Модели на машинно обучение за диагностичен и прогнозен анализи за казуса на рака на гърдата при интелигентната интегрирана дигитална платформа InSilcoKDD

Илиян Кордев, Пламенка Боровска

Резюме: *В статията са представени моделите на машинно обучение за диагностичен и прогнозен анализ на рака на гърдата, основани на анализ на големи биомедицински данни. Посредством тези модели са верифицирани и валидирани интелигентният метод за адаптивно откриване на in silico знания, както и интелигентната интегрирана дигитална платформа InSilicoKDD. Експерименталните модели са имплементирани на езика за програмиране Python във професионалната интегрирана среда за развитие PyCharm на JetBrains и е използвана библиотеката за машинно обучение MLlib на работната рамка Apache Spark. Наборите данни за обучение и валидиране, както и за оперативен анализ съдържат клинични изследвания от биопсията на клетките в тумора на реални пациенти. Анализът на експерименталните резултати, получени на дигиталната платформа InSilicoKDD, показват най-добри резултати за акуратност за метода на случайна гора на решенията (93.42%) при диагностичния анализ и за метода на логистичната регресия (79.35%) при прогнозния анализ за рецидив на рака на гърдата.*

Ключови думи: *in silco технологии, анализ на големи биомедицински данни, дигитални платформи, биоинформатика, машинно обучение*

Machine Learning Models for Diagnostic and Predictive Analytics for the Breast Cancer Issue on the Intelligent Integrated Platform InSilcoKDD

Ilijan Kordev, Plamenka Borovska

Abstract*: The article presents the machine learning models for diagnostic and prognostic analysis of breast cancer, based on Big biomedical data analytics. Using these models, the intelligent method for adaptive in silico knowledge data discovery as well as the intelligent integrated digital platform InSilicoKDD have been verified and validated. The experimental models were implemented in the Python programming language within JetBrains professional PyCharm integrated development environment, and the MLlib machine learning library of the Apache Spark framework was utilized. The training and validation data sets, as well as the operational analysis, comprise clinical data from cell biopsy in the tumor of real patients. The analysis of the experimental results conducted on the digital platform InSilicoKDD showed the best results for accuracy for the method of random forest (93.42%) in the diagnostic analysis and for the method of logistic regression (79.35%) in the prognostic analysis for breast cancer recurrence.*

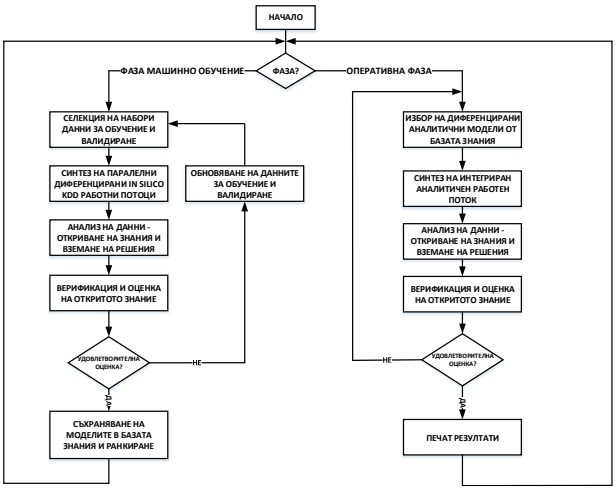
Key words: *in silco technologies, Big biomedical data analytics, digital platforms, bioinformatics, machine learning*

1. АКТУАЛНОСТ И ЗНАЧИМОСТ НА НАУЧНАТА ОБЛАСТ

Анализът на големи биомедицински данни е най-новата и перспективна дигитална технология в подкрепа на прецизната медицина. Информацията, която медиците трябва да обработят и анализират с цел персонализирана диагностика и терапия на пациентите, е огромна, и изисква неимоверни усилия и неприемливо голямо време. На основата на предложения интелигентен метод за адаптивно откриване на in silico знания и

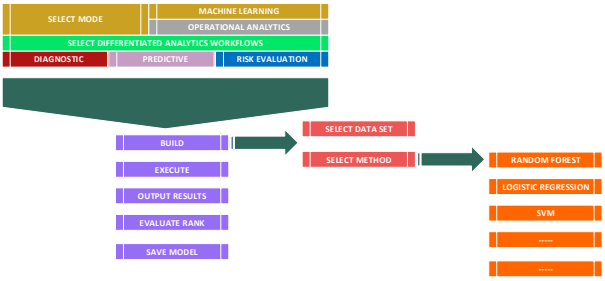
интелигентно вземане на решение посредством анализ на големи данни [1] е проектирана и имплементирана интелигентна интегрирана дигитална платформа InSilicoKDD [2].

Платформата InSilicoKDD е изградена на основата на фазово – паралелната парадигма и функционира в два режима: (1) машинно обучение, и (2) оперативен анализ. Във фазата на машинно обучение се създават модели за конструиране на диференцирани аналитични потоци при различна дълбочина на анализа и откритите знания – описателен, диагностичен, прогнозен, и предписателен (проактивен) анализи. Създадените модели се съхраняват в хранилище и се ранкират в съответствие с оценките за прецизност, акуратност и сензитивност. Във оперативната фаза се конструират и изпълняват интегрирани аналитични потоци посредством селекция и интегриране на диференцираните работни потоци (модели). Функционалността на интелигентната интегрирана дигитална платформа InSilicoKDD, както и йерархията на менютата в графичния потребителски интерфейс, са представени на фиг. 1 и фиг. 2, съответно.

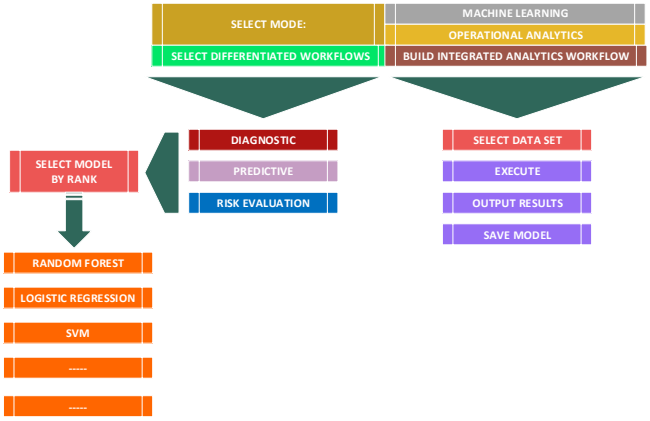


Фиг. 1. Фазово-паралелната парадигма при метода за адаптивно откриване на in silico знания и вземане на решения

Целта на настоящата статия е да представи синтеза на моделите за машинно обучение за анализ на големи биомедицински данни, както и експерименталните резултати за тяхната ефективност на интелигентната интегрирана дигитална платформа InSilicoKDD за казуса на рака на гърдата.



Фиг. 2а Йерархия на опциите в менюто на режима на машинно обучение



Фиг. 2б Йерархия на опциите в менюто на режима на оперативен анализ

2. МЕТРИКА ЗА ОЦЕНКА НА КАЧЕСТВАТА НА МОДЕЛИТЕ ЗА МАШИННО ОБУЧЕНИЕ.

Използват се следните оценъчни параметри за примера за диагноза на рак на гърдата:

- *True positive* (*TP*): диагнозата е положителна и пациентът има рак (тази оценка е желаната);
- *True negative* (*TN*): диагнозата е негативна и пациентът няма рак (тази оценка е желаната);
- *False positive* (*FP*) – фалшива тревога (*false alarm)*: диагнозата е положителна, а пациентът няма рак (тази оценка не е желаната);
- *False negative* (*FN*): диагнозата е негативна, а пациентът има рак (тази оценка е най-нежеланата);

Акуратност (Accuracy) **–** представлява отношението на правилно поставените диагнози (попадения) от модела към общият брой поставени диагнози в процеса на валидирането на модела. Оценката за акуратност е най-интуитивната.

Оценката за акуратност се прилага предимно за симетрични набори данни - броят на фалшивите позитивни диагнози FP и броят на фалшивите негативни диагнози FN поставени от модела са

близки, а също така и когато стойностите на FN и FP са близки.

Акуратността се изчислява по следната формула:

$$Accuracy = (TP+TN)/(TP+FP+FN+TN) \quad (1)$$

Прецизност (Precision) представлява отношението на правилно поставените позитивни диагнози (попадения) от модела към общият брой поставени диагнози (правилни и неправилни). Дава отговор на въпроса: колко от поставените позитивни диагнози в рамките на набора данни за валидиране кореспондират с действителните диагнози. Оценката за прецизност показва до каква степен сме сигурни, че положителните позитивни диагнози, поставени от модела, са правилни т.е. да няма здрави пациенти диагностицирани с рак Прецизността се изчислява по следната формула:

$$Precision = TP/(TP+FP) \quad (2)$$

Recall (Sensitivity) представлява отношението на правилно поставените положителни диагнози (попадения) от модела към сумата от общия брой действителни положителни диагнози и общия брой фалшиви негативни диагнози (недиагностицирани пациенти с рак) в рамките на набора данни за валидиране. Дава отговор на въпроса: *колко от пациентите с рак на гърдата в рамките на набора данни за валидиране са диагностицирани правилно (или са пропуснати)*. Параметърът *Recall* дава оценка за ефективността на модела при поставянето на положителна диагноза на специфицираното заболяване. Той се използва в случаите, когато получаването на фалшиви позитивни диагнози е за предпочитане пред получаването на негативни фалшиви диагнози, т.е. когато високата честота на фалшивите негативни диагнози (пациенти с рак да бъдат диагностицирани като здрави) е неприемлива (предпочита се фалшивите тревоги да са повече пред това да има болни пациенти с недиагностициран рак). Оценката на параметъра за сензитивност показва до каква степен сме сигурни, че моделът не пропуска положителните диагнози (което е изключително важно и недопустимо - болни пациенти да бъдат диагностицирани като здрави).

Параметърът *Recall* се изчислява по следната формула:

$$Recall = TP/(TP+FN) \quad (3)$$

F1-score (F-Score / F-Measure) обхваща оценките за прецизност и сензитивност и отразява техния баланс. Представлява хармонична средната стойност (harmonic mean(average)) на прецизността и сензитивността. Най-добрите стойности на F1 Score се получават при добър баланс между прецизността и сензитивността precision (p) & recall (r) в системата. При небалансирана система стойността на F1 Score не е висока в случай, че стойността на прецизността се подобрява за сметка на стойността на сензитивността или обратно. Например, при P=1 и R=0, стойността на *F1 score* е 0. Оценката на параметъра F1 е препоръчителна при неравномерно разпределение на класовете. F1 score се изчислява по формулата:

$$F1\ Score = 2*(Recall * Precision) / (Recall + Precision) \quad (4)$$

Специфичност (Specificity) представлява отношението на правилно диагностицираните здрави хора от модела към общия брой действително здрави хора във набора данни за валидиране. Дава отговор на въпроса: на колко от здравите хора е поставена правилна диагноза. Оценката на параметъра *Specificity* фокусира върху правилното диагностициране на здравите пациенти, т.е. фалшивите диагнози, позитивни или негативни, поставени от модела, са изключително нежелани. Не е допустимо здрави хора, диагностицирани погрешно с рак на гърдата (фалшива тревога), да бъдат подложени на лечение. Специфичността се изчислява по формулата:

$$Specificity = TN/(TN+FP) \quad (5$$

Интерпретация на примерни оценки на параметрите на модел за машинно обучение:

✓ *Accuracy = 90%*: за 100 пациента моделът е поставил 90 правилни диагнози и 10 неправилни диагнози (*обхваща както позитивните диагнози - болни пациенти, така и негативните диагнози - здрави пациенти*);

✓ *Precision = 80%*: средно на 100 пациента с рак на гърдата (болни), 80 пациента са

диагностицирани правилно от модела за машинно обучение, докато 20 болни пациенти са диагностицирани погрешно като здрави (*обхваща само позитивните диагнози – болни пациенти*).

✓ *Recall = 70%*: само 70 от 100 пациенти, болни от рак, са диагностицирани правилно, останалите 30 болни пациента моделът е пропуснал (не са диагностицирани като болни);

✓ *Specificity = 60%*: всеки 40 от 100 реално здрави пациенти са диагностицирани погрешно като болни от модела и 60 от 100 реално здрави пациенти са определени като здрави.

Оценката на полезност в имплементирания софтуерен модул за интерпретиране, оценка на полезността и визуализация на резултатите се включва и оценката на полезност от верификацията на данни, направена от експерт онколог.

Значимостта на откритите знания се оценява чрез хибриден подход - комбинация от изчислителни критерии (PAS – Precision, Accuracy & Sensitivity) и експертна оценка за стойност от учените от целевата област. За да се намали разликата между адекватността на изчислителните методи и биологичните и медицински аспекти на значимостта на откритите знания, ранкирането на моделите се основава на хибриден подход. Предлага се оценъчен критерий на базата на комбинация от претеглени стойности: PAS (прецизност, точност, сензитивност) и количествената оценка на експерта за значимостта на откритото познание. Двете стойности са мащабирани до нива от 1 до 10. Съотношението на тегловните коефициенти не са фиксирани и потребителят може да избира различни съотношения.

3. СИНТЕЗ НА МОДЕЛИ ЗА МАШИННО ОБУЧЕНИЕ ЗА ДИАГНОСТИЧЕН АНАЛИЗ.

За синтеза на експерименталните модели за машинно обучение са използвани:

✓ езика за програмиране *Python 3.6,* който е интерпретируем, интерактивен, обектно-ориентиран и се използва широко в научните изследвания, поради голямото разнообразие от достъпните библиотеки и платформи (NumPy, SciPy, Matplotlib, Spark, Tensorflow и др.);

✓ професионалната интегрирана среда за развитие *PyCharm* на JetBrains, осигуряваща фамилия от интегрирани среди за езиците Java, Ruby, Python, PHP, C, C++, C#, Go, JavaScript и SQL;

✓ библиотеката за машинно обучение MLlib на работната рамка *Apache Spark.*

За диагностиката на рака на гърдата в платформата InSilicoKDD са заложени 4 опции, базирани на три основни метода за класифициране на тумора (злокачествен или доброкачествен):

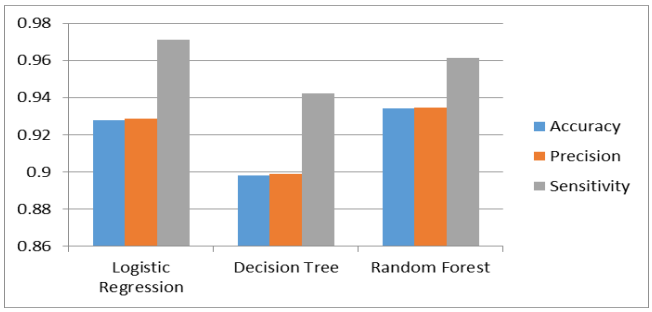
✓ *Опция 1: Дърво на решенията (Decision Tree)* – метод с контролирано обучение (supervised learning algorithms) и може да се използва за решаването на проблеми с регресионен или класификационен характер.

✓ *Опция 2: Произволна гора на решенията (Random Forest)* - обучават се много дървета на решенията, които впоследствие се комбинират;

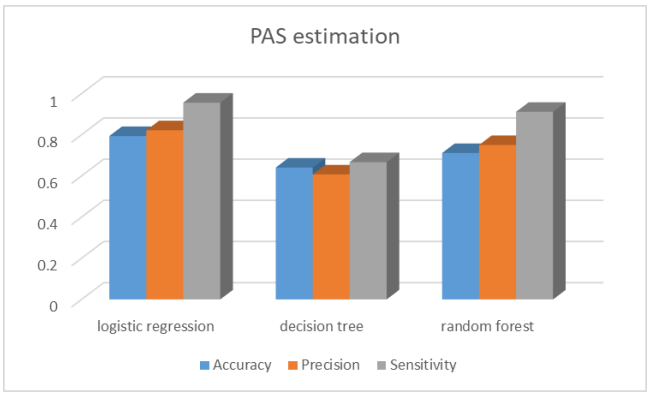
✓ *Опция 3: Логистична регресия (Logistic Regression)* - един от най - популярните методи за анализ на данни с котролирано обучение за бинарна класификация (болен/ здрав, доброкачествен/ злокачествен тумор);

✓ *Опция 4: Система от методи* – едновременно машинно обучение с избрани или всички методи.

За машинното обучение за диагностика на рака на гърдата се използва *набора данни Breast Cancer Wisconsin (Diagnostic) Data Set* (WDBC) на Университета на Уисконсин, САЩ [3]. Съдържа резултатите от биопсията на клетките в тумора на пациента. Характеристиките на данните са радиус, текстура, обиколка, площ, плавност, компактност, вдлъбнатост, вдлъбнати точки, симетрия, фрактално измерение. Данните за осреднени за всичките клетки, като за всяко свойство е дадена осредената, минималната и максималната стойности. Наборът данни обхваща 30 характеристики за пациент, като за всеки един пациент е отразена диагнозата на тумора – злокачествен (*malignant*) или доброкачествен (*benign*). При експериментирането е променено обозначението на диагнозата от *M* на *1* за злокачествен тумор, и от *B* на *0* за доброкачествен тумор. Наборът от данни WDBC обхваща данни за 569 пациента, като 357 са с доброкачествени тумори, а 212 са със злокачествени ракови тумори на гърдата.



Фиг.3. Експериментални резултати от машинното обучение със система от модели на дигиталната платформа InSilicoKDD за диагностика на рака на гърдата



Фиг.4 Експериментални резултати от диференцирания работен поток за прогнозен анализ със система от модели

4. СИНТЕЗ НА МОДЕЛИ ЗА МАШИННО ОБУЧЕНИЕ ЗА ПРОГНОЗЕН АНАЛИЗ

За прогнозирането на рецидив на рака на гърдата в платформата InSilicoKDD са заложени 4 опции, базирани на три основни метода за класифициране на *тумора (злокачествен или доброкачествен):*

✓ *Опция 1: Дърво на решенията (Decision Tree)*;

✓ *Опция 2: Произволна гора на решенията (Random Forest);*

✓ *Опция 3: Логистична регресия (Logistic Regression)*;

✓ *Опция 4: Система от методи* – едновременно машинно обучение с избрани или всички методи.

Експерименталните резултати от машинното обучение за прогноза за рецидив на рака на гърдата са представени на фиг. 4.

Анализът на експерименталните резултати показва най-добри резултати по отношение на акуратност, прецизност и сензитивност за метода на логистичната регресия.

5. СИНТЕЗ НА ИНТЕГРИРАН АНАЛИТИЧЕН ПОТОК ЗА ОПЕРАТИВЕН АНАЛИЗ.

Оперативен диагностичен анализ. Клиничните данни на пациента за провеждането на експеримента във фазата на оперативния диагностичен анализ са представени в табл.1.

За експерименталните пациентски данни експерименталните резултати от диагностичната фаза в оперативен режим са следните:

Табл. 1. Клинични данни на пациента за провеждането на експеримента във фазата на оперативния диагностичен анализ

Радиус (radius_mean)	Текстура (texture_mean)	Обиколка (perimeter_mean)	Площ (area_mean)	Гладкост (smoothness_mean)
15.22	30.62	103.4	716.9	0.1048
Компактност (compactness_mean)	Вдлъбнатост (concavity_mean)	Точки на вдлъбнатост (concave_points_mean)	Симетрия (symmetry_mean)	Размери на фрактала (fractal_dimension_mean)
0.2087	0.255	0.09429	0.2128	0.07152

1. Дърво на решенията – диагнозата е злокачествен рак на гърдата, което е вярно;
2. Произволни гора на решенията - диагнозата е злокачествен рак на гърдата, което е вярно;
3. Логистична регресия - диагнозата е злокачествен рак на гърдата, което е вярно.

Оперативен прогнозен анализ. Клиничните данни на пациента за провеждането на експеримента във фазата на оперативния прогнозен анализ са представени в табл.2.

Табл.2 Клинични данни на пациента, подложени на оперативен прогнозен анализ.

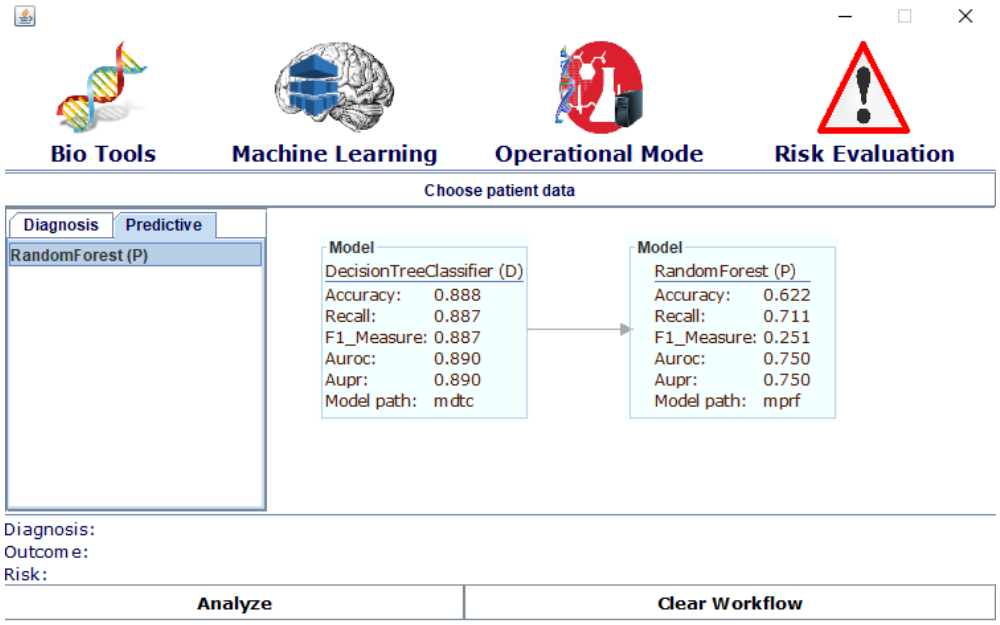
Радиус (radius_mean)	Текстура (texture_mean)	Обиколка (perimeter_mean)
14.72	25.26	99.28
Вдлъбнатост (concavity_mean)	Точки на вдлъбнатост (concave_points_mean)	Симетрия (symmetry_mean)
0.1729	0.09465	0.2079
Площ (area_mean)	Гладкост (smoothness_mean)	Компактност (compactness_mean)
657.5	0.1174	0.2112
Размери на фрактала (fractal_dimension_mean)	Размер на тумора (tumorsize)	Състояние на линфните възли (lymphnodestatus)
0.07496	1.7	21

За експерименталните пациентски данни експерименталните резултати от прогнозната фаза в оперативен режим са следните:

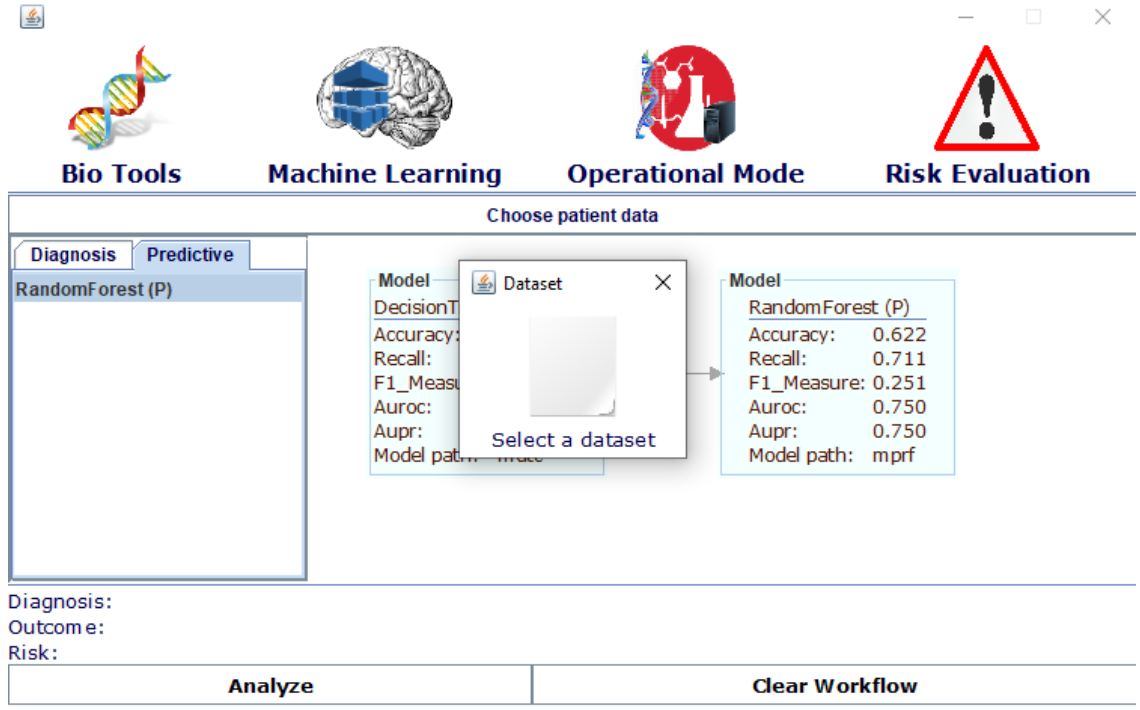
1. Дърво на решенията – прогнозата е рецидив на рак на гърдата, което е грешно;
2. Произволни гора на решенията - прогнозата е, че не се очаква рецидив на рак на гърдата, което е вярно;
3. Логистична регресия - прогнозата е, че не се очаква рецидив на рак на гърдата, което е вярно.

5. ВИЗУАЛИЗАЦИЯ НА ЕКСПЕРИМЕНТАЛНИТЕ РЕЗУЛТАТИ НА ПЛАТФОРМАТА *INSILICOKDD*.

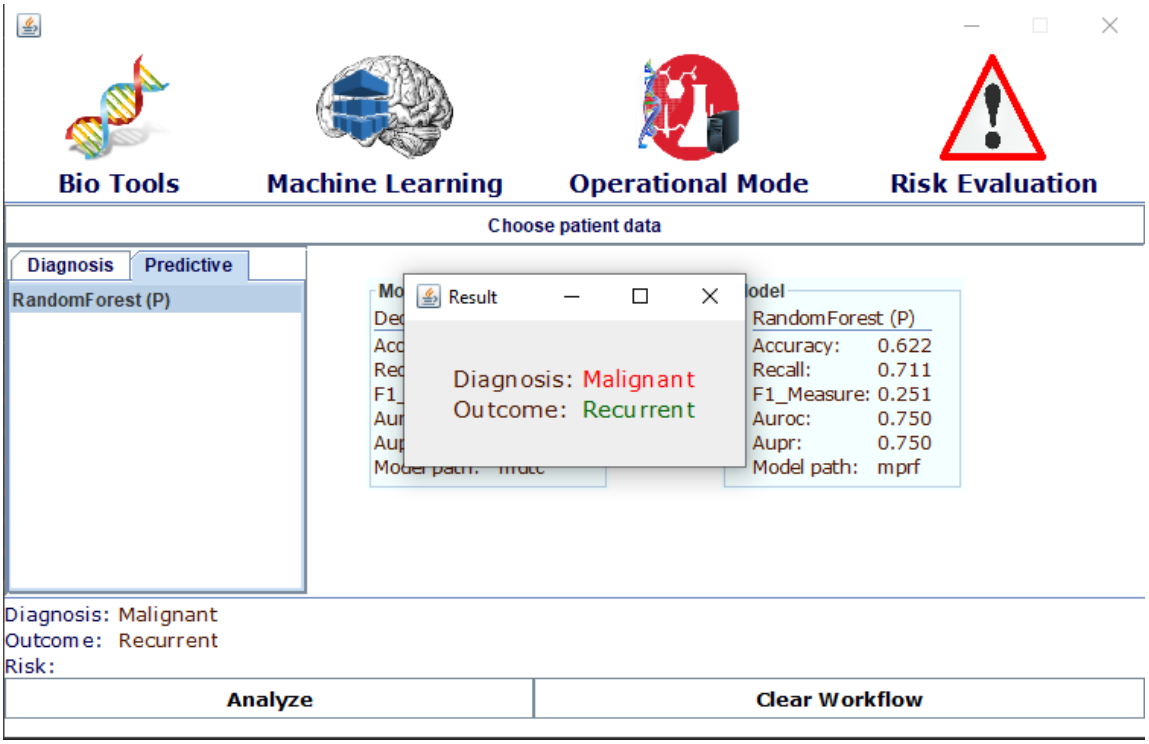
В рамките на интелигентната дигитална платформа InSilicoKDD е интегриран софтуерен инструмент за визуализация като част от потребителския графичен интерфейс. На фиг. 5, фиг. 6, фиг.7 и фиг.8 са представени екранни снимки от визуализацията на експерименталните резултати.



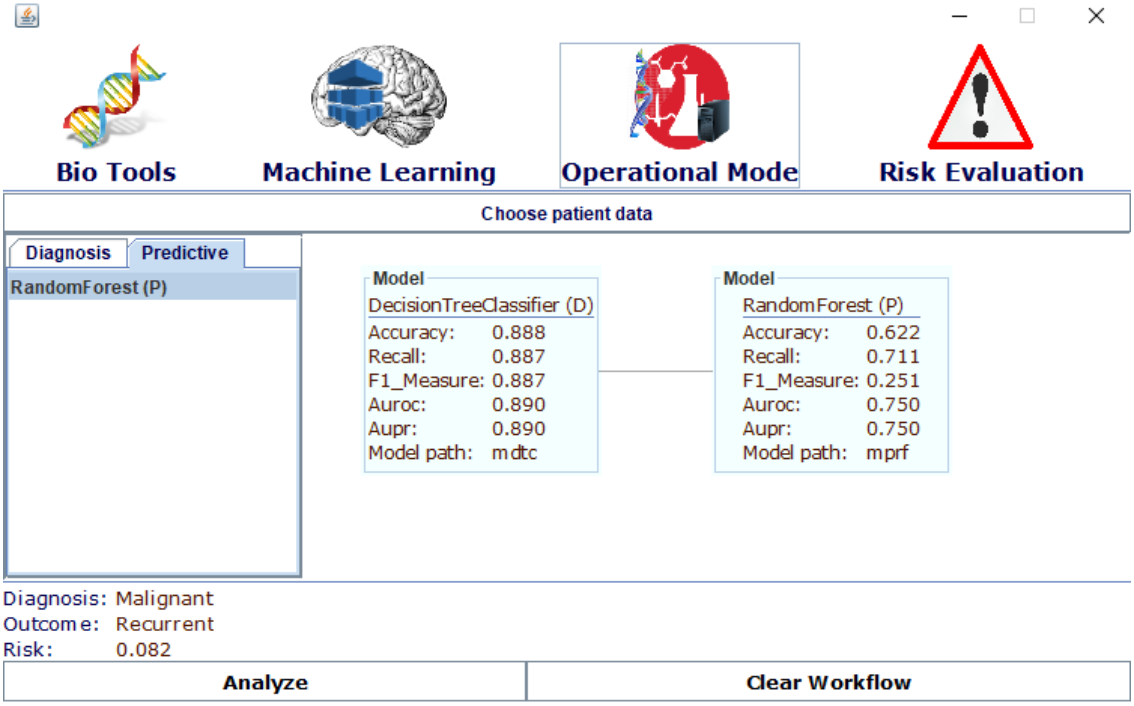
Фиг.5. Екранна снимка на конфигуриране на интегриран аналитичен работен поток обхващащ: (1) диференциран работен поток за оценка на риска (модел на Клаус), (2) диференциран диагностичен работен поток (дърво на решенията), и (3) диференциран прогнозен работен поток (произволна гора на решенията)



Фиг. 6 Екранна снимка на избор на данни на пациента при изпълнение на интегриран работен поток в режим на оперативен анализ на дигиталната платформа InSilicoKDD.



Фиг. 7 Екранна снимка на визуализация на резултатите за диагностика и прогноза на рака на гърдата на пациент след изпълнението на интегриран работен поток в режим на оперативен анализ на дигиталната платформа InSilicoKDD



Фиг. 8 Екранна снимка на визуализация на резултатите за оценка на риска, диагностика и прогноза на рака на гърдата на пациент след изпълнението на интегриран работен поток в режим на оперативен анализ на дигиталната платформа InSilicoKDD.

6. ЗАКЛЮЧЕНИЕ

В статията са представени моделите на машинно обучение за диагностичен и прогнозен анализ на рака на гърдата, основани на анализ на големи биомедицински данни. Посредством тези модели са верифицирани и валидирани интелигентният метод за адаптивно откриване на in silico знания, както и интелигентната интегрирана дигитална платформа InSilicoKDD. Експерименталните модели са имплементирани на езика за програмиране Python във професионалната интегрирана среда за развитие PyCharm на JetBrains и е използвана библиотеката за машинно обучение MLlib на работната рамка Apache Spark. Наборите данни за обучение и валидиране, както и за оперативен анализ съдържат клинични изследвания от биопсията на клетките в тумора на реални пациенти. Анализът на експерименталните резултати, получени на дигиталната платформа InSilicoKDD, показват най-добри резултати за акуратност за метода на случайна гора на решенията (93.42%) при диагностичния анализ и за метода на логистичната регресия (79.35%) при прогнозния анализ за рецидив на рака на гърдата..

БЛАГОДАРНОСТИ

Резултатите от това изследване са финансирани от научно-изследователски проект ДН07/24/2016г. „Интелигентен метод за адаптивно откриване на in silico знания и взимане на решения, базирани на анализ на потоци големи данни за целите на научните изследвания“, ФНИ, МОН.

REFERENCES

[1] P. Borovska, D. Ivanova, “Intelligent method for adaptive in silico knowledge discovery based on big genomic data analytics”, AIP Conference Proceedings vol. 2048, 060001(2018); ISBN: 978-073541774-8, pp. 060001-1 - 060001-9, American Institute of Physics

[2] Plamenka Borovska, Iliyan Kordev, Boris Borowsky (2020), Intelligent Integrated Digital Platform InSilicoKDD in Support of Precision Medicine for Scientific Research. *International Journal of Advanced Science and Technology*, *29*(7s), 2101-2109, ISSN: 2005-4238 (Print),

ISSN: 2207-6360 (Online), Publisher: Science and Engineering Research Support Society, Australia
[3] Ivanova D., Borovska P., Design and implementation of bioinformatics KDD workflow for breast cancer risk evaluation, AIP Conference Proceedings, Volume 2172, 13 November 2019, Article number 020006, 45th International Conference on Application of Mathematics in Engineering and Economics, AMEE 2019; Sozopol; Bulgaria; 7 June 2019 through 13 June 2019; ISBN: 978-073541919-3 ISSN:0094-243XE-ISSN:1551-7616 pp. 020007-1-020007-7
[4] Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on? https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124
[5] William H. Wolberg, Breast Cancer Database, University of Wisconsin Hospitals, Madison, Wisconsin, USA. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)
[6] E. Biganzoli, A. Vellido, F. Ambrogi, R. Tagliaferri, "Computational Intelligence Methods for Bioinformatics and Biostatistics", Springer Science & Business Media, Dec. 2012, e-ISBN 978-3-642-35686-5
[7] PyServer https://pypi.org/project/pyserver/

За авторите:

Илиян Кордев
Студент, Специалност Информатика и софтуерни науки
Факултет по приложна математика и информатика
Технически университет – София

Пламенка Боровска, pborovska@tu-sofia.bg
Професор, доктор на науките
катедра Информатика
Факултет по приложна математика и информатика
Технически университет - София

About the authors:

Ilijan Kordev
Student, Specialty Informatics and Software Sciences
Faculty for Applied Mathematics and Informatics
Technical University of Sofia

Plamenka Borovska, pborovska@tu-sofia.bg
Professor, DSc, Informatics Dept.,
Faculty for Applied Mathematics and Informatics
Technical University of Sofia