Development of ML model for coding of economic activities and occupations in household surveys (project in progress)

Mark Feldman
Deputy Director of Senior Microeconomics
Central Bureau of Statistics
markf@cbs.gov.il

It's All in the Numbers

In the Presentation :

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

- Summary and recommendations
- 2019 data test results
- Machine learning model integration into a coding system
- Requirements for staff for the development and ongoing maintenance of a model
- Technical model details (Amit Shkolnik, Machine Learning Specialist)

Summary and recommendations

- Israel CBS is soon to launch an Artificial Intelligence model for automatic classification of economic activity and occupation according to ISIC & ISCO standards.
- The model has been developed by Amit Shkolnik – a natural language processing and machine learning advisor, under the supervision of Mark Feldman- deputy director of senior micro economic department.
- On 2021 the AI model will run under supervision of CBS team of classifiers and it supposes to save up to 30% of manual work.

Summary and recommendations

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

- The model was trained using dataset containing 700,000 records manually coded in previous years (2013-2018) and tested on an 80,000 records test set (2019).
- K-nearest neighbors (KNN) model was trained to classify economic activity and occupation.
- Today, the model reached 74.1% accuracy on economic activity and 68.5% accuracy on occupation.

Summary and recommendations

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

- The results of the model implementation on the data for 2019 will be presented below. The project was carried out in collaboration with the Information Systems Department and the Statistical Methodology Department. After integrating this model into the CBS' production processes, it will need the Information Systems Department's support and maintenance and the Statistical Methodology Department's model improvement.
- At this moment the project is at the stage of preparing integrating this model into the CBS' production processes.

Summary and recommendations

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

- **Main conclusion:** even at this stage this model implementation will make it possible to increase the work efficiency and about 30% of cases will not be sent to the manual coding. This figure is very important considering the preparations for the 2021 census and the need to recruit coders for an economic activity and occupation coding in the census. In my opinion, the model can be improved even more, and as a result we can reach even higher efficiency levels.

2019 data test results

The file includes 80,000 records - data from 2019, for each record there are codes (4 digits) of economic activity and occupation from three different sources: from the automatic coding system, from the ML and final code from coding division, a total of six variables per record.

The table below shows the percentage of cases where the systems failed to code them completely (at least one of the digits received – X, which means that the economic activity or occupation is not completely recognized according to the existing classifications):

		The number of not completely coded cases	The percentage of not completely coded cases
ML	economic activities	5728	7.2
Automatic coding system	economic activities	18719	23.4
Final coding	economic activities	5266	6.6
ML	occupations	6053	7.6
Automatic coding system	occupations	32355	40.4
Final coding	occupations	5377	6.7

2019 data test results

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

Test results: In some cases there is an exact fit between the codes coded by the ML system to the final coding.

The testing shows that the ML system succeeds in coding the "correct" economic activities in 74.1% of cases, occupations in 68.5% of cases, and the "correct" coding of economic activities and of occupations at the same time in 54.1% of cases.

		The number of completely coded cases in both systems	The number of cases with exact fit	The percentage of cases with exact fit	The percentage out of 80,000 records
ML, final coding	economic activities	72528	53724	**74.1**	**67.2**
ML, final coding	occupations	71671	49063	**68.5**	**61.3**
ML, final coding	economic activities and occupations	68939	37722	**54.7**	**47.2**

2019 data test results

Test results: In some cases there is an exact fit between the codes coded by the automatic coding system and the ML.

The testing shows that 22,145 cases were coded with the same codes by both systems, which is about 27.7% of the total cases. Notably, 19,899 (89.9%) of the same cases were coded in the same way in the final coding.

In 53.6% of the cases the code of the economic activity was the same, and in 38.4% of the cases the code of the occupation was the same.

		The number of completely coded cases in both systems	The number of cases with exact fit	The percentage of cases with exact fit	The percentage out of 80,000 records
ML, automatic coding system	economic activities	58975	42875	72.7	**53.6**
ML, automatic coding system	occupations	46093	30706	66.6	**38.4**
ML, automatic coding system	economic activities and occupations	38926	22145	56.9	**27.7**

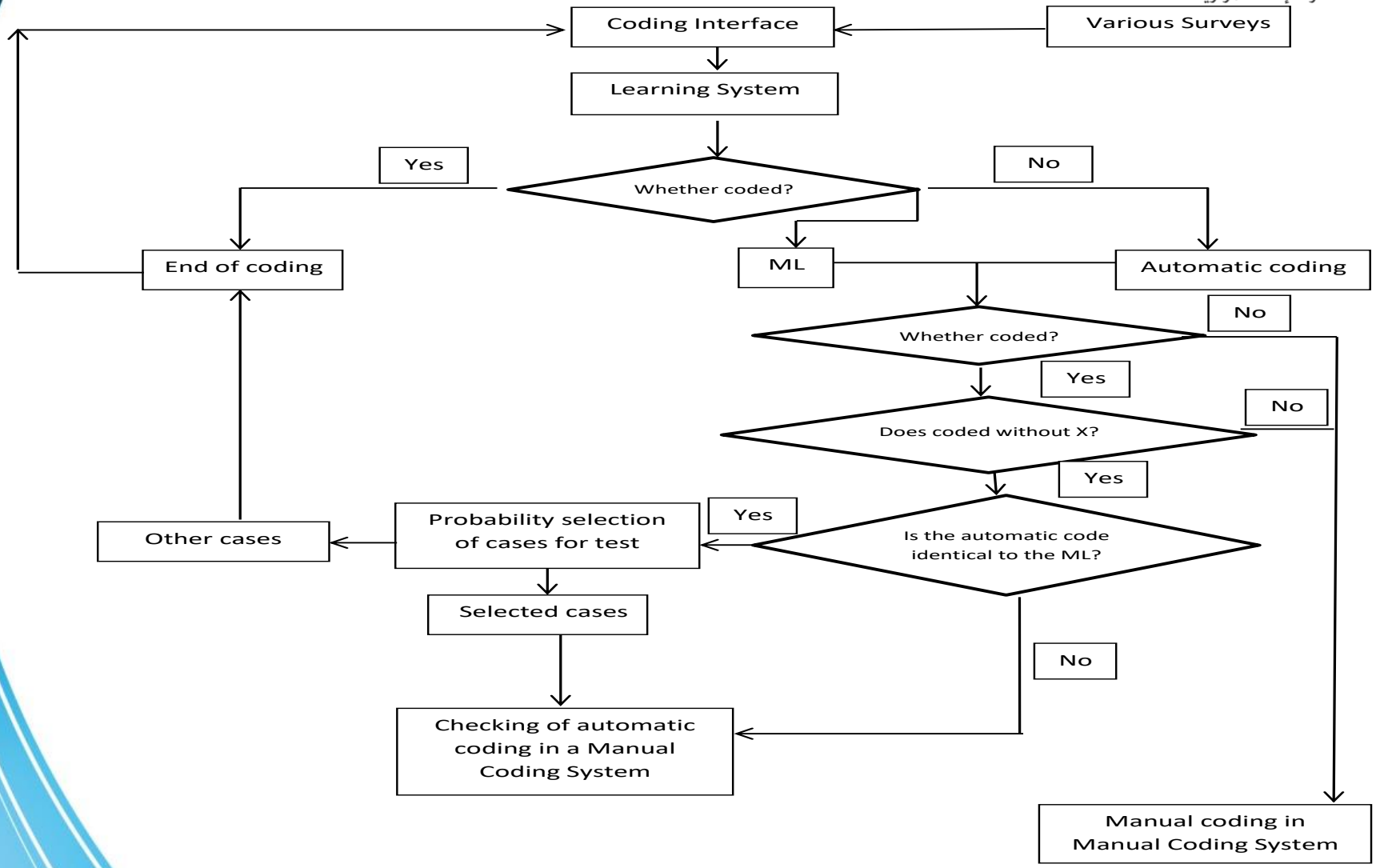
Conclusion: At this stage, if the ML system is integrated into the CBS coding processes, 27.7% of the cases will not be sent to the coding in a manual coding system. If by the 2021 census we manage to improve the system, which includes more case learning, the savings percentage will be even greater.

ML model integration into a coding system

03.09.2020

Draft version 1

Flow chart of a new coding system as of September 2020 (including ML)



The term "coded", "code" refers to having a full symbol of 4 digits both in economic activities and occupations.

Requirements for staff for the development and ongoing maintenance of a model

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

- Relevant departments :

- **Senior Department of Micro-Economic** (LFS Department, Coding Division). Thematic support for economic activity and occupation, actual manual coding. The function of some of the coders in the coding division will change and additional level of knowledge (academic) will be required to identify problems in the machine learning system, to make conclusions and to give feedback for continuous improvement of the model.
- **Senior Information Systems Department**. Routine maintenance of a coding system including a machine learning model.
- **Senior Department of Statistical Methodology**. Methodological support for the methodological improvement of a machine learning model and an increase in the percentage of coded cases.

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

Technical model details (Amit Shkolnik, Machine Learning Specialist)

Software and hardware requirements

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

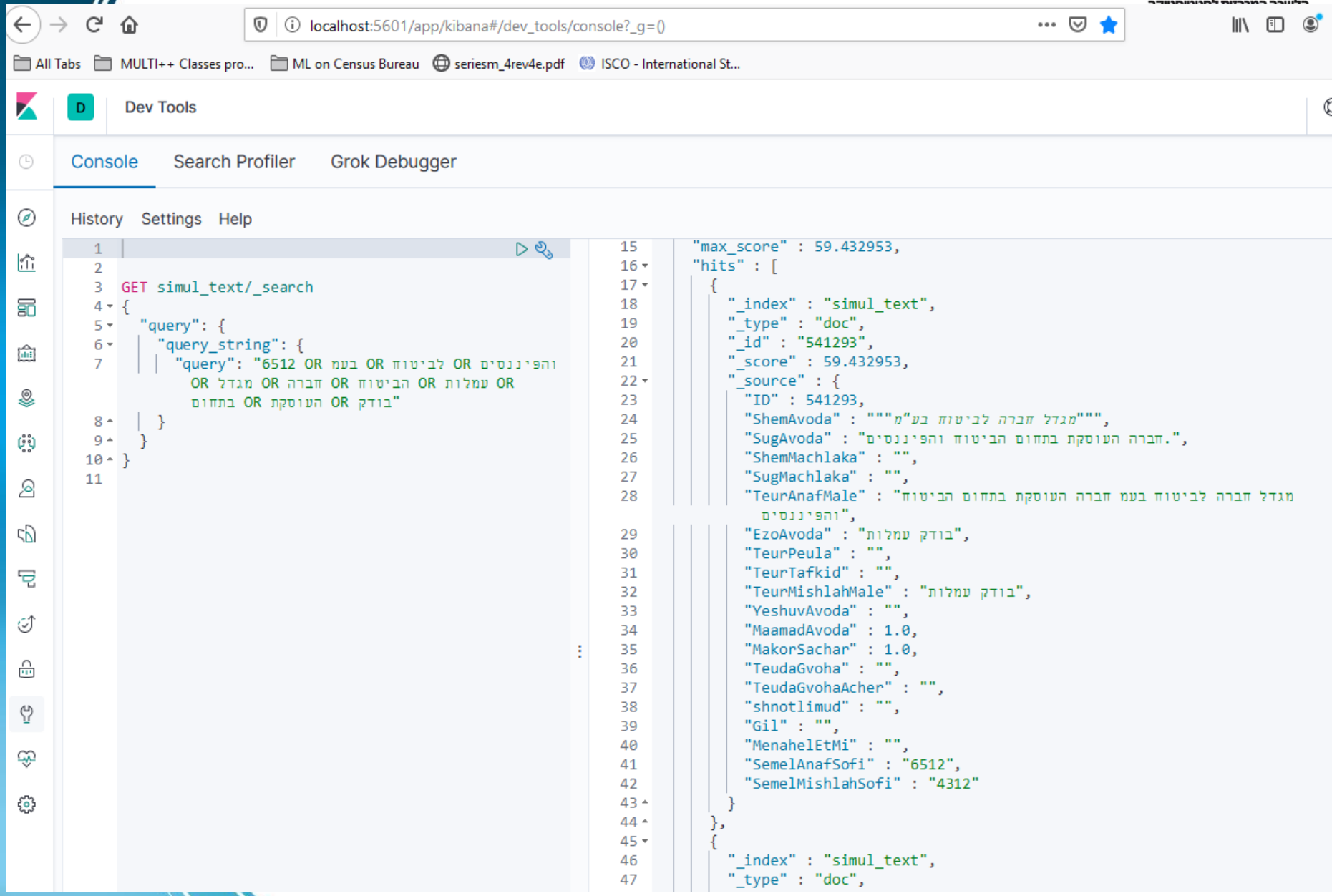
- Operation system: WINDOWS 10 or Server WINDOWS 2016.
- Programming language: Anaconda Python 3.7.
- Searching server: ELASTIC 7.5.
- Visualization server for ELASTIC: KIBANA 7.5.2

Determining the level of accuracy and confidence

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

Final_Precision	Finaly_handled_prcnt	F1	accuracy	threshold
0.7647	1	0.867	0.765	1
0.795	0.942	0.878	0.791	1.1
0.8018	0.924	0.878	0.793	1.2
0.8112	0.906	0.88	0.799	1.3
0.8176	0.892	0.88	0.802	1.4
0.8243	0.872	0.878	0.801	1.5
0.8332	0.852	0.878	0.803	1.6
0.8391	0.837	0.877	0.803	1.7
0.8484	0.823	0.879	0.808	1.8
0.8572	0.804	0.879	0.809	1.9
0.8721	0.769	0.875	0.808	2
0.8812	0.749	0.872	0.806	2.1
0.8826	0.739	0.868	0.801	2.2
0.8838	0.734	0.865	0.798	2.3
0.8844	0.728	0.863	0.795	2.4
0.8875	0.721	0.861	0.794	2.5
0.8896	0.715	0.86	0.793	2.6
0.8905	0.712	0.858	0.791	2.7
0.893	0.704	0.856	0.788	2.8
0.8961	0.694	0.853	0.785	2.9
0.9018	0.686	0.853	0.786	3
0.9026	0.681	0.85	0.784	3.1
0.9049	0.675	0.849	0.782	3.2
0.9045	0.672	0.846	0.779	3.3
0.9092	0.667	0.847	0.781	3.4

Server ELASTIC and Interface KIBANA



Determining the level of accuracy and confidence

Final_Precision	Finaly_handled_prcnt	F1	accuracy	threshold
0.7647	1	0.867	0.765	1
0.795	0.942	0.878	0.791	1.1
0.8018	0.924	0.878	0.793	1.2
0.8112	0.906	0.88	0.799	1.3
0.8176	0.892	0.88	0.802	1.4
0.8243	0.872	0.878	0.801	1.5
0.8332	0.852	0.878	0.803	1.6
0.8391	0.837	0.877	0.803	1.7
0.8484	0.823	0.879	0.808	1.8
0.8572	0.804	0.879	0.809	1.9
0.8721	0.769	0.875	0.808	2
0.8812	0.749	0.872	0.806	2.1
0.8826	0.739	0.868	0.801	2.2
0.8838	0.734	0.865	0.798	2.3
0.8844	0.728	0.863	0.795	2.4
0.8875	0.721	0.861	0.794	2.5
0.8896	0.715	0.86	0.793	2.6
0.8905	0.712	0.858	0.791	2.7
0.893	0.704	0.856	0.788	2.8
0.8961	0.694	0.853	0.785	2.9
0.9018	0.686	0.853	0.786	3
0.9026	0.681	0.85	0.784	3.1
0.9049	0.675	0.849	0.782	3.2
0.9045	0.672	0.846	0.779	3.3
0.9092	0.667	0.847	0.781	3.4

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

Thank you for you attention!