УДК 004

С.В. Знахур, Л.В. Знахур

Харківський національний економічний університет ім. С. Кузнеця, Харків

ОСОБЛИВОСТІ РЕАЛІЗАЦІЇ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ДЛЯ АНАЛІЗУ ЕКОНОМІЧНОГО ПОТЕНЦІАЛУ ПІДПРИЄМСТВ РЕГІОНУ НА БАЗІ AZURE MACHINE LEARNING

Робота присвячена особливостям побудови інтелектуальної системи для обробки даних підприємств регіону на основі використання сервісів Azure та Machine Learning на базі ML Studio Azure.

***Ключові слова**: Azure, SQL, Data Mining, Machine Learning, економічний потенціал, хмарні сервіси, кластеризація.*

Вступ

Інтелектуальний аналіз даних є процесом виявлення закономірностей у великих обсягах даних. При цьому найбільша цінність і нетривіальність одержуваних знань можлива при аналізі великих обсягів структурованих та неструктурованих даних, які можуть надходити з різних джерел (БД, РБД, XML, HTML, TXT). Значним кроком розвитку інтелектуального аналізу стало створення платформ SaaS (Software as a service), на яких розробник розміщує створений ним програмний продукт, забезпечує його повне обслуговування та розвиток, а замовники отримують доступ до ресурсів через Інтернет, а також різні типи хмарних сервісів. Прикладом є використання NoSQL-технології створення сховища даних Google BigQuery, сервіс обробки цих даних використовує алгоритми машинного навчання Google Prediction API; платформа хмарних сервісів корпорації Microsoft Azure з сервісом Azure Machine Learning Center [1]. Завдяки подібним рішенням дослідники отримують можливість реалізовувати складні обчислювальні експерименти в оперативному режимі, застосувати алгоритми Data Mining, ML (maching lerning) та штучного інтелекту, аналізувати якість навчання і ступінь корисності отриманих результатів.

Основна частина

Метою роботи є аналіз можливостей Azure та хмарних сервісів для інтелектуальної обробки неструктурованих та структурованих даних, які накопичуються з різних джерел (на прикладі аналізу потенціалу та інтелектуального капіталу організацій та підприємств Харківського регіону).

На ринку хмарних платформ існує жорстка конкуренція між рішеннями Amazon, Google, Microsoft, IBM, Oracle, які надають сервіси для обробки та аналізу даних, але Windows Azure має низку переваг, які можуть бути цікаві з практичної точки зору. Наприклад, Google та Amazon Web Services пропонують використовувати тільки хмарні сервіси, а на базі Windows Azure доступні також технології гібридних хмар, що можуть використовувати ресурси замовника і ресурси хмар, також Windows Azure пропонує досить різноманітні технології для хмарної інтелектуальної обробки даних [1; 2]: HDInsight, Data Factory, Machine Learning, Stream Analytics. Сервіс HDInsight базується на платформі Apache Hadoop та дозволяє обробляти великий обсяг інформації в умовах масштабування системи. HDInsight має можливості роботи зі структурованими даними, частково структурованими і неструктурованими. Сервіс використовують для задач пошуку і класифікації, типових задач обробки даних, завдань управління потоками даних, аналітичних задач. Data Factory – це служба для зберігання та обробки даних, що працює із структурованими, та неструктурованими даними, які отримані з локальних та хмарних джерел. Data Factory використовується для підключення розташованих у різних джерелах даних різних типів, введення в експлуатацію конвеєрів даних, інтеграції даних. Machine Learning (ML) – сервіс машинного навчання у хмарі, який призначений для рішення задач статистичного аналізу та Data Mining. Сервіс складається з двох компонентів: Machine Learning Studio (клієнтська частина) і Machine Learning API Service (серверна частина). Сервіс Machine Learning використовується для рішення задач кластерізації, класифікації, прогнозування, пошуку асоціацій та інших. Stream Analytics – це сервіс для аналізу потоку даних в реальному часі. Для отримання потоку даних сервіс взаємодіє з Azure Event Hub і сховищем даних, а для зберігання результатів аналізу з Event Hubs, Blob Storage, Azure SQL Database. Даний сервіс в комплексі з концентраторами подій (Event Hubs) дає можливість обробляти великі обсяги даних в режимі реального часу.

Для нашого дослідження більш докладно розглянемо можливості машинного навчання Machine Learning на базі платформи Azure для рішення зада-

чі побудови інтелектуальної системи щодо аналізу даних підприємств регіону. Загальна схема використання технології Machine Learning в дослідженні наведена на рис. 1. Пропонується використовувати дані, які накопичуються в БД MS SQL DB Azure та дані, які зберігаються на статистичних сайтах, та мають або прямий доступ к форматам CSV, XML, або доступ на основі використання API [2]. Доступ до даних пропонується реалізувати на базі ML Studio Azure. Загальна послідовність інтелектуальної обробки та візуалізації даних включає етапи:

1) формування первинних даних на основі використання WEB технологій роботи з БД (підключення к MS SQL Azure за допомогою PDO ("sqlsrv:server = tcp: proectsqldb.database.windows.net, 1433; Database = ensero", "adminsql", "Proectsqldbhneu"). В роботі введення даних було реалізовано за допомогою WEB-інтерфейсу (форми на лендінговій сторінки проекту), яка також була опублікована на ресурсах Azure;

2) моделювання та розгортання включає підключення до сервісу Machine Learning на базі ML Studio Azure; формування та візуалізацію логіки статистичного експерименту, розробку сценаріїв експерименту, створення WEB-сервісів для доступу до результатів моделювання (формування ключа API);

3) підключення до даних та моделей за допомогою API та формування звітів для візуалізації результатів моделювання на клієнтських пристроях відповідно цілям та цінностям аналізу даних.

На рис. 2 показано інтерфейс підключення до сервісу Machine Learning на порталі Azure.

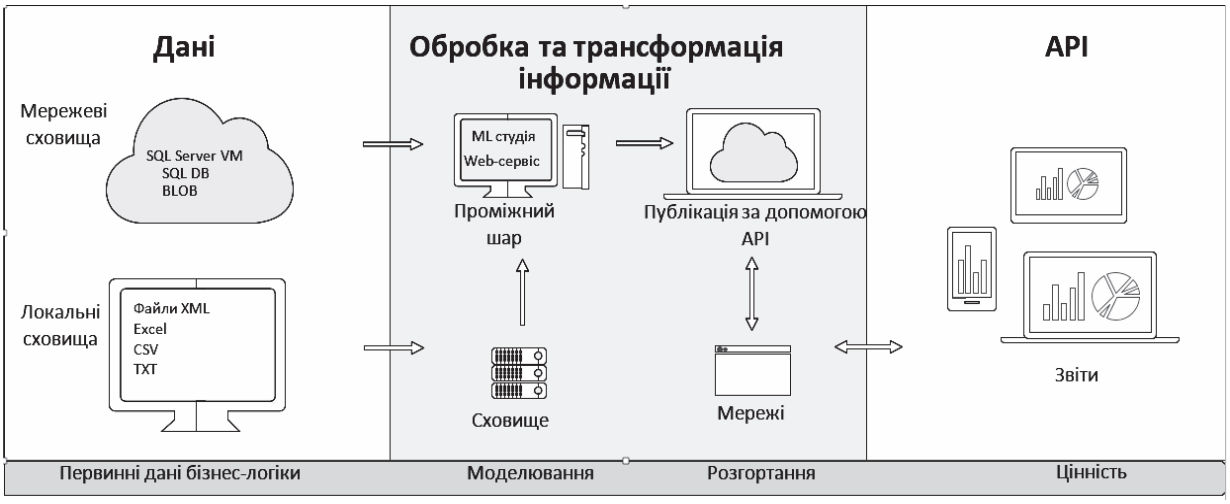


Рис. 1. Загальна схема обробки та візуалізації даних на базі Windows Azure

Рис. 2. Сервіс Azure Machine Learning

Розглянемо загальну постановку задачі статистичного експерименту на базі Azure Machine Learning. Необхідно провести інтелектуальний аналіз даних організацій (знайти сховані закономірності) на основі індикаторів методики оцінки інноваційного потенціалу підприємства [4], які розраховуються спеціалістами організацій та зберігаються у БД (використовується онлайн режим вводу даних за допомогою WEB-форм та БД MS SQL Azure), та статистичних даних України в форматі XML, CSV (дані сайту http://www.economywatch.com/economic-statistics, та http://datacatalog.worldbank.org/ (наприклад, на основі API http://api.worldbank.org/countries/ukr/indicators/NY.GDP.MKTP.CD).

Здатність організацій до інновацій визначається станом їх управлінської системи, якістю персоналу, станом матеріально-технічної бази і маркетингом (ринковою активністю). Відповідно до такого підходу виділяються наступні об'єкти оцінювання інноваційного потенціалу: організаційно-управлінська система; персонал; виробнича і науково-технічна база; ринкова активність; показники поточної фінансово-господарської діяльності. Для інтелектуального аналізу (пошуку закономірностей) було обрано каскад методів [3]: кластерний аналіз (K-means), класифікація на основі методу LM Two Class Locally Deep Support та Regression на базі результатів кластерного аналізу (регресійні моделі для кожного кластеру).

На етапі завантаження первинних даних були імпортовані дані по підприємствах (рис. 3), які включають значення наступних показників (які відібрані експертами згідно методики оцінки інноваційного потенціалу підприємства [4]): середня заробітна плата виробничого персоналу, рівень освіти і кваліфікації менеджерів вищої ланки, середній вік працівників основного виробництва, середній вік ITP, індекс зміни обсягу реалізованої продукції, частка інноваційної продукції в загальному обсязі виробництва, рентабельність виробництва, частка інноваційних витрат в загальних виробничих витратах, знос основного виробничого обладнання (%), частка витрат на навчання персоналу в загальних виробничих витратах, наявність сертифікації виробництва, участь у внутрішніх та міжнародних виставках, ярмарках, конкурсах. База даних включає 48 спостережень, які структуровані за роками. Для аналізу впливу макроекономічних факторів були використані показники: індекс зміни обсягу промислового виробництва, індекс зміни обсягу інвестицій в промисловість, середній вік працездатного населення, рівень інфляції. Наступним етапом є робота з неповними даними, так як за деякими підприємствами були пропуски за значеннями показників. Для їх усунення використовувався елемент Clean Missing Data (рис. 4), за допомогою якого були вилучені недостовірні спостереження.

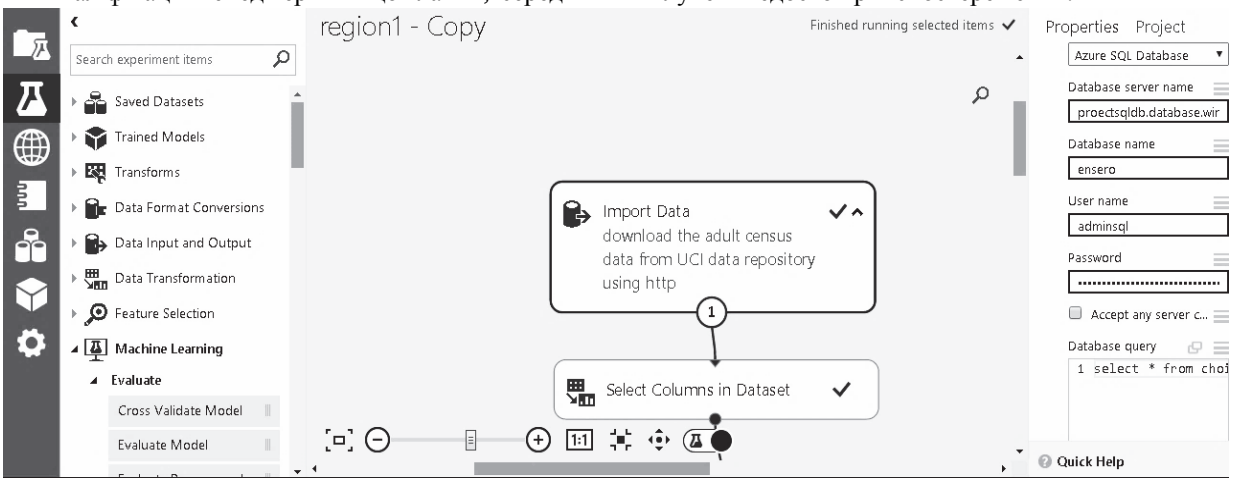


Рис. 3. Імпорт структурованих даних

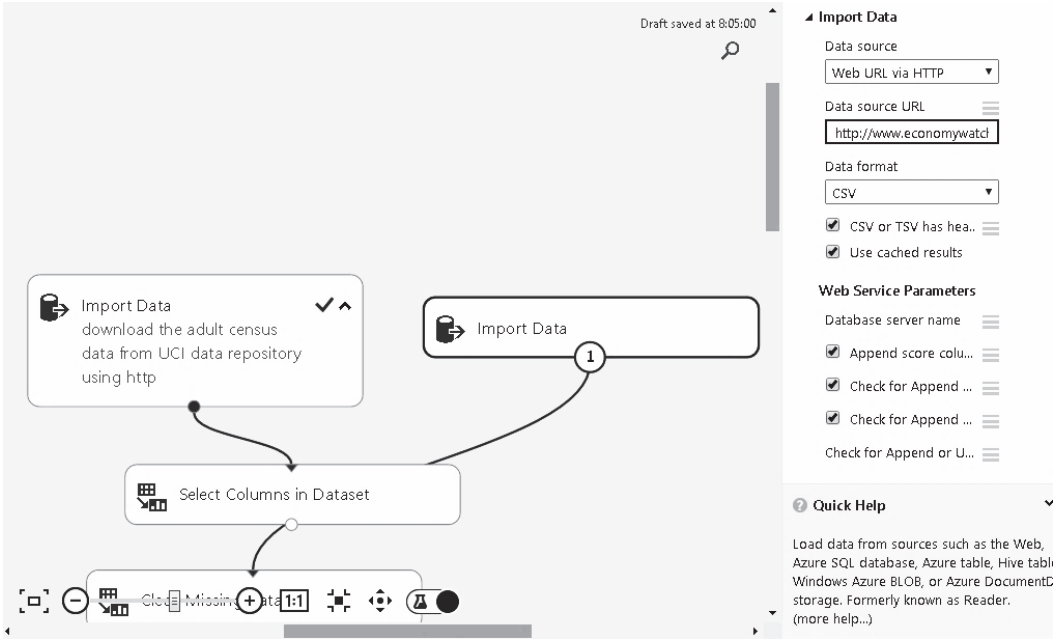


Рис. 4. Імпорт неструктурованих даних та очистка даних

Після цього були видалені дані, які дублюються, оскільки вони можуть помітно знизити точність моделі. Для можливості проведення достовірного аналізу (групування та кластеризації) додатково був поставлено фільтр (на рівні sql запиту) для отримання даних тільки відповідного року. Важливим етапом первинної обробки є нормалізація (стандартизація) даних. Для рішення задач кластеризації це може бути принциповим щодо отримання якісної моделі. У нашому випадку варіація за показниками досить значна, тому необхідно зробити нормалізацію (обрано метод MinMax).

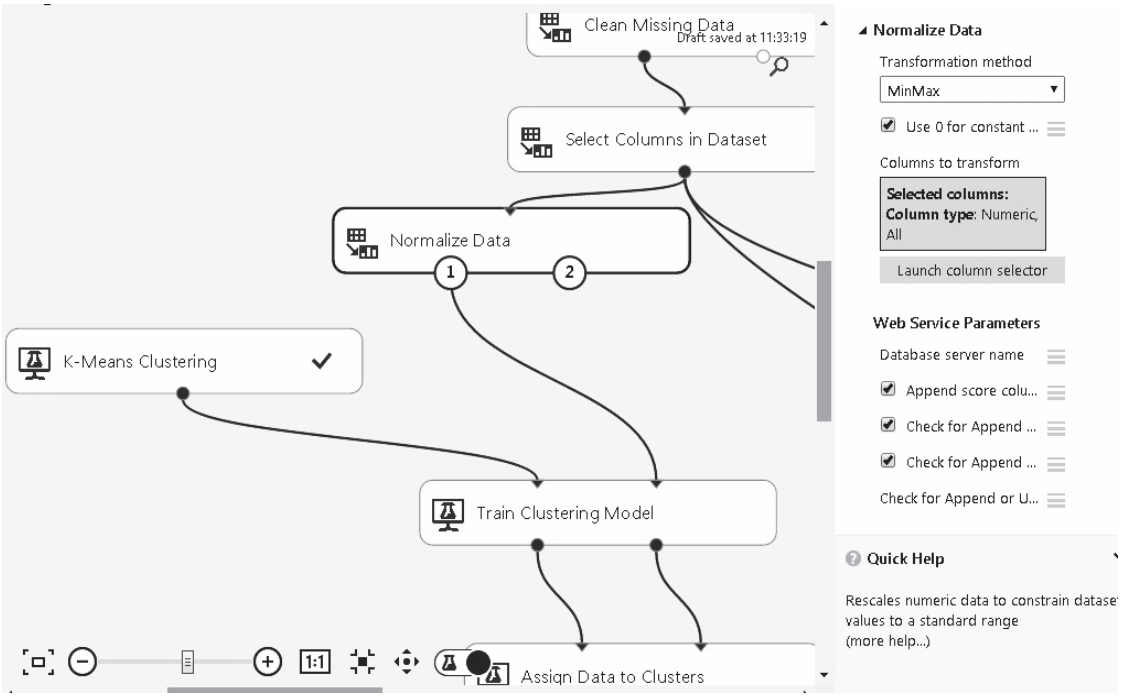


Рис. 5. Нормалізація даних

Ключовим методом інтелектуального аналізу в роботі було обрано метод K-means (для первинної кластерізації даних за ознаками). На рис. 6 показано, що було обрано 3 центри кластерів та Евклідова відстань у якості параметрів методу.

Фрагмент загальної схеми проведення експериментів щодо формування train моделей, evaluate моделей експлуатації та формування параметрів WEB-сервісу наведено на рис. 7. Важливим є здійснення асоціації отриманих результатів кластеризації (Assign Data to Clusters) з даними для можливості подальшого аналізу кожного кластеру (в рамках проведення регресійного аналізу).

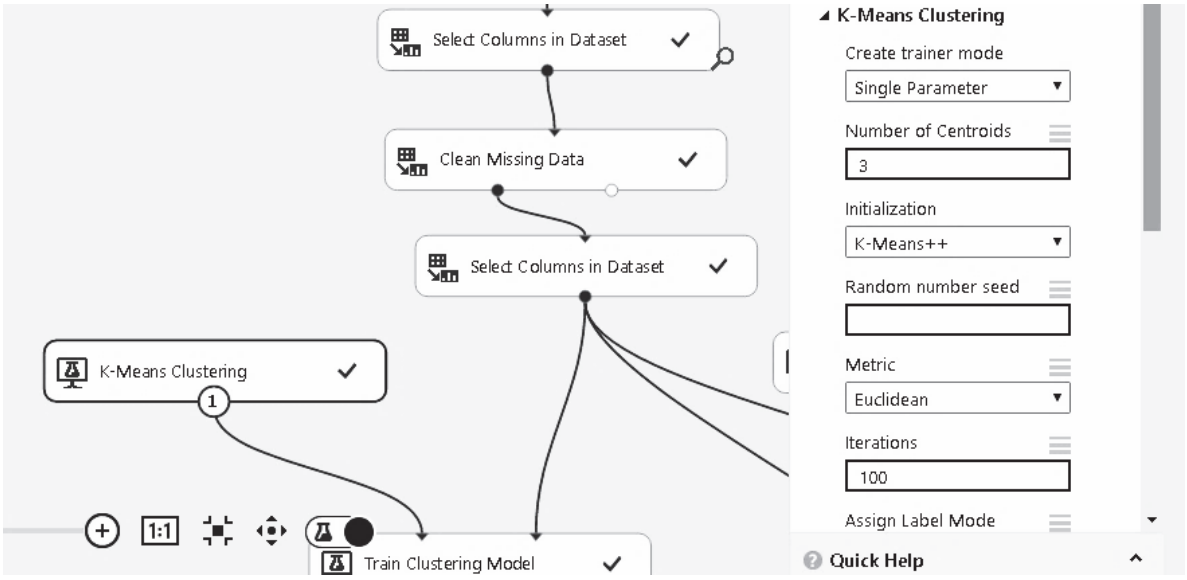


Рис. 6. Налаштування методу K-means

Після проведення моделювання згідно схеми (рис. 7) необхідно визначити механізм публікації результатів. Доступ до даних аналізу було організовано за допомогою API ключа (рис. 8). Також можливо використати клієнтський додаток (наприклад, Excel) для підключення моделей, параметрів, даних на рівні кожного користувача. Приклад доступу до результатів кластерного аналізу наведено на рис. 9.

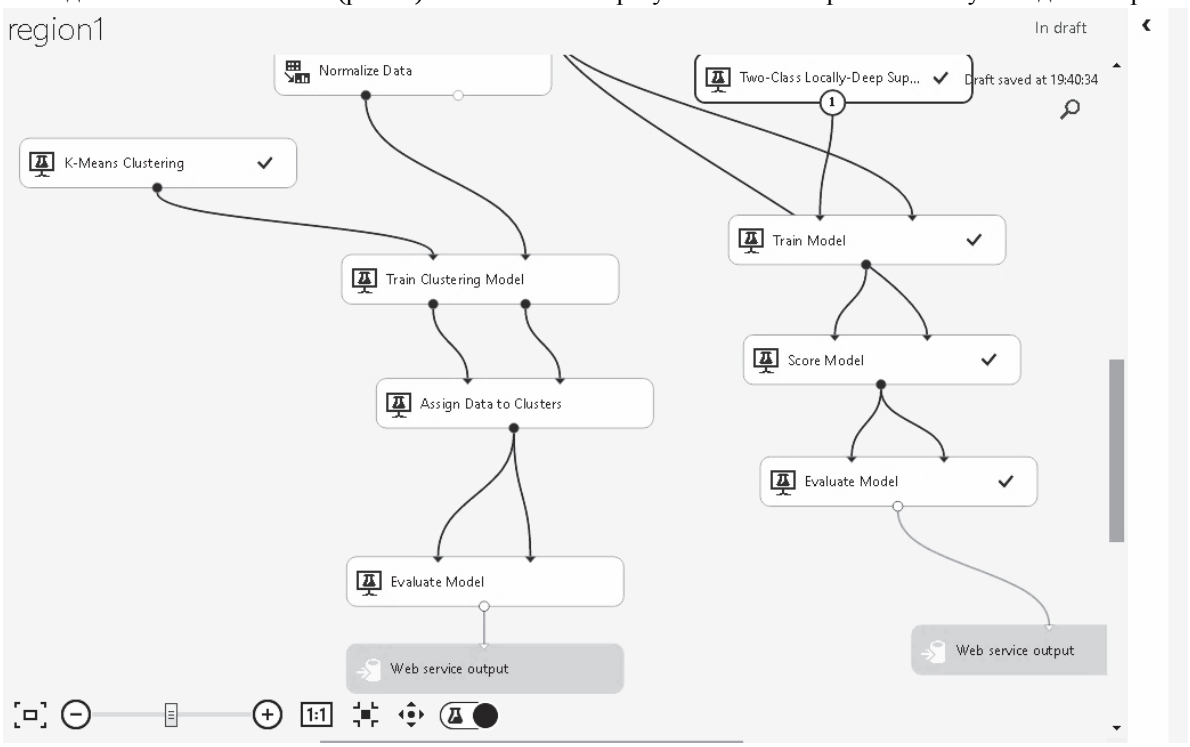
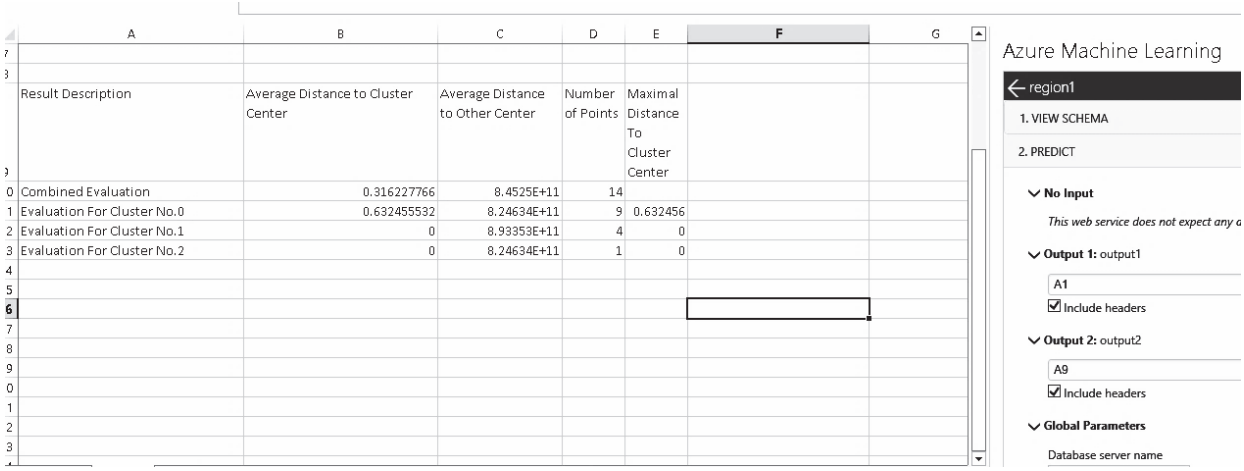


Рис. 7. Фрагмент загальної моделі щодо налаштування методів, train моделі, моделі експлуатації та формування параметрів WEB-сервісу

Рис. 8. Генерація API ключа та клієнтських додатків (Excel)



	A	B	C	D	E	F	G
	Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center		
0	Combined Evaluation	0.316227766	8.4525E+11	14			
1	Evaluation For Cluster No.0	0.632455532	8.24634E+11	9	0.632456		
2	Evaluation For Cluster No.1	0	8.93353E+11	4	0		
3	Evaluation For Cluster No.2	0	8.24634E+11	1	0		

Рис. 9. Результати в Excel

Висновок

Наведена у роботі задача розробки онлайн орієнтованої інтелектуальної системи була реалізована за допомогою хмарної платформи Microsoft Azure Machine Learning для рішення задач кластерізації, класифікації, регресійного аналізу даних підприємств регіону в рамках проекту оцінки їх економічного потенціалу. У процесі роботи були проаналізовані сучасні технології обробки великих обсягів даних; алгоритми Data Mining для каскадного аналізу даних в Azure ML. Також була побудована система моделей та проаналізовані можливості Azure щодо підключення до результатів моделювання через API для масштабування інтелектуальної системи та можливості підключення значної кількості користувачів (спеціалістів підприємств). Наступним кроком є розвиток системи моделей та визначення інтелектуальних тригерів щодо вибору найкращих алгоритмів для аналіза згруппованих за часом даних (наприклад, панельних даних).

Список літератури

1. Машинное обучение [Електронний ресурс]. – Режим доступу: https://azure.microsoft.com/ru-ru/services/machine-learning/.

2. Построение гибридных приложений в облаке на платформе Microsoft Azure http://download.microsoft.com/download/0/F/B/0FBFAA46-2BFD-478F-8E56-7BF3C672DF9D/.

3. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – 2-е изд., перераб. и доп. / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко. – СПб.: БХВ-Петербург, 2007. – 384 с.

4. Оценка инновационного потенциала предприятия (Методические рекомендации) [Електронний ресурс]. – Режим доступу: www.inno.kharkov.ua/wp-content/uploads/.../metodika.doc.

Надійшла до редколегії 13.03.2017

Рецензент: д-р техн. наук проф. М.Д. Годлевський, НТУ «ХПІ», Харків.

ОСОБЕННОСТИ РЕАЛИЗАЦИИ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ АНАЛИЗА ЭКОНОМИЧЕСКОГО ПОТЕНЦИАЛА ПРЕДПРИЯТИЙ РЕГИОНА НА БАЗЕ AZURE MACHINE LEARNING

С.В. Знахур, Л.В. Знахур

Работа посвящена особенностям построения интеллектуальной системы для обработки данных предприятий региона на основе использования сервисов Azure и Machine Learning на базе ML Studio Azure.

Ключевые слова: *Azure, SQL, Data Mining, Machine Learning, экономический потенциал, облачные сервисы, кластеризация.*

FEATURES IMPLEMENTATION OF THE ECONOMIC ANALYSIS OF INTELLECTUAL SYSTEM BUILDING ON ENTERPRISES OF THE REGION BAZEILLES AZURE MACHINE LEARNING

S.V. Znahur, L.V. Znahur

The work is devoted to the features of building an intelligent system for processing data of enterprises in the region based on the use of Azure and Machine Learning services based on ML Studio Azure.

Keywords: *Azure, SQL, Data Mining, Machine Learning, economic potential, cloud services, clustering.*