https://orcid.org/0000-0002-6928-2080
https://orcid.org/0000-0003-1604-6129
https://orcid.org/0000-0002-1256-6138

УДК 004.4, 004.6, 004.021

О.В. ГУДЗЬ[*], О.В. БУШМА[**], Б.Л. ГОЛУБ[*]

ТЕСТУВАННЯ СИСТЕМИ АНАЛІЗУ ДАНИХ ДЛЯ ТЕХНОЛОГІЇ ПРОГНОЗУВАННЯ ШКІДЛИВИХ РЕЧОВИН

[*]Національний університет біоресурсів та природокористування України. м. Київ, Україна
[**]Київський університет імені Бориса Грінченка, м. Київ, Україна

Анотація. *Питання безпечності продуктів гостро стоїть у багатьох країнах та компаніях. Адже від вживання продуктів, що контаміновані шкідливими речовинами, щороку велика кількість людей отримають різні захворювання. Щоб підвищити безпечність продуктів, необхідні технології, які дають змогу швидко провести тестування продукції, та на основі отриманих результатів застерегти появі шкідливих речовин. Сьогодні у світі активно розвиваються біосенсорні технології для проведення експрес-тестування продуктів сільськогосподарського призначення. Використання технологій інтелектуального аналізу даних може забезпечити виявлення нових знань про характер та залежність від зовнішніх чинників різних шкідливих речовин. У статті розглядаються питання побудови та тестування технології прогнозування, до яких входять аналіз рішень, що вже існують, пошук прогностичних моделей, використання методів Data Mining для знаходження якісно нових знань та гіпотез, тестування гіпотез. У першому розділі статті проводиться огляд існуючих моделей та методів, представлено формули розрахунків ризика появи афлотоксину. У другому розділі проводиться короткий опис системи збору даних, що складається із сховища даних, сенсорів та мобільного додатка як джерела даних. Результати роботи інтелектуального аналізу даних із використанням методів Data Mining представлені у третьому розділі. Як результат, було отримано гіпотезу впливу на появу мікотоксинів у посівах. Підтвердження або спростування гіпотези виконувалось за рахунок її перевірки з використанням OLAP-технології. Результати перевірки гіпотези представлено в останньому, четвертому розділі. Основним результатом роботи є отримання системи аналізу даних, яка готова до апробації в реальних умовах. Проте слід зазначити, що необхідно провести більш розширене тестування на великих об'ємах даних, але це можливо накопичити за декілька років спостережень.*

Ключові слова: *управління інформацією, сховище даних, аналіз даних у режимі реального часу, інтелектуальний аналіз даних, технології OLAP і Data Minin, мікотоксини.*

Abstract. *The issue of product safety is acute in many countries and companies. Indeed, every year a large number of people catch various diseases after using products contaminated with harmful substances. To increase food safety, there should be implemented some technologies which would enable quick testing of products and, using the obtained knowledge, to prevent appearance of harmful substances. Biosensor technologies which are utilized for conducting quick tests for agricultural purposes are being nowadays actively developed in the world. The use of data mining technologies can provide the identification of new knowledge about the nature and dependence of various harmful substances on external factors. The article considers the issues of construction and testing of forecasting technology, including analysis of existing solutions, search for forecasting models, usage of Data Mining methods to find qualitatively new knowledge and hypotheses, and hypothesis testing as well. The first section of the article provides an overview of existing models and methods, formulas for calculating the risk of aflatoxin. The second section contains a brief description of the data collection system which consists of a data warehouse, sensors and a mobile application as a data source. The results of data mining obtained after the use of Data Mining methods are presented in the third section of the paper. As a result, there was obtained a hypothesis of the effect on the appearance of mycotoxins in crops. Confirmation or refutation of the hypothesis was performed by testing it using OLAP-technologies. The results of the hypothesis test are presented in the last section. The main result of the work was obtaining a data analysis system ready for testing in real*

conditions. However, it should be noted that there is a need for more extensive testing on large amounts of data which can be accumulated over several years of observations.
Keywords: *information management; data warehouse; real-time data analysis; data mining; OLAP and Data Mining technologies; mycotoxins.*

DOI: 10.34121/1028-9763-2021-3-113-120

1. Вступ

Аграрне виробництво в Україні складається із двох головних галузей: рослинництво та тваринництво. Також є третя – це кормовиробництво, яка є проміжною, що у великих господарствах має свою структуру, специфіку, організаційно-економічні основи. Близько 93% орних земель в Україні припадає на рослинництво та кормовиробництво, з яких до 30% відведено під кормові культури. Сумарно 40–50% виробництва у рослинництві становить побічна продукція: стебла кукурудзи й сорго, солома хлібів, жом, патока та ін. Ця продукція через проміжну галузь, кормовиробництво, використовується у тваринництві. Рослинництво в Україні все більше набуває рис біологічного виробництва, тобто такого, що ґрунтується на широкому використанні альтернативних, біологічних і пов'язаних з ними агротехнічних методів вирощування сільськогосподарських культур із мінімальним застосуванням засобів хімізації в системі захисту рослин та з максимальним біологічних джерел живлення рослин [1].

У свою чергу, у рослин серйозні порушення фізіологічних процесів обумовлюють гриби, бактерії та віруси, що поселяються на їхній поверхні, у тканинах або у клітинах. Таких істот називають патогенними організмами. Зазвичай вони викликають захворювання рослини-хазяїна та навіть її загибель. Найбільш поширеними патогенами є гриби. Вже виявлено понад 10 000 видів паразитичних грибів. Тільки в сільськогосподарських культурах вони знижують урожай на 20%. На відміну від цього, відомо тільки близько 200 видів бактерій, які вражають рослини.

Паразитичні гриби та мікроорганізми поділяють на дві групи [2]:

• поліфаги, які паразитують на різних видах рослин;

• монофаги, що здатні вражати рослини чітко одного виду.

Як результат, у процесі вирощування будь-яких рослин існує дуже багато факторів, які необхідно відслідковувати та аналізувати. Основним результатом впровадження та роботи цієї системи є збільшення врожайності та безпечності продуктів.

Метою статті є представлення системи аналізу даних, отриманих шляхом експрес-тестування сільськогосподарської продукції під час її вирощування.

2. Моделі, що існують, та методи прогнозування

У результаті пошуку наявних рішень було виявлено, що на сьогодні існують підходи, на зразок [3], які враховують умови перед цвітінням, що дозволяє передбачити рівень зараження колоса спорами грибів. Однак постійний моніторинг контамінації зерна та погодних умов у процесі вирощування сівозміни дасть змогу прогнозувати як інтенсивність спороношення пліснявих, так і потенційний рівень забруднення.

Підхід до прогнозування та розповсюдження мікотоксинів полягає у розробці алгоритму (моделі прогнозування), що аналізуватиме зібрані поточні дані щодо рівня зараження, координат точок збору зразків та стадії вегетації рослин, поточних погодних умов та метеорологічних прогнозів відповідно до відомих закономірностей розповсюдження пліснявих грибків. Було знайдено прогностичну модель [4], що передбачає використання у розрахунку, окрім параметрів температури та вологості, параметр періоду посухи.

Модель прогнозування появи афлатоксину була апробована в Кенії. В Африці на південь від Сахари існує велика проблема з афлатоксином, який виробляється грибком Aspergillus flavus. У Кенії з 1980 року померло понад 500 людей від афлатоксикозу. Саме

цим обґрунтовується потреба в подібній прогностичній моделі, яка маркувала б на карті ризик появи афлатоксину, а також відсутність недорогих і легких у використанні у полі тест-пристроїв.

Розглянута модель базується на Agriculcultural Production Systems sIMulator (APSIM) modelling framework. APSIM моделює зростання кукурудзи, фенологію, врожайність і водний баланс ґрунту, використовуючи щоденні дані про максимальну та мінімальну температури, радіацію та опади. APSIM змоделює водне співвідношення попиту та пропозиції (SDR, без одиниці вимірювання) як індикатор посухи. Визначається, що ключовими параметрами фактора ризику появи афлатоксину є «hot and dry condition» (спека та сухість). Також розглядається взаємодія температури та води із грибком. Є інші моделі (Battiliani 2013), представлені в [5], що використовують як параметри спороношення, інфекційність, зростання грибків і продукцію AFLA при різниці температур і активності води. Але вона не враховує вологість ґрунту та період посухи в моделі. Модель Chauhan 2008 [6] ґрунтується на температурі та періоді посухи.

Визначення фактора температури у моделі зображено на рис. 1.

$$\text{Aflo\_temp\_factor} = \frac{T_{\text{mean\_aflo}} - T_{\text{min\_aflo}}}{T_{\text{opt\_aflo}} - T_{\text{min\_aflo}}};$$

and when $T_{\text{mean\_aflo}} > T_{\text{opt\_aflo}}$ and $< T_{\text{max\_aflo}}$ then

$$\text{Aflo\_temp\_factor} = \frac{T_{\text{max\_aflo}} - T_{\text{mean\_aflo}}}{T_{\text{max\_aflo}} - T_{\text{opt\_aflo}}};$$

and when $T_{\text{mean\_aflo}} < T_{\text{min\_aflo}}$ or $> T_{\text{max\_aflo}}$ then

$$\text{Aflo\_temp\_factor} = 0.$$

Рисунок 1 – Визначення фактора температури

Максимальна температура складає 42,5 градуса, оптимальна – 32,5, мінімальна 11,5. Визначення SDR у моделі не більше 0,2. Стадія розвитку культури дорівнює 8 за визначенням APSIM. Розрахунок індексу ризику появи афлатоксину показано на рис. 2.

When the SDR ≤ 0.20 and maize growth stage ≥ 8

$$\sum \text{Aflo\_risk} = \text{Aflo\_risk} + (1 \times \text{Aflo\_temp\_factor})$$

$$\text{ARI} = \sum \text{Aflo\_risk} \times 10$$

Рисунок 2 – Розрахунок індексу ризику появи афлотоксину

Було оцінено здатність цієї моделі у використанні наборів даних, що генерували багаторазові випробування в Кенії. Змодельований лінійний зв’язок ARI із спостережуваним середнім рівнем забруднення афлатоксинами, який коливався від <1 до 7143 ppb у п’яти різноманітних середовищах, був значним і пояснював високий ступінь варіації середнього рівня забруднення. Модель очікує подальшого застосування як інструмента досліджень та підтримки прийняття рішень для мінімізації забруднення афлатоксинами у цій важливій основній харчовій культурі багатьох країн, що розвиваються [4].

Пошук аналогічних систем на ринку не дав позитивних результатів, проте існують компанії, які пропонують розробку подібних систем, використовуючи методи «Machine learning».

3. Опис системи збору даних

Першим етапом розробки системи було створення загальної архітектури. Результати роботи представлені у статті [7]. В рамках цієї роботи було розроблено сховище даних та описано розгортання системи. Додатково був розроблений оптичний сенсор, що визначає концентрацію мікотоксинів флуоресцентним методом. Результати роботи опубліковані у статті [8]. Також було розроблено мобільний додаток на платформі «Android», що слугує одним із інструментів для збору даних про стан посівів. Слід також відзначити створення web-додатка, що використовується як місце відображення результатів роботи системи.

Наразі відбуваються впровадження та апробація прогностичної моделі на тестових наборах даних. Паралельно відбувається збір даних у режимі реального часу.

4. Результати роботи інтелектуального аналізу даних Data Mining

Використовуючи метод пошуку асоціативних правил, було визначено взаємозв'язки між кількістю мікотоксинів і температурою та вологістю. Дані, які використовуються, є тестовими наборами. Вони зберігаються у багатовимірному вигляді – гіперкубі. Виконуючи пошук асоціативних правил, необхідно запустити процес обробки гіперкуба у середовищі BI MS SQL Server.

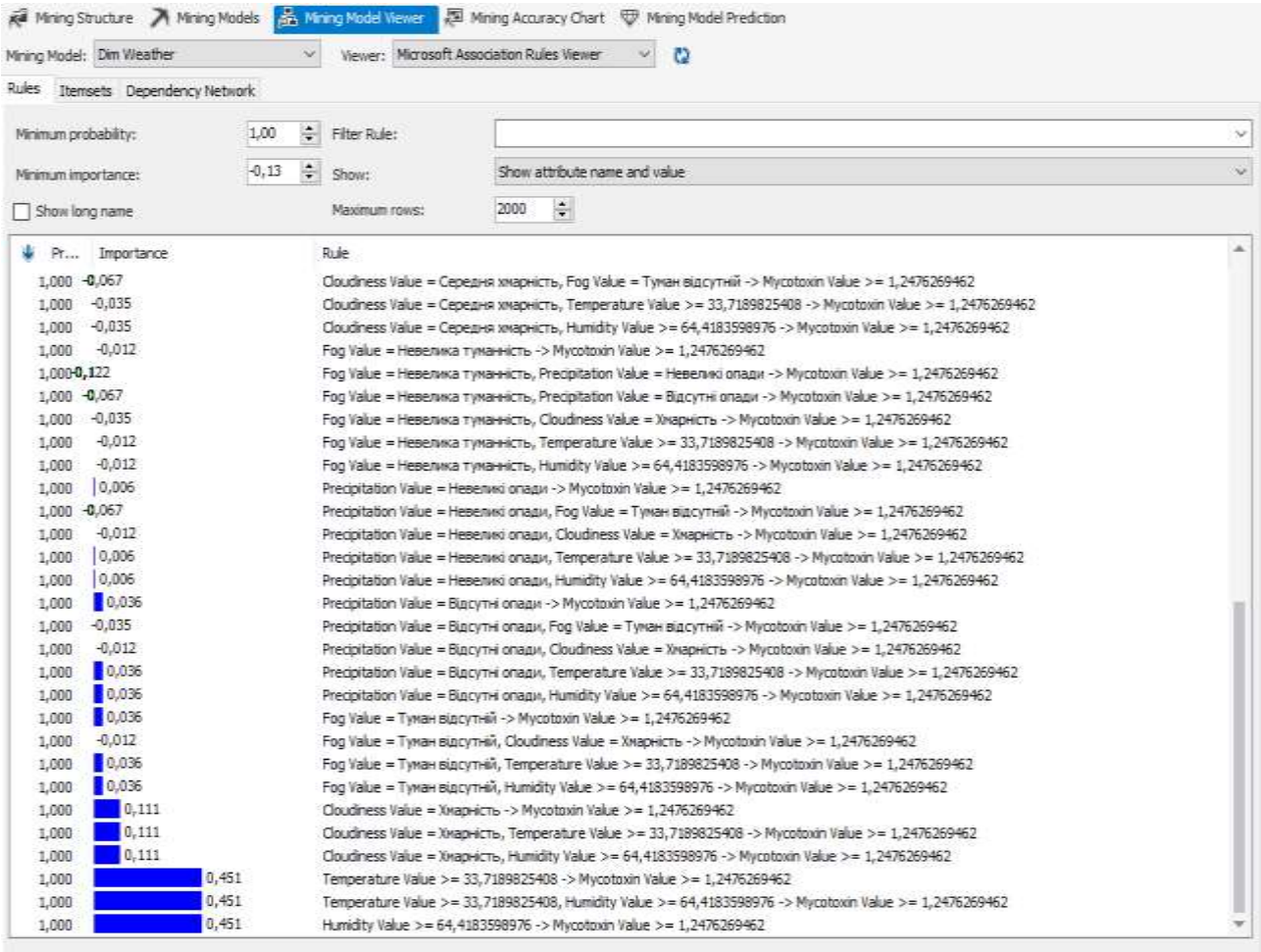

Рисунок 3 – Асоціативні правила

У результаті отримуємо:

• асоціативні правила з визначеною ймовірністю настання та важливістю;

• набори входжень;

• підтримку наборів.

Знайдені правила зображено на рис. 3, набори входжень та їхня підтримка – на рис. 4.

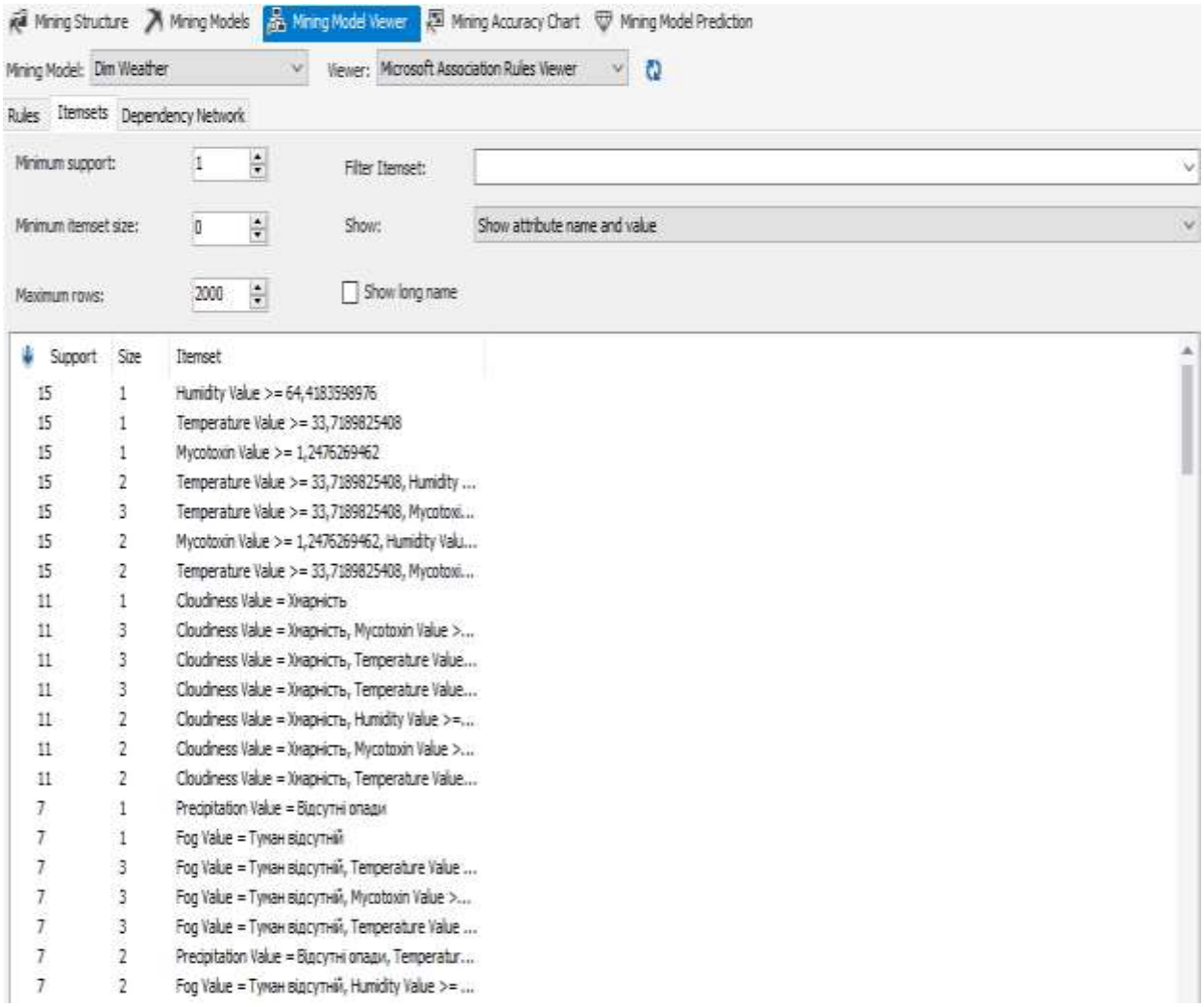

Рисунок 4 – Набори входжень

Аналізуючи результати, бачимо, що з найбільшою важливістю та ймовірністю настання майже 100% правило таке: при вологості повітря понад 64,4% концентрація мікотоксинів більша за 1,24мг/кг. При цьому підтримка наборів, які входять до цього правила, одна з найвищих і становить 15 (з 2000 записів). Враховуючи отриманий результат, за гіпотезу можна взяти це правило.

5. Результати аналізу даних на основі OLAP-технологій

Одним із основних засобів проведення OLAP-аналізу є розрахунок показника KPI. Цей показних показує коефіцієнт досягнення певної мети, поставленої аналітиком. Також він дає можливість визначити тренд на майбутнє.

Аналітичну цікавість викликає визначення впливу погодних умов, а саме вологості, на поширення мікотоксинів. Тому як параметри розрахунку будуть використовуватися се-

реднє арифметичне значення концентрації мікотоксинів у продукції та середнє значення вологості повітря. Параметри братимуться за певний період часу.

При розрахунку система використовує декілька виразів:

• вираз значення – для відображення поточного значення;

• вираз цілі – задає значення, якому має відповідати поточне значення в ідеальних умовах;

• вираз статусу – визначає статус поточного значення відповідно до цільового значення;

• вираз тренду – визначає умови, необхідні для розрахунку тренду.

Як вираз значення використовується середнє значення вмісту мікотоксинів. Код виразу:

[Measures].[Mycotoxin Value]/[Measures].[Fact Mycotoxins Count].

Цільовим значенням є умова, при якій за середнім значенням вологості більше за 40% значення вмісту мікотоксинів повинно бути більше за 0,1мг/кг. Код виразу:

```
case
when [Measures].[Humidity Value]/[Measures].[Fact Mycotoxins Count]>40
then [Measures].[Mycotoxin Value]/[Measures].[Fact Mycotoxins Count]>.1
 end
```

Вираз Статус представляє результат у діапазоні значень від -1 до 1. Це означає, що при від'ємному значенні поточне не відповідає цільовому значенню, а при додатному – навпаки. Код виразу Статусу:

```
case
when [Measures].[Humidity Value]/[Measures].[Fact Mycotoxins Count]>40 and
[Measures].[Mycotoxin Value]/[Measures].[Fact Mycotoxins Count]>.1 then 1
when [Measures].[Humidity Value]/[Measures].[Fact Mycotoxins Count]>40 and
[Measures].[Mycotoxin Value]/[Measures].[Fact Mycotoxins Count]<=.1 then 0
when [Measures].[Humidity Value]/[Measures].[Fact Mycotoxins Count]>40 and
[Measures].[Mycotoxin Value]/[Measures].[Fact Mycotoxins Count]<=.005 then -1
end.
```

Тренд також визначається коефіцієнтом як Статус. Проте він визначає трохи інший діапазон умов, щоб змоделювати тренд на майбутнє. Код виразу:

```
case
when [Measures].[Humidity Value]/[Measures].[Fact Mycotoxins Count]>30 and
[Measures].[Mycotoxin Value]/[Measures].[Fact Mycotoxins Count]>.5 then 1
when [Measures].[Humidity Value]/[Measures].[Fact Mycotoxins Count]>40 and
[Measures].[Mycotoxin Value]/[Measures].[Fact Mycotoxins Count]>.1 then 0
when [Measures].[Humidity Value]/[Measures].[Fact Mycotoxins Count]>50 and
[Measures].[Mycotoxin Value]/[Measures].[Fact Mycotoxins Count]<=.001 then -1
end
```

Результати розрахунку виявились такими:

• середнє значення вмісту мікотоксинів дорівнює 0,34 мг/кг;

• поточне значення відповідає цільовому;

• статус додатній;

• тренд нульовий.

Результати зображені на рис. 5.

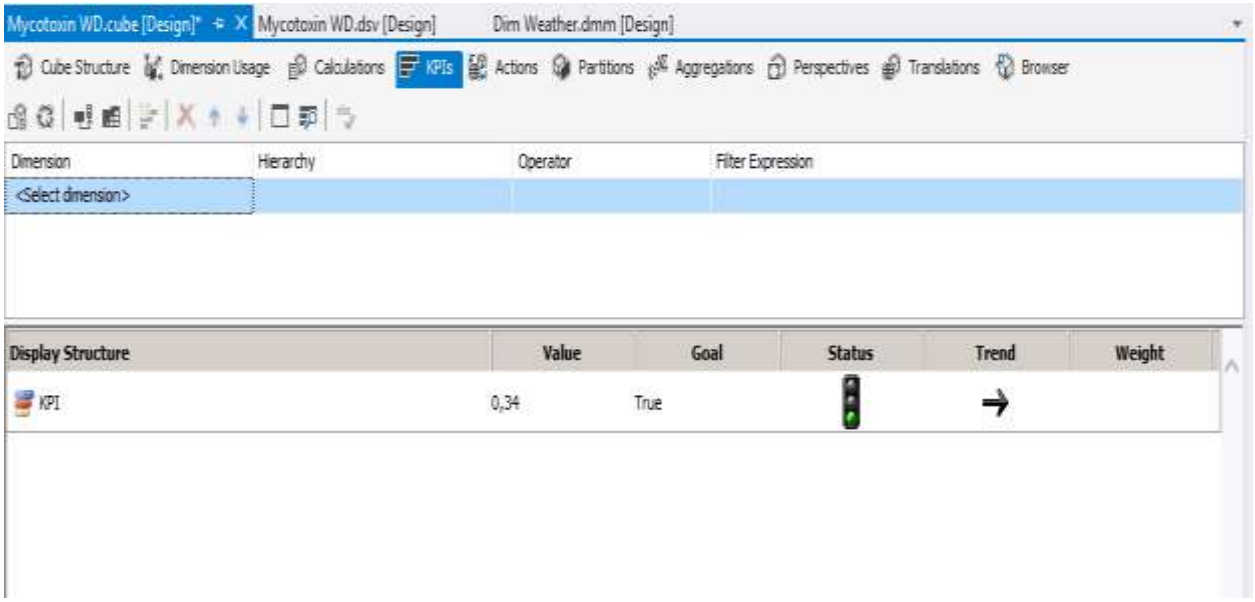

Рисунок 5 – Результати розрахунку KPI

Отже, за результатами розрахунку можна зробити висновок, що при більшій вологості концентрація мікотоксинів теж більша. Вплив вологості на поширення мікотоксинів існує.

6. Висновки

Отримавши гіпотезу шляхом використання методу пошуку асоціативних правил, з'явилась необхідність підтвердити чи відхилити дану гіпотезу. Перевірку гіпотез можна проводити засобами OLAP-аналізу.

Було проведено OLAP-аналіз, який дав значення показника KPI впливу вологості на концентрацію. За даними показника виявляється, що при збільшенні вологості збільшується і концентрація мікотоксинів.

Гіпотеза базується на правилі, що при вологості понад 64,4% концентрація більша за значення 1,24 мг/кг. Ця гіпотеза не суперечить показникам KPI OLAP-аналізу. Отже, таким чином гіпотеза підтверджується.

Проте слід зауважити, що аналіз проводився на тестовій невеликій кількості даних. Тому при збільшенні даних ці показники можуть змінитися.

СПИСОК ДЖЕРЕЛ

1. Зінченко О.І., Салатенко В.Н., Білоножко М.А. Рослинництво: підручник. К.: Аграрна освіта, 2001. 591 с.
2. Скляр В.Г. Екологічна фізіологія рослин: підручник / під заг. ред. Ю.А. Злобіна. Суми: ВТД «Університетська книга», 2015 . 271 с. ISBN 978-966-680-759-8.
3. Klem K., Váňová M., Hajšlová J., Lancová K., Sehnalová M. A neural network model for prediction of deoxynivalenol content in wheat grain based on weather data and preceding crop. *Plant Soil and Environment.* 2007. Vol. 53 (10). P. 421–429. DOI:10.17221/2200-PSE.
4. Chauhan Ya., Tatnell J., Krosch S., Karanja J., Gnonlonfin B., Wanjuki I., Wainaina J., Harvey J. An improved simulation model to predict pre-harvest aflatoxin risk in maize. *Fields Crops Research.* 2015. Vol. 178. P. 91–99. DOI: 10.1016/j.fcr.2015.03.024.
5. Battilani P., Camardo L.M., Rossi V., Giorni P. AFLA-maize, a mechanistic model for Aspergillus flavus infection and aflatoxin B1 contamination in maize. *Comput. Electron. Agric*. 2013. Vol. 94. P. 38–46.
6. Chauhan Y.S., Wright G.C., Rachaputi N.C. 2008. Modelling climatic risk of aflatoxin contamination in maize. *Aust. J. Exp. Agric.* 2008. Vol. 48. P. 358–366.

7. Golub B.L., Gudz A.V., Bushma A.V. Decision support information system in the process of growing biotechnical objects. *Mathematical machines and systems.* 2018. N 4. P. 26–35.
8. Hudz O.V., Karpiuk A.D., Holub B.L., Dudnyk A.O., Bushma A.V. Optical sensor for the detection of mycotoxins. *Semiconductor Physics, Quantum Electronics & Optoelectronics*, 2021. Vol. 24, N 2. P. 227–233.

Стаття надійшла до редакції 14.07.2021