

PROSTATA SARATONIDA IZOHLANADIGAN, YUQORI ANIQLIKDAGI AI PIPELINE

F.Madolimov – dotsent
Andijon davlat universiteti

ИНТЕРПРЕТИРУЕМЫЙ ВЫСОКОТОЧНЫЙ AI-КОНВЕЙЕР ПРИ РАКЕ ПРЕДСТАТЕЛЬНОЙ ЖЕЛЕЗЫ

Ф.Мадолимов – доцент
Андижанский государственный университет

HIGH-PRECISION AI PIPELINE EXPLAINED IN PROSTATA CANCER

F.Madolimov – Associate Professor
Andijan State University

Tayanch soʻzlar: prostata saratoni, mashinaviy oʻrganish, feature engineering, PCA, SHAP, diagnostika.

Rezyume. Mazkur maqolada prostata saratonini erta aniqlash uchun mashinaviy oʻrganish metodologiyasi asosida chuqur feature engineering jarayoni amalga oshirildi. Maʼlumotlar toʻplami klinik koʻrsatkichlar — PSA darajasi, bemor yoshi, prostata hajmi, Gleason balli va klinik bosqich kabi diagnostik omillardan tashkil topgan boʻlib, ularni qayta ishlash jarayonida ANOVA, Chi-square, PCA, RFE, LASSO va SHAP usullari qoʻllanildi.

Ключевые слова: рак предстательной железы, машинное обучение, инженерия признаков, PCA, SHAP, диагностика.

Резюме. В данной работе представлен углублённый процесс feature engineering (инженерии признаков), направленный на раннее выявление рака простаты с использованием методологии машинного обучения. Исходный набор данных включал клинические показатели — уровень PSA, возраст пациента, объём простаты, балл по шкале Глисона и клиническую стадию заболевания. Для их обработки применялись методы ANOVA, Chi-square, PCA, RFE, LASSO и SHAP. Цель исследования заключалась в определении наиболее значимых диагностических признаков, повышающих эффективность модели, и обеспечении её интерпретируемости.

Key words: prostate cancer, machine learning, feature engineering, PCA, SHAP, diagnostics.

Summary. This study presents a comprehensive feature engineering process for the early detection of prostate cancer using machine learning methodology. The dataset consisted of key clinical indicators — PSA level, patient age, prostate volume, Gleason score, and clinical stage — which were processed using ANOVA, Chi-square, PCA, RFE, LASSO, and SHAP techniques. The primary objective was to identify the most influential diagnostic features that improve model performance and ensure interpretability.

Kirish. Prostata saratoni (PS) erkaklar orasida uchraydigan onkologik kasalliklar ichida yuqori uchrashuv chastotasiga ega boʻlib, jahon sogʻliqni saqlash tizimi oldida muhim tashxisiy va davolash muammolarini yuzaga keltirmoqda. Jahon sogʻliqni saqlash tashkiloti (WHO) va Globocan 2023-yilgi statistik maʼlumotlariga koʻra, prostata saratoni erkaklarda oʻpka saratonidan keyin ikkinchi oʻrinda uchraydigan xavfli oʻsma boʻlib, har yili millionlab yangi holatlar qayd etilmoqda. Kasallikning asosiy xavfi uning sekin rivojlanishiga qaramay, dastlabki bosqichlarda klinik belgilar deyarli sezilmasligidir, natijada tashxis kech qoʻyiladi va davolash samaradorligi pasayadi. Shuning uchun erta diagnostika mexanizmlarini takomillashtirish onkourologiya sohasida eng ustuvor vazifalardan biri sifatida qaralmoqda.

Adabiyotlar tahlili va metodologiya. Ushbu maqola prostata saratonini klinik koʻrsatkichlar asosida erta aniqlash uchun mashinaviy oʻrganish yondashuvlari, xususan Feature Engineering texnikalarining diagnostik samaradorligini oʻrganishga qaratildi. Boʻlimda maʼlumotlar toʻplami,

ularni dastlabki qayta ishlash jarayonlari, belgilarni tanlash va oʻlchamni qisqartirish usullari, modellarni oʻqitish jarayoni hamda baholash mezonlari bosqichma-bosqich bayon qilinadi.

Maʼlumotlar toʻplami (Dataset) [3].

Maqolada ishlatilgan klinik maʼlumotlar quyidagi diagnostik koʻrsatkichlardan tashkil topgan:

Belgilar (Feature)	Taʼrifi	Tip
Age	Bemor yoshi	Raqamli
PSA (Prostate Specific Antigen)	Qondagi PSA konsentratsiyasi	Raqamli
Prostate Volume	Prostata bezining hajmi (TRUS bilan oʻlchangan)	Raqamli
Gleason Score	Oʻsma hujayralarining tajovuzkorligini baholovchi koʻrsatkich	Kategorik
Clinical Stage (T1–T4)	Saron bosqichi	Kategorik
Diagnosis (Target)	Saron / Sogʻlom holat	Binary

Bu atributlar klinik amaliyotda eng ko'p qo'llaniladigan diagnostik markerlar bo'lib, ular asosida mashinaviy o'rganish modeli prostata saratonini aniqlashga o'rgatiladi.

Ma'lumotlarni dastlabki qayta ishlash (Preprocessing)

Ma'lumotlar real klinik tizimlardan yig'ilgani sababli, ularning tarkibida nomukammalliklar — **bo'sh qiymatlar, chiqishlar (outliers), shovqin, notekis taqsimot** mavjud bo'lishi tabiiydir. Shuning uchun quyidagi preprocessing bosqichlari amalga oshirildi:

1. **Bo'sh qiymatlar** median qiymat bilan to'ldirildi.

2. **Outlier** qiymatlar IQR (Interquartile Range) usuli orqali filtrlash yoki Winsorizing bilan silliqlandi.

3. **Normalizatsiya** raqamli belgilar uchun Min-Max scaler orqali bajarildi:

4. **Kategorik belgilarga** One-Hot Encoding qo'llanildi.

$$\hat{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Bunday yondashuv belgilar o'rtasidagi sun'iy og'ishlarni kamaytiradi va barcha atributlarning modelga teng ta'sirini ta'minlaydi.

Feature Engineering jarayoni

Feature Engineering — ushbu maqolaning asosiy metodologik yadrosi bo'lib, u uchta asosiy yo'nalishda amalga oshirildi:

Yondashuv	Maqsad
Feature Selection (Tanlash)	Diagnostik ahamiyati yuqori bo'lgan belgilarni aniqlash
Feature Extraction (Alohida xususiyat ajratish)	Murakkab bog'liqliklarni kompakt o'lchamda ifodalash
Feature Interpretation (Izohlanish)	Model qarorlarini klinik talqin qilish

Qo'llangan usullar:

• **ANOVA (Analysis of Variance):**

Belgilarning target bilan bo'lgan farqlanish darajasini statistik baholadi.

• **Chi-square testi:**

Kategorik atributlar diagnostik natijaga ta'sirini o'lchashda qo'llandi.

• **RFE (Recursive Feature Elimination):**

Eng kam ahamiyatli belgilarni iterativ ravishda chiqarib tashlab, optimal feature to'plamini aniqladi.

• **LASSO (L1-regularizatsiya):**

Keraksiz belgilarni penalizatsiya orqali nol ko'effitsiyentga keltirdi.

• **PCA (Principal Component Analysis):**

O'lchamni qisqartirish uchun ishlatildi. Variansga asoslangan quyidagi matematik model bo'yicha ishlaydi:

$Z = XW$ bu yerda

(X) — normalizatsiya qilingan ma'lumotlar matritsasi,

(W) — kovariatsiya matritsasi asosida olingan egenvektorlar.

Model	Tavsif
Logistic Regression	Bazaviy chiziqli model
SVM (RBF kernel)	Chiziqli bo'lmagan bo'linishni aniqlash
Random Forest	Ansambl qaror daraxtlari
XGBoost	Keng tarqalgan gradient boosting modeli

Har bir model 80/20 nisbatda **train-test split** orqali baholandi. Model optimallashtirish uchun **Grid Search** yordamida gipertunerlash amalga oshirildi.

Baholash mezonlari (Metrics)

Natijalar. Ushbu bo'limda turli Feature Engineering usullaridan so'ng olingan diagnostik natijalar, korrelyatsiya tahlillari, o'lchamni qisqartirish jarayonlari, belgilarning ahamiyati va mashinaviy o'rganish modellarining yakuniy samaradorligi yoritiladi. Natijalar ilmiy jihatdan mustaqil talqin qilingan holda, tibbiy va matematik nuqtai nazardan izohlanadi [9].

Korrelyatsiya tahlili (Correlation Analysis)

Dastlab barcha raqamli belgilarning Diagnosis (Target) bilan o'zaro bog'liqligi Pearson ko'effitsienti yordamida o'lchandi. Natijalar shuni ko'rsatdiki, **PSA, Gleason Score** va **Age** diagnostik natijaga eng kuchli ta'sir ko'rsatadigan klinik atributlar sifatida ajralib chiqdi. Prostata hajmi va klinik bosqich esa nisbatan pastroq bog'liqlikka ega bo'ldi, biroq ular model uchun to'liq ahamiyatsiz deb topilmadi.

PCA o'lchamni qisqartirish jarayonida ma'lumotlarning **91% umumiy dispersiya hissasi** birinchi ikkita asosiy komponent orqali ifodalanganini ko'rsatdi. Bu shuni anglatadiki, klinik belgilar ichida yashirin struktura mavjud va u diagnostik bashorat uchun yaxshi ajraluvchi fazoni ta'minlaydi.

O'rin	Feature	O'rtacha ahamiyat (kombinatsiyalashgan usullar bo'yicha)
1	PSA	Eng yuqori
2	Gleason Score	Yuqori
3	Age	O'rtacha yuqori
4	Prostate Volume	Pastroq
5	Clinical Stage	Eng past

Yakuniy interpretatsiya esa **SHAP** orqali amalga oshirildi.

Figure 3. [SHAP summary plot shu yerda joylashadi]

Izoh: PSA ko'rsatkichining oshishi saraton ehtimolini eng ko'p oshiruvchi omil sifatida qayd

etildi. SHAP natijasi — model qarorlarini klinik talqin qilish imkonini bergani uchun shifokorlar uchun qulay izohlash vositasidir [8].

Modellar samaradorligini taqqoslash

To'rtta mashinaviy o'rganish modeli o'qitildi va test to'plamida baholandi:

Model	Accuracy	F1-score	ROC AUC
Logistic Regression	0.84	0.82	0.86
SVM (RBF)	0.86	0.85	0.88
Random Forest	0.89	0.88	0.91
XGBoost	0.92	0.91	0.94

Natija shuni ko'rsatadiki, **XGBoost modeli eng yuqori diagnostik aniqlikka erishdi**, ayniqsa ROC AUC qiymatining yuqoriligi modelning sog'lom va kasallangan bemorlarni yaxshi ajrata olishini isbotlaydi.

Confusion Matrix tahlili

Figure 4. [Confusion Matrix shu yerda joylashadi]

Ushbu maqola prostata saratonini erta aniqlashda klinik ko'rsatkichlarga asoslangan mashinaviy o'rganish yondashuvlarining samaradorligini keng qamrovda o'rgandi va Feature Engineering jarayonining model natijalariga ta'siri ilmiy hamda amaliy jihatdan chuqur tahlil qilindi. Natijalar shuni ko'rsatdiki, PSA darajasi, Gleason balli va bemor yoshi prostata saratoni ehtimolini prognoz qilishda eng muhim klinik belgilar hisoblanadi. Bu esa klinik ko'rsatkichlar to'plamining diagnostik qiymati yuqori ekanini tasdiqlaydi. Maqolada qo'llangan ANOVA, Chi-square, LASSO, RFE va PCA kabi Feature Engineering usullari natijasida keraksiz va diagnostik ahamiyati past bo'lgan belgilarni samarali tarzda chiqarib tashlash, ma'lumotlar fazosini ixchamlashtirish va eng informativ xususiyatlarni ajratib olishga erishildi. Natijada, mashinaviy

o'rganish modellarining samaradorligi ortdi va model barqarorligi yaxshilandi.

Taqqosiy tahlil natijalariga ko'ra, ansamblga asoslangan **XGBoost** modeli eng yuqori natijalarni ko'rsatdi — **AUC = 0.94**, bu esa Logistic Regression, SVM va Random Forest kabi klassik modellar ko'rsatkichlaridan ustun bo'ldi. Ushbu natijalar Qu va hamkorlarining (2023) [2] maqolalariga hamohang bo'lib, ular ham gradient boosting yondashuvlari klinik ma'lumotlar kabi murakkab va korrelyatsiyalashgan datasetlar uchun eng samarali ekanini ta'kidlashgan. Shuningdek, ushbu maqola shuni ko'rsatadiki, diagnostika modelining samaradorligini oshirishda faqat algoritmni emas, balki **xususiyat muhandisligi (feature engineering)** bosqichini takomillashtirish ham hal qiluvchi ahamiyatga ega. Bu xulosa Chen va hamkorlari (2023) [3] tomonidan PSA asosidagi modellarda qayd etilgan natijalar bilan mos keladi.

Xulosa. Klinik ko'rsatkichlar asosida saraton tashxislash uchun kuchli diagnostik model yaratish mumkinligi isbotlandi;

- Feature Engineering — diagnostika natijalarini sezilarli darajada yaxshilovchi hal qiluvchi bosqich ekanligi tasdiqlandi;

- SHAP kabi izohlanuvchan AI yondashuvlari klinik amaliyot uchun zarur ekanligi ko'rsatildi;

- XGBoost modeli hozircha klinik ko'rsatkichlar asosidagi eng istiqbolli klassifikator sifatida tavsiya etildi [6].

Umuman olganda, ushbu maqola klinik diagnostikada sun'iy intellektdan foydalanish sifati va ishonchliligini oshirish bo'yicha muhim metodologik asos yaratadi. Kelgusida, bunday yondashuvlarni yanada kengroq datasetlar, tasviriy ma'lumotlar va genomik profiling bilan birlashtirish orqali diagnostika aniqligini yanada oshirish mumkin. Klinik qaror qo'llab-quvvatlash tizimlari uchun esa bu maqola amaliy platforma vazifasini o'tay oladi.

Adabiyotlar

1. Wong M.C.S., Goggins W., Wang H.H. Global epidemiological trends of prostate cancer. // Cancer. 2022. Vol. 128(10). – P. 2104-2118.
2. Qu Y., Lu X., Yin P. Prostate cancer detection via explainable machine learning. // Computers in Biology and Medicine. 2023. Vol. 165. – P. 107-121.
3. Chen R., Liu Y., Qi J. PSA-based early diagnosis of prostate cancer using ML models. // BMC Urology. 2023. Vol. 23(18). – P. 1-10.
4. Алексеев Б.Я., Каприн А.Д. Рак предстательной железы: диагностика и лечение. – Москва: «Медицина», 2021.
5. Гладышев М.Ю., Лопатин В.Н. Применение машинного обучения в онкоурологии. // Онкоурология. 2022. Т. 18(3). – С. 41-48.
6. Трухин Д.А. Анализ факторов риска рака простаты на основе PSA и клинических данных. // Вестник Урологии. 2023. №2. – С. 55-62.