıq, yanlış pozitiv hallar, doğru pozitiv hallar, f-ölçü, dəqiqlik meyarlarından istifadə olunmuşdur [11].

```
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 1. Qaraciyər xərçəngi məlumat dəstinin *MinMaxScaler* miqyaslama

Həssaqlıq *(P)* həqiqi müsbətlərin sayı kimi müəyyən edilir və aşağıdakı kimi təyin edilir:

$$P = \frac{Tp}{Tp+Fp} \tag{1}$$

Burada: *Tp* – 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:

$$R = \frac{Tp}{Tp+Fn} \tag{2}$$

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

***F*1-ölçü** (F1-Score) və geri çağırmanın harmonik ortası kimi müəyyən edilir və aşağıdakı düsturla hesablanır:

$$F1 = 2 * \frac{P*R}{P+R} \tag{3}$$

***Dəqiqlik* (Accuracy)** aşağıdakı kimi müəyyən edilir:

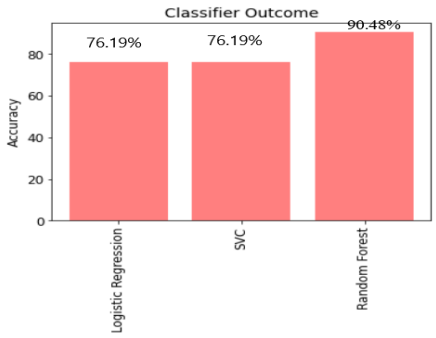
$$A = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (4)$$

Modelin uyğunluğunun qarşısını almaq üçün verilənlərin 90%-i təlim verilənləri, qalan 10%-i isə test verilənləri kimi seçilmişdir. *LR, SVM və RF* maşın təlimi alqoritmləri tətbiq edilmişdir. Nəticənin təhlili üçün *Anaconda* mühitində *Jupite*r proqramından istifadə edilmişdir. Cədvəl 1-də xətalar matrisinin meyarlarının qiymətləri və klassifikatorların tətbiqindən alınan dəqiqlik meyarlarının nəticələri verilmişdir.

Klassifikatorların tətbiqindən alınan dəqiqlik meyarının qiymətinin qrafik təsviri şəkil 2-də göstərilmişdir. Şəkil 2-dən göründüyü kimi, xəta matrisi meyarları üzrə *RF* ən yüksək dəqiqliyə malik nəticə göstərmişdir

Cədvəl 1. Xəta Matrisi və performansın ölçülməsi

Classifier	TP	FP	FN	TN	P	R	F1
SVM	9	3	2	7	0.78	0.70	0.74
RF	10	1	1	9	0.90	0.90	0.90
LR	9	3	2	7	0.78	0.70	0.74



Şəkil 2. Təsnifləndirmədən alınan dəqiqlik meyarının qiymətinin qrafik təsviri

3.3. HSK üzrə toplanmış milli verilənlər bazasında boşluqlar doldurulması

Azərbaycan Tibb Universitetinin I cərrahi xəstəliklər kafedrasının Türkiyənin Malatya İnönü Universitetinin Qaraciyər Nəqli İnstitutu ilə birgə fəaliyyəti əsasında yaradılmış HSK üzrə kliniki xəstələr haqqında verilənlər bazası (*milli HCC Dataset*) tədqiqat mənbəyi seçilmişdir. Verilənlər bazası 27 göstərici üzrə 556 kliniki xəstənin məlumatları əsasında formalaşdırılmışdır (şəkil 3) (bazadakı boşluqlar şəkil 1-də qırmızı halqalarla göstərilmişdir).

```
ilk100 = Tb.head(50)

display(ilk100)
```

	Gender	Age	HBsAg	HCVAb	HIV	Varices	Spleno	PVT	Metastasis	Encephalopathy	...	Total_Bil	ALT	AST	GGT	ALP	Creatinine	Nodule	N
0	1	42.0	1.0	0.0	0.0	1.0	1.0	1.0	NaN	0.0	...	0.6	23.0	36.0	49.0	79.0	0.8	1.0	
1	1	67.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0.0	...	0.5	58.0	111.0	478.0	423.0	1	4.0	
2	0	56.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0.0	...	0.6	51.0	38.0	48.0	84.0	0,7	4.0	
3	1	60.0	1.0	0.0	0.0	1.0	0.0	0.0	NaN	0.0	...	0.51	39.0	37.0	114.0	112.0	1,2	2.0	
4	1	57.0	1.0	0.0	0.0	1.0	0.0	0.0	0	0.0	...	1.44	32.0	48.0	128.0	117.0	0,77	2.0	
5	0	60.0	1.0	0.0	0.0	1.0	0.0	0.0	1	0.0	...	0.93	72.0	75.0	252.0	222.0	0.62	4.0	
6	1	61.0	1.0	0.0	0.0	1.0	1.0	0.0	NaN	0.0	...	0.62	72.0	40.0	46.0	57.0	0.77	2.0	
7	1	75.0	1.0	0.0	0.0	1.0	0.0	0.0	NaN	0.0	...	0.8	39.0	41.0	207.0	167.0	0,83	1.0	

Şəkil 3. M*illi HCC Dataset*-inin fraqmenti

Qeyd edək ki, bazadakı 556 yazıdan (row) yalnız 50 yazıda proqnoz nəticələr (0 (ölüm) və 1 (yaşayan)) göstərilmişdir. Bu baza əsasında kliniki xəstələrin vəziyyətinin proqnozlaşdırılması (yəni, hər bir yazı üzrə proqnoz nəticənin alınması) üçün ilk növbədə istifadəçinin müdaxiləsi ilə bazadakı müxtəlif tipli verilənlər eyni tipə (float) gətirilmiş, bazadakı boşluqların doldurulması üçün MICE (*ом англ. multivariate imputation by chained equations)*alqoritmi seçilmişdir. Bu alqoritm ilə boşluqlarda olacaq dəyişənlər üçün posterior ehtimal paylamalarını əldə etmək məqsədilə Markov sxemi üzrə

Monte-Karlo metodu istifadə edilir. Bu metod verilənlər bazasında çoxlu sayda dəyişənlər üzrə boşluqlar olduğu halda istifadə üçün yararlıdır. Boşluqların doldurulması üçün orta qiymətə görə boşluğun doldurulması, yaxın qonşular metodu, boşluqların reqression modelləşdirilməsi və s. kimi müxtəlif metodlar vardır [13, 14]. Beləkilə, MICE alqoritminin tətbiqi ilə milli HSK bazasındakı boşluqların doldurulmasından sonrakı təsviri şəkil 4-də verilmişdir.

	Gender	Age	HBsAg	HCVAb	HIV	Varices	Spleno	PVT	Metastasis	Encephalopathy	...	Total_Bil	ALT	AST	GGT	ALP	Creatinine	Nodu
0	1	42.0	1.0	0.0	0.0	1.000000	1.0	1.000000	0.271354	0.000000	...	0.60	23.0	36.0	49.0	79.0	0.800000	1.00000
1	1	67.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	...	0.50	58.0	111.0	478.0	423.0	1.000000	4.00000
2	0	56.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	...	0.60	51.0	38.0	40.0	84.0	0.700000	4.00000
3	1	60.0	1.0	0.0	0.0	1.000000	0.0	0.000000	0.283048	0.000000	...	0.51	39.0	37.0	114.0	112.0	1.200000	2.00000
4	1	57.0	1.0	0.0	0.0	1.000000	0.0	0.000000	0.000000	0.000000	...	1.44	32.0	49.0	128.0	117.0	0.770000	2.00000
5	0	60.0	1.0	0.0	0.0	1.000000	0.0	0.000000	1.000000	0.000000	...	0.93	72.0	75.0	252.0	222.0	0.620000	4.00000
6	1	61.0	1.0	0.0	0.0	1.000000	1.0	0.000000	0.262002	0.000000	...	0.62	72.0	40.0	46.0	57.0	0.770000	2.00000
7	1	75.0	1.0	0.0	0.0	1.000000	0.0	0.000000	0.438280	0.000000	...	0.80	39.0	41.0	207.0	167.0	0.830000	1.00000

Şəkil 4. MICE alqoritminin tətbiqindən sonra *milli HCC Datasetin*-dən bir fraqmentin təsviri

3.4. Milli verilənlər bazasına daxil olan kliniki xəstələrin vəziyyətinə uyğun proqnoz nəticələrin alınması üçün seçilmiş maşın təlimi alqoritminin tətbiqi

Milli verilənlər bazasında proqnoz nəticələrin əksəriyyətinin göstərilməməsi (50/556) maşın təlimi metodları əsasında proqnozlaşdırma məsələsinin həlli üçün 3.2. bəndində aldığımız nəticəyə görə *RF* alqoritmi tətbiq edilmişdir. Bu məqsədlə milli bazadan proqnoz nəticələri məlum olan 50 yazı *Kaggle şirkətinin HCC Datasetinə* əlavə edilməklə sonuncu genişləndirilir (204+50 yazı). Genişləndirilmiş HCC Dataseti əsasında proqnoz nəticələr almaq üçün verilənlərin *Kaggle şirkətinin HCC Datasetindəki* 204 yazı (80%) təlim məlumatları, milli bazadan əlavə olunan 50 yazı (20%) isə test məlumatları kimi qəbul edilir və *RF* alqoritmi tətbiq edilməklə proqnoz nəticələrin alınması yerinə yetirilir (şəkil 5).

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

# Bazanı yükləmək və null dəyərlərə sahib olan sütunu ayırmaq
NT_not_null = NT[NT['Class'].notnull()]
NT_null = NT[NT['Class'].isnull()]

# təlim və test hissələrini yaratmaq  (null olanları test setinə daxil etmək)
NT_təlim, NT_test = train_test_split(NT_not_null, test_size=0.2, random_state=42)

# təlim setindən null olan dəyişənləri çıxarmaq
NT_təlim = NT_təlim.dropna(subset=['Class'])

# təlim setini X və y olaraq ayırmaq
X_təlim = NT_təlim.drop('Class', axis=1)
y_təlim = NT_təlim['Class']

# Modeli öyrətmək (təlim setindəki non-null dəyərlərlə)
model = RandomForestClassifier(random_state=42)
model.fit(X_təlim, y_təlim)

# Test setini X və y olaraq ayırın
X_test = NT_test.drop('Class', axis=1)
y_test = NT_test['Class']

# Modelin performansını qiymətləndirmək (test setindəki non-null dəyərlərlə)
doğruluq = model.score(X_test, y_test)
print("Model Doğruluğu:", doğruluq)

# Modeli istifadə edərək null olan dəyərləri təxmin et
təxminlər_null = model.predict(NT_null.drop('Class', axis=1))

# Təxminləri yazdırın
print("Null olan gözlemler için təxminlər:", təxminlər_null)
```

```
Model Doğruluğu: 0.7560975609756098
Null olan gözlemler için təxminlər: [1. 0. 1. 1. 1. 1. 0. 1. 1. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0. 0. 0. 1. 0. 0.
 1. 1. 1. 1. 1. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.
 0. 1.]
```

Şəkil 5. Genişləndirilmiş *HCC Datasetinə* RF alqoritminin tətbiqi və alınmış nəticələr

Alınan proqnoz nəticələrin dəqiqliyi 75,6% olmuşdur, retrospektiv nəticələrlə müqayisə 44/6 nisbətində (yəni 50 yazı üçün alınan nəticədən 47-sinin üst-üstə düşdüyü (xəta 14%)) olmuşdur. Xətanın yüksək olması səbəbindən genişləndirilmiş *HCC Dataseti* əsasında alınan nəticələrin yaxşılaşdırılması üçün *XGBClassification* tətbiq edilmişdir. Genişləndirilmiş *HCC Datasetinə XGBClassificatio*n tətbiqindən alınan proqnoz nəticələrin daha yüksək dəqiqlik (78,05%) göstərdiyi, nəticələrin retrospektivlə müqayisəsinin 47/3 nisbətində (50 nəticədən 47-nin üst-üstə düşdüyü (xəta 6%)) olduğu müəyyənləşdirilmişdir. Beləliklə, milli HCC Datasetindəki digər kliniki xəstələrin məlumatlarının hissə-hissə (genişləndirilmiş *HCC Datasetindəki* yazıları hər dəfə 80% qəbul edilməklə,

milli bazadan 20% yazının seçilməsi) seçilərək genişləndirilmiş *HCC Datasetinə daxil edilməsi* və *XGBClassification* tətbiqi əsasında proqnoz nəticələrin alınması yerinə yetirilə bilər.

4. Nəticə

- *Milli HCC Dataseti* əsasında proqnozlaşdırma məsələsinin maşın təlimi metodlarının tətbiqi ilə həlli üçün HSK üzrə daha mükəmməl baza olaraq *Kaggle* saytından götürülmüş *HCC Dataseti* seçilmişdir;
- *HCC Datasetin*-dəki yazıların 90% təlim və 10%-i test yazısı kimi seçilmiş, *LR, SVM və RF* maşın təlimi alqoritmləri tətbiq edilmiş, *RF* alqoritminin ən yüksək dəqiqliyə malik nəticə göstərdiyi təyin olunmuşdur;
- HSK üzrə milli bazadakı boşluqların doldurulması üçün MICE alqoritmi seçilmiş və tətbiq edilmişdir;
- HSK üzrə milli bazadakı 50 yazı (proqnoz nəticələri məlum olan) *Kaggle* saytından götürülmüş *HCC Datasetinə* əlavə edilmiş və baza genişləndirilmişdir. Genişləndirilmiş *HCC Datasetində* proqnoz nəticələrin alınması üçün RF alqoritmi və *XGBClassification* tətbiqi edilmişdir, milli *HCC Datasetindəki* hər bir yazı üzrə proqnoz nəticənin alınması imkanı göstərilmişdir.

Ədəbiyyat

1. Bayramov, N. Y.: Surgical diseases of the liver. Baku: Qismet (2012).
2. Huang, D. Q., Tran, A., Tan. E.X., Nerurkar, S.N., Teh, R., Teng, M.L.P., Yeo, E.J., Zou, B., Wong, C., Esquivel, C.O., Bonham, C.A, Nguyen, M.H.: Characteristics and outcomes of hepatocellular carcinoma patients with macrovascular invasion following surgical resection: a meta-analysis of 40 studies and 8,218 patients. Hepatobiliary Surg Nutr 11(6), 848-860 (2022).
3. Jennifer, J. R., Brit, L.: Suffering in Silence: Medical Error and its Impact on Health Care Providers. The Journal of Emergency Medicine, 54(4), 402–409 (2017).
4. Attia, B., Rehan, A.K., Ahsan, W.R.: Medical errors: causes, consequences, emotional response and resulting behavioral change. Pakistan Journal of Medical Sciences, 32(3), 523–528 (2016).
5. 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, 11(4), 56–61 (2016).
6. 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, 108(2), 570–579 (2015).
7. Mammadova, M. G., Bayramov, N. Y., Jabrayilova, Z. G.: Development principles of fuzzy rule-based system for hepatocelular carcinoma staging. Eureka: physics and engineering, 3, 3–13 (2021).
8. Mammadova, M. G., Bayramov, N. Y., Jabrayilova, Z. G., Manafli, M. I., Huseynova, M. R.: Knowledge transformation in the intelligent system for hepatocellular carcinoma staging. Proceedings of the 8th International Conference on Control and Optimization with Industrial Applications (COIA'2022), 24-26 August 2022, Baku, Azerbaijan, vol.1, 318–321.
9. Mammadova, M. G., Jabrayilova Z. G., Garayeva L. A., Ahmadova A. A. Prediction of hepatocellular carcinoma using a machine learning // The 16th IEEE International Conference Application of Information and Communication Technologies (AICT-2023), Washington DC, 12-14 Oct 2022, INSPEC Accession Number: 22541899 (2023).
10. Santos, M., Henriques P.A., Garc´ıa-Laencina, P., Simao, A., and Carvalho, A. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients, Journal of biomedical informatics, vol. 58, pp. 49–59, (2015).

11. Yadav, S., Shukla, S. Analysis of k-fold cross-validation over holdout validation on colossal datasets for quality classification, 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 78–83 (2016).

12. Harris, C. R., Millman, K. J., and et.al. Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online].

13. Little RJA, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 408 p. (2014).

14. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Annals of Translational Medicine, vol. 4, no. 2, pp. 30. (2016).