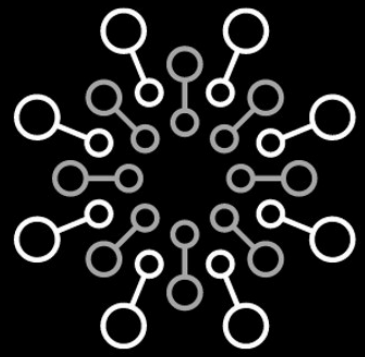


Hierarchical Multi-label Classification (HMLC) Delivery Accelerator

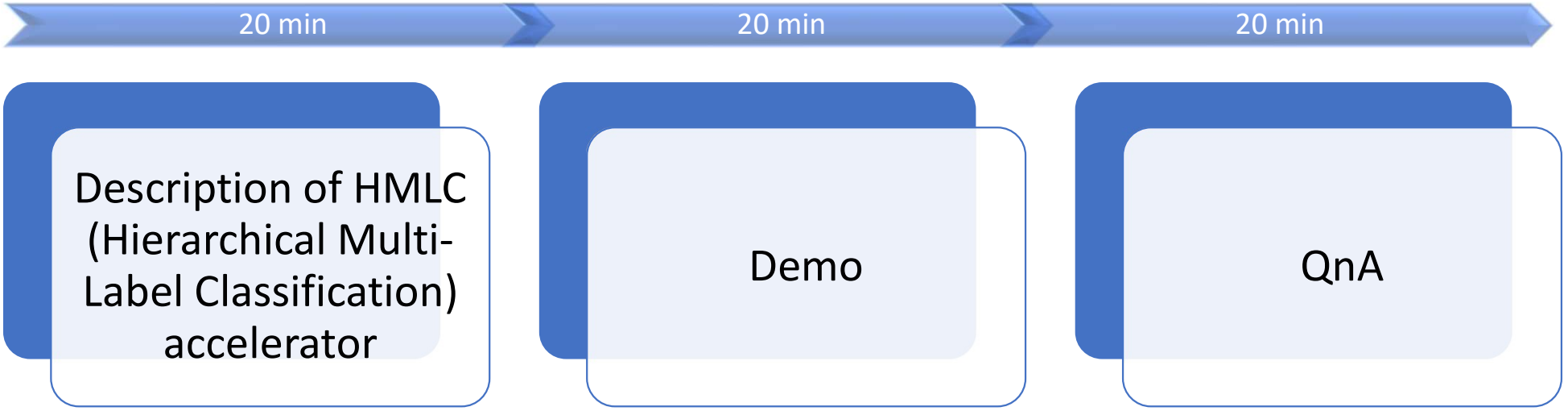
Hierarchical Multilabel Classification (sharepoint.com)

Senani Nori

Tarun Dugar

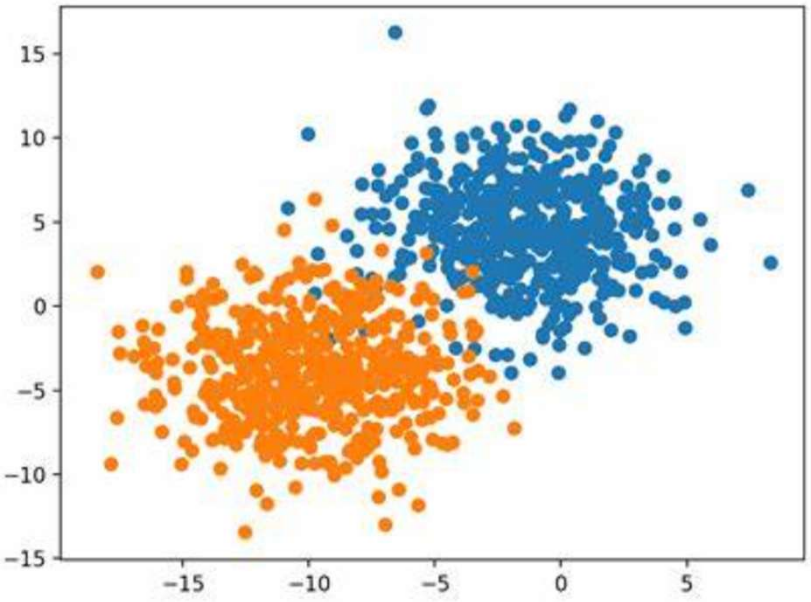
13 July 2022

Agenda

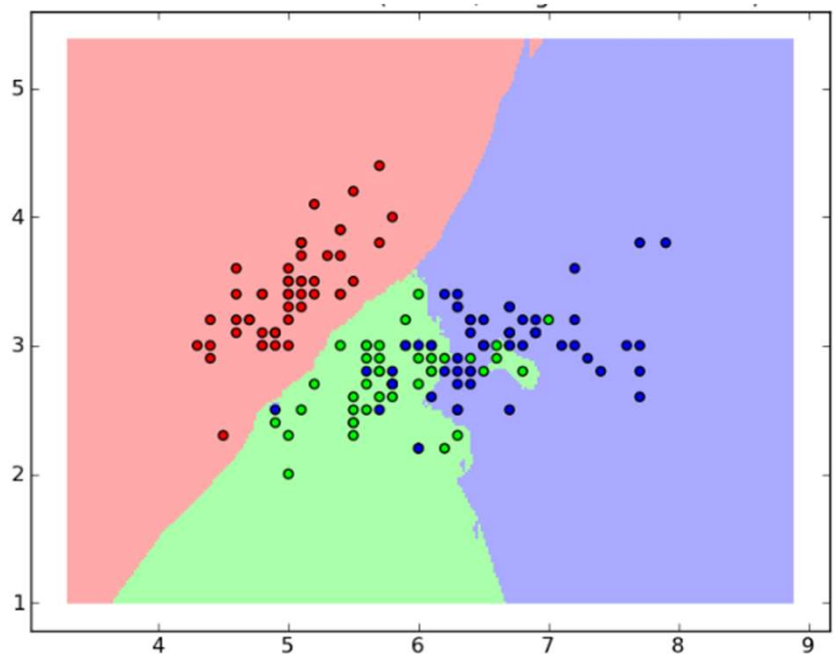


- What is Multi-Label and hierarchical?
  - Examples
- What is the complexity of the problem?
- Description of the solution

Binary Classifier



Multi-class Classifier



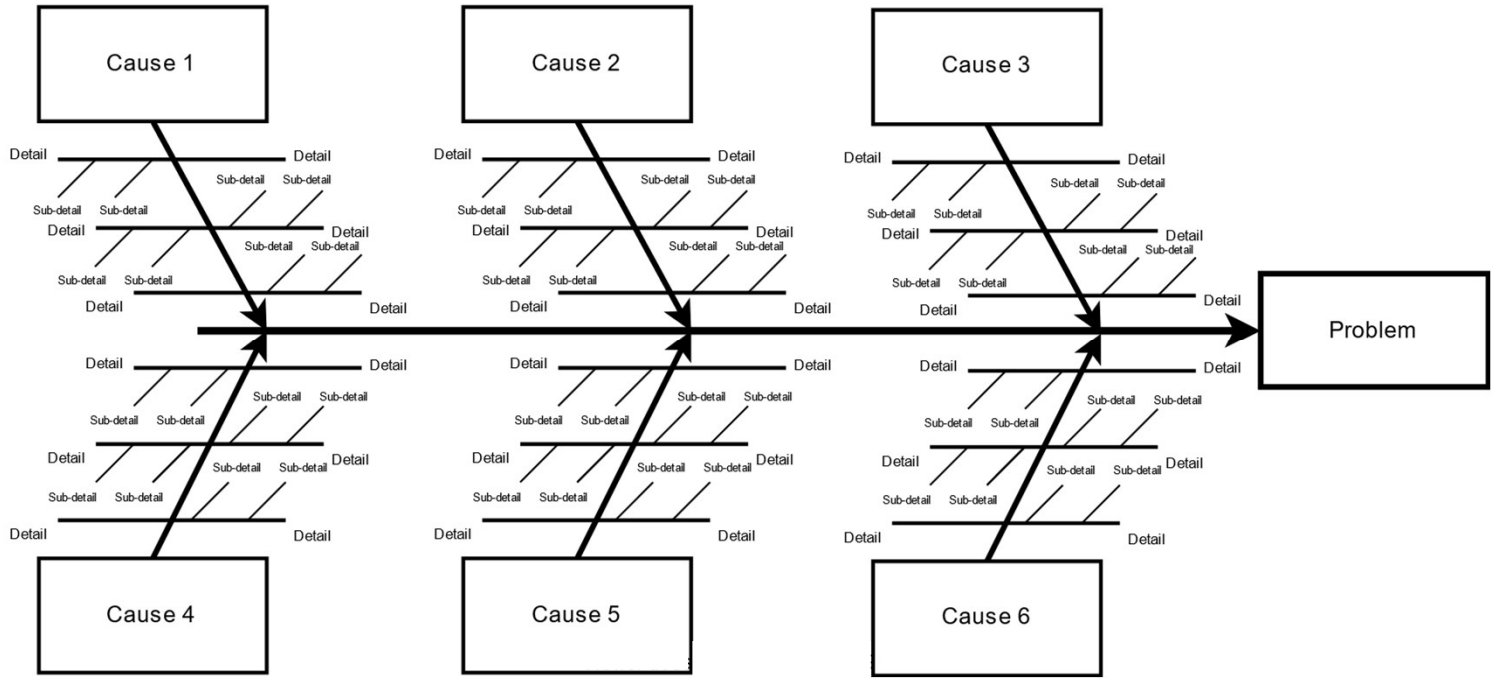
Multi-label / Multi-level Classifier

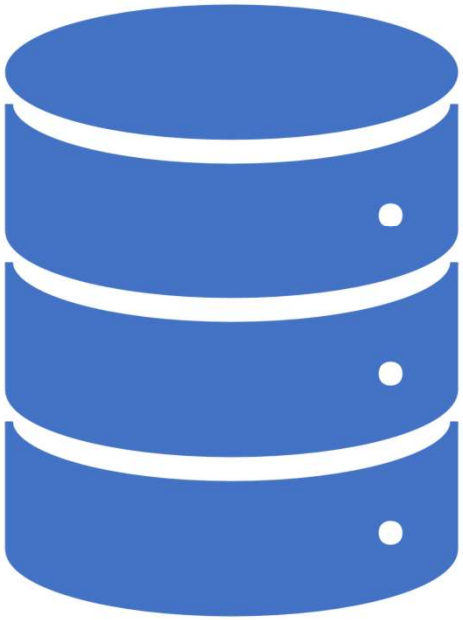
Hierarchical Multi-label Classifier

Description	Market	Sector	Brand	Sub-brand
VASELINE BLOT ADVANCED STRGTH 4X4X400ML	HAND & BODY CARE	HAND & BODY EXCL SUN	VASELINE	VASELINE BODY LOTION ADVANCE STRENGTH
CITRA BRIGHT WH AURA GEL 30X4X35ML	HAND & BODY CARE	HAND & BODY EXCL SUN	CITRA	CITRA BL BRIGHT WHITE AURA GEL
VASELINE ALOE SOOTHE LOT N 12X550ML	HAND & BODY CARE	HAND & BODY EXCL SUN	VASELINE	VASELINE BODY LOTION ALOE SOOTHE
OMO PLUS PERFUME SCARLETT 18B 6X6X220G	FABRIC CLEANING	FABRIC SOLUTION WASH	OMO	OMO NM CONC POWDER PLUS PERFUME
BREEZE EXCEL GOLD SOLAR 12X900G	FABRIC CLEANING	FABRIC SOLUTION WASH	BREEZE	BREEZE EXCEL GOLD
COMFORT FW BLOOMING CLEAN 12X12X75G	FABRIC CLEANING	PREMIUM CARE DETERGENT	COMFORT	COMFORT PRM CR NMC PWDR BLOOMING CLEAN
BREEZE EXCEL GOLD GAI 6X2500G	FABRIC CLEANING	FABRIC SOLUTION WASH	BREEZE	BREEZE EXCEL GOLD
PONDS MENFO ACNE OIL CONTROL P2X6X100G	FACE	FACE CLEANSING	PONDS MEN	PONDS MEN FOAM OIL GREEN
KNORR CUBE SHITAKE (NEW) P12X6X60G	BOUILLONS & SEASONINGS	BOUILLONS	KNORR	KNORR BASIC CUBE
TRESEMME MASQUE DEEP REPAIR 4X3X180ML	WASH & CARE	TREATMENTS	TRESEMME	TRESEMME TREATMENT
TRESEMME SERUM SPLIT REMEDY IMP 6X97ML	STYLING	OTHER HAIR STYLING	TRESEMME	TRESEMME OTH STYLING SERUM SPLIT REMEDY
LUX SHW CRM MAGICAL SPELL BT P8X3X220ML	SKIN CLEANSING	BODY CLEANSING	LUX	LUX SHOWER CREAM

This data snippet is shown only to give a feel of actual problems that can be solved using the HMLC accelerator, but it needs to be emphasized that neither the problem nor data may be shared with non-Microsoft FTEs

Root Cause Analysis





Example Datasets

All examples are public datasets, which are acknowledged in the accelerator.

Class, Level and Hierarchy

In most classification problems, there is only one level of labels

- Binary classifier
  - Covid positive, fraud, pass, …
- Multi-class classifier
  - Dog, neutral sentiment, category 6, indemnity clause …

In multi-label classification, each row has multiple labels

- Level 1 Sentiment. DS Toolkit. Say Very Good, ..., …., …, …, Very Bad
- Level 2. Feature being commented upon. Say, ease of deployment.
- Level 3. Specific attribute of the feature commented. Say, Deployment on Azure / GCP / AWS

In Hierarchical Multi-Label Classification

- There are multiple levels of classification (more than one label per row)
- With a hierarchy amongst the levels

DBPedia Dataset

	x_text	y_l1	y_l2	y_l3
0	William Alexander Massey (October 7, 1856 – Ma...	Agent	Politician	Senator
1	Lions is the sixth studio album by American ro...	Work	MusicalWork	Album
2	Pirqa (Aymara and Quechua for wall, hispaniciz...	Place	NaturalPlace	Mountain
3	Cancer Prevention Research is a biweekly peer-...	Work	PeriodicalLiterature	AcademicJournal
4	The Princeton University Chapel is located on ...	Place	Building	HistoricBuilding

DBPedia Dataset – Classification of Subject Matter

Description: This dataset contains 342k rows. In each row, the input is a text – an entry in Wikipedia – is classified in three levels. The first level categories of 'Work', 'Place' etc. have sub-categories such as 'NaturalPlace' and 'Building' in the second level, under 'Work'. The third level is a more specific category such as 'Moutain' or 'HistoricalBuilding'. Number of Levels: 3 (9, 70 and 219 classes in Levels 1, 2 and 3 respectively). Classes in level 3 are unique across Level2 categories. Therefore, this is equivalent to a single-level, multi-class classification problem.

Amazon Product Reviews

	x_productId	x_Title	x_userId	x_Helpfulness	x_Score	x_Time	x_Text	y_Cat1	y_Cat2	y_Cat3
0	B000E46LYG	Golden Valley Natural Buffalo Jerky	A3MQDNGHDJU4MK	0/0	3.0	-1	The description and photo on this product need...	grocery gourmet food	meat poultry	jerky
1	B000GRA6N8	Westing Game	unknown	0/0	5.0	860630400	This was a great book!!!! It is well thought t...	toys games	games	unknown
2	B000GRA6N8	Westing Game	unknown	0/0	5.0	883008000	I am a first year teacher, teaching 5th grade....	toys games	games	unknown
3	B000GRA6N8	Westing Game	unknown	0/0	5.0	897696000	I got the book at my bookfair at school lookin...	toys games	games	unknown
4	B00000DMDQ	I SPY A is For Jigsaw Puzzle 63pc	unknown	2/4	5.0	911865600	Hi! I'm Martine Redman and I created this puzz...	toys games	puzzles	jigsaw puzzles

Description: Based on the text of reviews in Amazon, the task is to identify three levels of hierarchical categories. Number of levels: 3 (Level 1 classes are: health personal care, toys games, beauty, pet supplies, baby products, and grocery gourmet food.) Nature of input columns: Text, primarily. Numerical and Categorical columns are also available.

Brazilian Legislation Dataset

```
1 brazil_df.groupby(['y_Tema', 'y_Sub1', 'y_Sub2', 'y_Sub3', 'y_Sub4', 'y_Sub5'], dropna=False).size()
```

executed in 38ms, finished 22:41:21 2021-10-06

```
y_Tema       y_Sub1    y_Sub2    y_Sub3  y_Sub4  y_Sub5
Assistencia  ale       exp       vtr     NaN     NaN        1
                       rhs-imt   NaN     NaN     NaN        1
                       NaN       NaN     NaN     NaN        5
             asaout    NaN       NaN     NaN     NaN        3
             doc       ppp       ges     NaN     NaN        1
                                                         ...
Prevencao    vtr       rhs-imt   NaN     NaN     NaN        2
                       san-res   NaN     NaN     NaN        1
                       vep       NaN     NaN     NaN        1
                       NaN       NaN     NaN     NaN      157
Prevenção    vis-med   NaN       NaN     NaN     NaN        1
Length: 246, dtype: int64
```

Description: This contains details of Brazilian Legislation classified into Themes (Tema) and five further sub-classifications. The text is in Portuguese. One feature of this dataset is that a large number of NaNs occur in the sub-categories

Offensive Tweets Dataset

	x_tweet	y_subtask_a	y_subtask_b	y_subtask_c
5773	#Elections News: Gun control group's political...	NOT	NaN	NaN
4511	@USER You are a big blot on the dharmic Kashmi...	OFF	TIN	IND
2446	1/ Resists newest tactic against conservatives...	OFF	UNT	NaN
9257	@USER "Yes. No one should make threats." Reall...	NOT	NaN	NaN
3052	@USER It's funny. You're claiming gun control ...	NOT	NaN	NaN

Offensive/Not-offensive Targeted / Untargeted Individual / Group / Other

Description: At the first level tweets are classified as Offensive and Not; Offensive tweets are further classified as Targeted Insults & Threats and Untargeted. In the third level, Targeted Insults & Threats are further classified into those targeted at Individual, Group or Other.

Bushveld Stratigraphic Layers Dataset

u_ICP_ppm	x_Pt_ICP_ppm	x_Pd_ICP_ppm	x_Rh_ICP_ppm	x_Ir_ICP_ppm	x_Ru_ICP_ppm	y_Stratigraphy	x_Filter	y_Level1	y_Level2
0.01	0.53	0.16	0.14	0.09	0.32	LG1	0	LG	1
0.01	1.56	0.60	0.42	0.13	0.38	LG2	0	LG	2
0.01	0.04	0.02	0.10	0.04	0.26	LG3	0	LG	3
0.01	0.10	0.02	0.07	0.04	0.34	LG4	0	LG	4
0.01	0.55	0.19	0.21	0.08	0.47	LG5	0	LG	5

Description: The Bushveld Complex (in South Africa), the largest layered mafic-ultramafic intrusion worldwide, is host of numerous, laterally continuous and chemically similar chromitite layers. Based on their stratigraphic position the layers are subdivided into a lower, middle and upper group (LG, MG and UG). Within these groups the layers are numbered successively – from the base to the top of each group. Based on the chemical composition, the requirement is to classify the layer.

Applications of HMLC

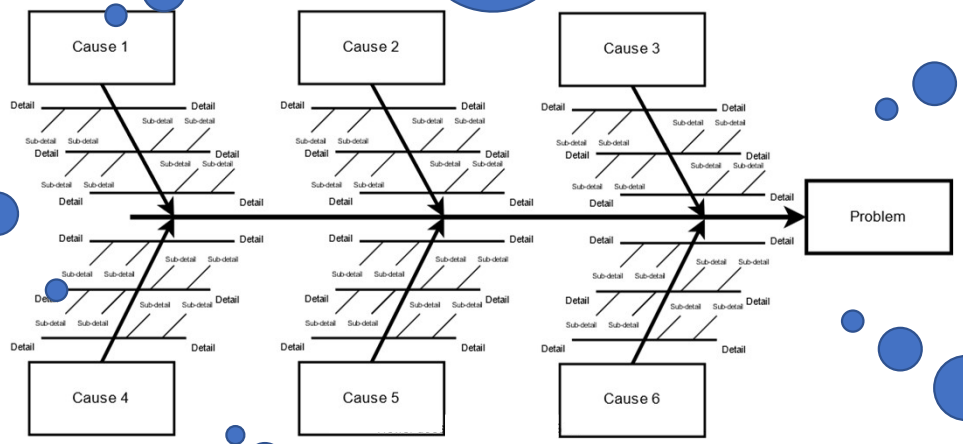
- Text classification
- Annotation of medical images
- Protein and gene prediction tasks
- Financial environment
  - RCA for budget deviations: Instead of text classification, use a general classifier. Feature engineering in finance is customer-specific. Common data model could be used to generate some features.

Complexity of the Problem

I want to classify at the lowest level, but there are hundreds of them. I don't have enough data – accuracy would be low

It would be very useful for my customer if the high-level classification is very accurate

If I classify Levels 1, 2 and 3 separately, would accuracy multiply like Bayesian probability?
0.8 * 0.7 * 0.6 = 0.336



But what approach would work best? Is the best approach different for different datasets? How do I find out?

Can I mix and match the approaches?

If I make independent models, would there be a mix-up of classes? Can I prevent that?

Approach	Model	Input	Output	Classes
Approach 1	M1	Operation	NPT Obs	5 classes
	M2	Op, Obs	NPT Cause	49 classes
	M3	Op, Obs, Cause	NPT Subcause	249 classes
Approch 2	M4	Operation	NPT Obs	5 classes
	M5	Operation	NPT Cause	49 classes
	M6	Operation	NPT Subcause	249 classes
Approach 3	M7	Operation	Obs, Cause, Subcause	303 classes
Approach 4	M8	Operation	Obs, Cause	49 classes
	M9	Op, Obs, Cause	NPT Subcause	303 classes
Approach 5	M10=M8	Operation	Obs, Cause	49 classes
	M11=M6	Operation	NPT Subcause	303 classes
Approach 6	Best of A1 to A5	…	…	150 classes
	Others-Model			153 classes

Chained models

Independent models

Powerset labels

Mix and Match Appraoch

This data snippet is shown only to give a feel of actual problems that can be solved using the HMLC accelerator, but it needs to be emphasized that neither the problem nor data may be shared with non-Microsoft FTEs

Solution: HMLC Accelerator

Solution Features

- Given a dataset, HMLC finds the best approach and the best models (Logistic Regression, Random Forest etc.)
  - Models to be used can be specified, including their hyper-parameters
  - Can be time-boxed
- Returns an artefact which acts like a trained model (though internally it is a combination of models)
  - SKLearn-like interface
  - Import and use straight away
  - Can extract model instances within the Hmlc() class
- Documented with an example notebook

Demonstration

Usage

```
from hmlc import HMLC
hmlc_obj = HMLC()
best_approach = hmlc_obj.fit(dt_train[input_col_list], dt_train[output_col_list])
```

Parameters

time_limit: float, default = 30

ngram: tuple, default = (1, 1)

stop_words: str, default = 'english'

estimators_: ['lrc', 'knn', 'dtc', 'gnb', 'mnb', 'rfc', 'abc', 'gbc', 'etc'], default = ['rfc', 'etc', 'gnb']

methods: ['independent_models', 'chained_models', 'powerset_models'], default = ['independent_models', 'chained_models', 'powerset_models']

additional_colms: list, default = []

validation_split: float, default = 0.2

max_features: int, default = 5000

token_pattern: str, default = r'([a-zA-Z0-9/+-]{1,})'

abbr_dict: dict, default = {}

Methods

predict(X):

Predict classes for X

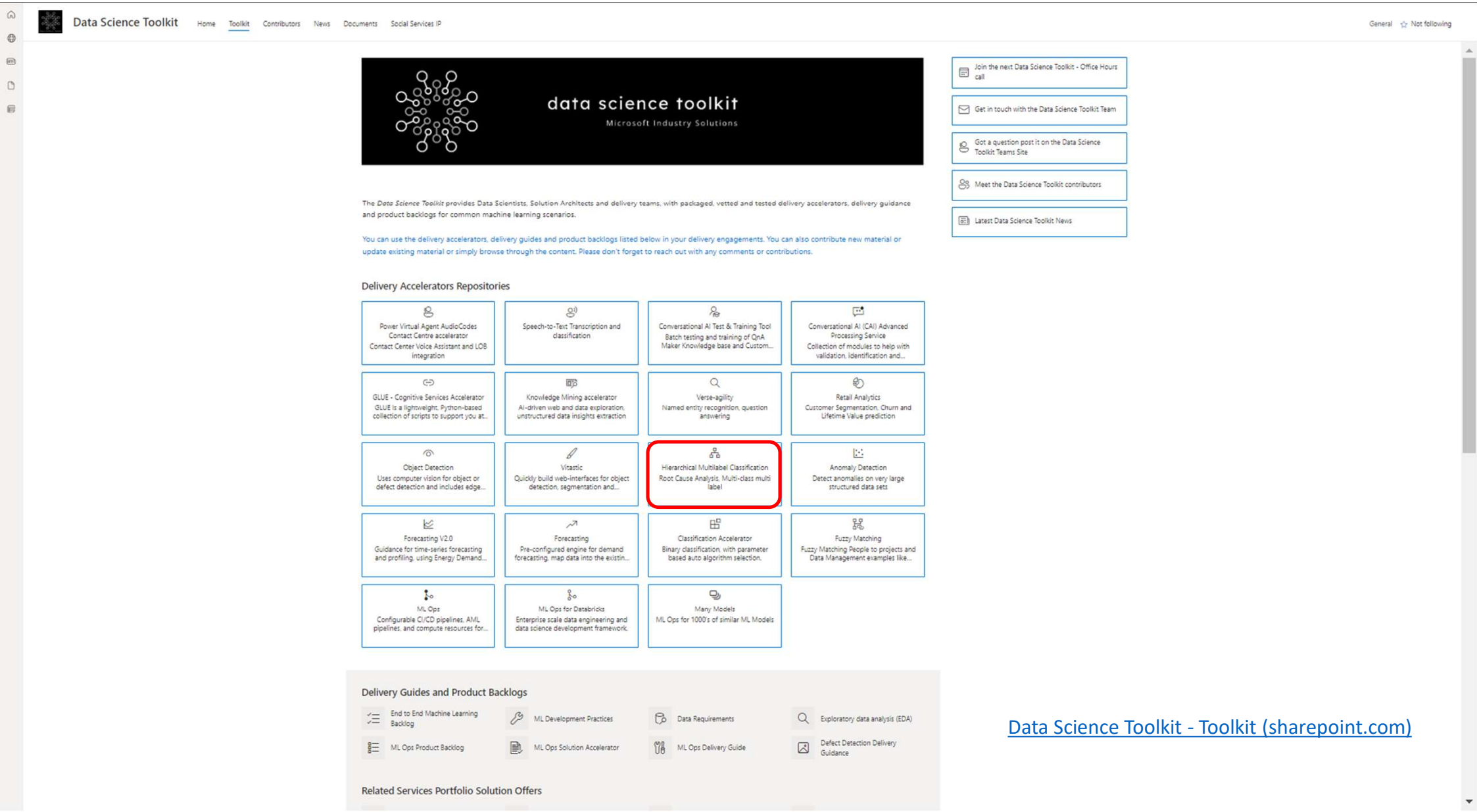
predict_proba(X):

Predict class probabilities for each class in each level, returns a nested dictionary

score(X, y):

Returns a dictionary containing accuracy and 1 – Hamming Loss for a given test data set and labels

[Hierarchical Multilabel Classification (sharepoint.com)]



[Data Science Toolkit - Toolkit (sharepoint.com)]

Send to Immersive Reader

data science toolkit

Microsoft Industry Solutions

Hierarchical Multilabel Classification

Willie Ahlers
DIR BPM MCS

Access the Accelerator

Hierarchical Multilabel Classification
https://github.com/microsoft/dstoolkit-hierarchical-multilabel-classification

Related Services Portfolio Offer

End-to-End Machine Learning

Industries

- Finance Operations
- Oil and Gas
- Healthcare
- Manufacturing

Use-case

- Root-cause analysis for finance budget deviations or forecast deviations
- Analyses of non-productive drilling platforms
- Multiclass, multilabel, text classification
- Image annotation
- Bio-informatics

Accelerator description

The Hierarchical Multi-label classification accelerator can be applied to a variety of use-cases where classes are hierarchically structured, and object can be assigned to multiple paths of the class hierarchy at the same time. Applications include text classification, image annotation, and in bioinformatics problems such as a protein function prediction.

Text classification problems can be described as, when given a text, classify the text into one of the labels for example:

- Website articles - > Sports, Politics, etc.
- Clauses in a contract - > Termination, Indemnity, etc.
- Sentiment analysis etc.

When more than one label is to be provided, it becomes multi-label classification. A single label, multi-class has one label, but more than 2 classes. Multi-label classification has multiple labels provided, each could have 2 or more classes.

Example.

- Level 1 Sentiment. DS Toolkit. Say Very Good, .., ..., .., ..., Very Bad
- Level 2. Feature being commented upon. Say, ease of deployment.
- Level 3. Specific attribute of the feature commented. Say, Deployment on Azure / GCP / AWS

Hierarchical multilabel classification can therefore be applied to:

- Determine root cause analysis of budget variations in Finance.
- Determining the most profitable product configurations for industrial manufacturing equipment.
- Classification of subject matter
- Product classification based on reviews.
- Offensiveness classification of tweets.
- Classification of legislation into themes
- Classification of geo-chemical layers of a region.

Contributors

Senani Nori
SENIOR DELIVERY DATA SCIENTIST

Tarun Dugar
ASSC CONSULTANT

[Hierarchical Multilabel Classification (sharepoint.com)]

Accelerator guidance

Related Accelerators

Discussion