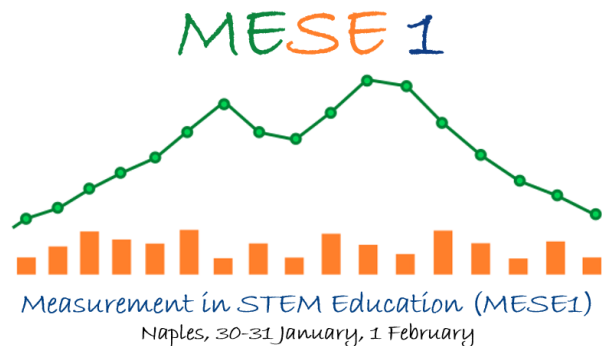


Automated corrections of open ended question of INVALSI tests

Michele Marsili (INVALSI)



WHAT IS

- ✓ INVALSI - Italian: National Institute for the Evaluation of Education Systems

MAIN TASKS

- ✓ Carry out periodic and systematic evaluation on the knowledge and skills of the students and on the educational system
- ✓ Study the causes of school failure and early school leaving with reference to the social context and the types of educational offer
- ✓ Try to measure the added value achieved by the schools
- ✓ Ensure Italian participation in European and international research projects (PISA, TIMSS...)
- ✓ Provide support to schools, regions and local authorities

WHY THE TESTS ARE DONE?



School system have a duty not to allow the existence of “A league” and “B league” areas, schools, or classes



Using the same tests for everyone helps us to identify areas that need improvement



Measure the learning outcomes of some key competences



Measure some basics of critical thinking: ability to understand texts, logical faculties, and the ability to solve new problems

The skills by the National Guidelines

ITALIAN

The tests on Italian measure the ability to understand written texts taken from literature, non-fiction or everyday life. This understanding, examined using closed questions (with a choice of predefined answers) or open questions, relates to the nature of the text, any explanations, the meaning of different passages or specific expressions, or the author's intention

MATHEMATICS

Mathematics tests measure the ability to use mathematical knowledge to solve problems, real or otherwise, logical skills, interpretation of graphs, interpretation of phenomena with a quantitative dimension, modelling, or use in various scientific disciplines

LISTENING

English Listening tests measure the ability to understand listening passages

READING

English Reading tests Measure the ability to understand written texts



**Policy
Makers**



School Staff



**Scientific
Community**



**Parents,
Journalists...**



Paper-based – Primary Schools:

Every pupil of the same grade get the same set of questions in the paper tests



Computer-based – Lower and Upper Secondary Schools:

The computer tests present them in different sets of items that are equivalent in terms of the skill measured and the difficulty level, taken from an ad-hoc “item bank”

COMPUTER-BASED TESTS



With computer-based tests, the measurement is more accurate and is presented in the form of the level reached for each skill



Teachers the laborious task of marking and data entry, with an annual saving of some 22 truck-loads of paper, formerly needed to distribute around 2.5 million test booklets



Statistical analysis of this “big data” makes it possible to calculate the probability with which a pupil would correctly answer other items that appear in the item bank

HOW THE TESTS ARE GENERATED

**Teachers
School principals
Researchers**

Creation of questions

The items proposed by the authors are examined by groups working collaboratively. Each item must correspond to a specific skill pursuant to the Reference Framework for the educational level examined

**Students
Teachers
School principals
INVALSI's researchers**

Pre-test

To check all these requirements, the booklet is pre-tested on a few thousand pupils. The pre-tests involve about 30,000 children per year, which is more than enough for a statistical analysis of the results that enables INVALSI researchers to identify any remaining problems

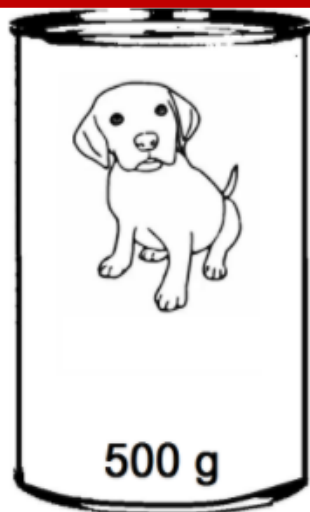
**Teachers
School principals
Researchers**

Final booklet

Authors and experts work together until the final phase. By this time, two years have passed since the work was begun, and the booklet finally becomes the actual booklet made up of the items that have "survived" all these checks

Domanda

Filippo, per il suo bassotto, compra sempre scatole di cibo per cani da 500 grammi, come quella mostrata in figura.



Ogni giorno il bassotto mangia 200 grammi di cibo per cani. Filippo conserva ogni scatola aperta finché non l'ha completamente svuotata. Filippo oggi non ha più scatole di cibo per cani e quindi deve comprarle. Quante scatole almeno dovrà comprare se vuole che gli bastino per una settimana?

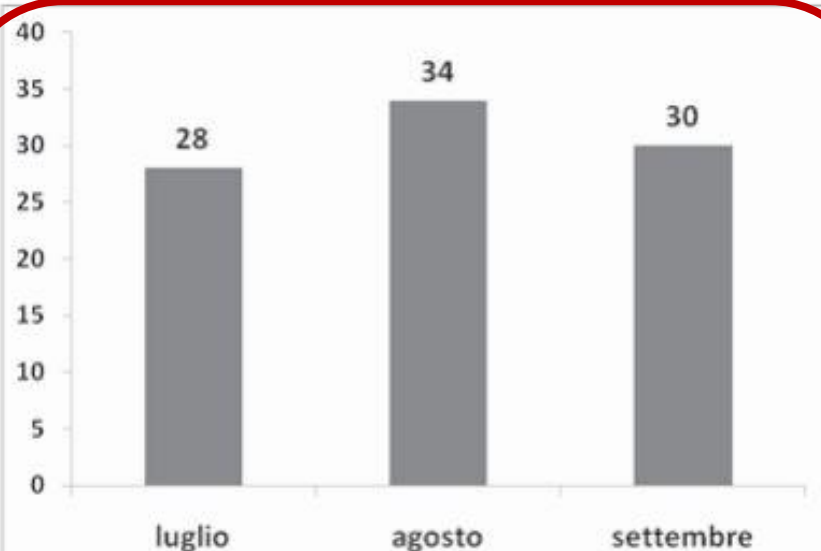
Scrivi come hai fatto per trovare la risposta e poi riporta sotto il risultato.

Digita il procedimento qui sotto.

Risultato: scatole

PURPOSE: To solve a PROPORTIONALITY PROBLEM AND INTERPRET THE RESULT IN THE CONTEXT OF THE PROBLEM

D10. Il grafico riporta il numero di *e-book reader* (lettori di libri elettronici) venduti nei mesi di luglio, agosto e settembre da un negozio di informatica. Negli altri nove mesi dell'anno lo stesso negozio ha venduto in media 18 *e-book reader* al mese.



Qual è il numero medio mensile di *e-book reader* venduti in quell'anno dal negozio?

- A. ☐ Circa 31
- B. ☐ Circa 28
- C. ☐ Circa 21
- D. ☐ Circa 24

PURPOSE: TO CALCULATE THE WEIGHTED MEAN, USING THE INFORMATION PRESENT IN A BAR GRAPH.

Manual coding VS ASAG



CBT introduces the important task of grading the short answers to the open questions of the test

Manual correction

**MORE accuracy
MORE time consuming**



**Degree of
automation**



ASAG

**LESS accuracy
LESS time consuming**

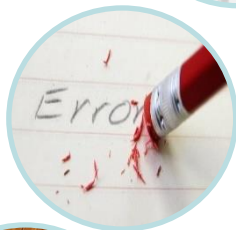
The automatic supervised correction procedure



Before the test administration period



The correction team and items authors group discuss to define the correction criteria



The correction criteria have been translated in logical patterns



Check the behavior of the correction criteria on pupils' answers to pre-test questions



Planning of control activities and meetings with the authors during the test administration period

Correction procedure



**Acquisition
of the
answers'
database**

**Pre-
processing
operations**

**Answers
classification
according to
the fixed
correction
criteria**

**Production
of reports**



Pre-processing is a generic term used for the different activities that you undertake to get your texts ready to be analysed

PRE-PROCESSING TECHNIQUES

- ✓ Converting your text to lower case
- ✓ Punctuation and non-alphanumeric character removal
- ✓ Stopwords removal
- ✓ Tokenisation
- ✓ Parts of speech tagging
- ✓ Stemming and lemmatization
- ✓ Automated correction of spelling and typing errors with identification and replacement of words «out of vocabulary» (OOV)

Pre-processing- Stop Words

Stop words are a set of commonly used words in a language.

Examples of stop words in English are “a”, “the”, “is”, “are” and etc.

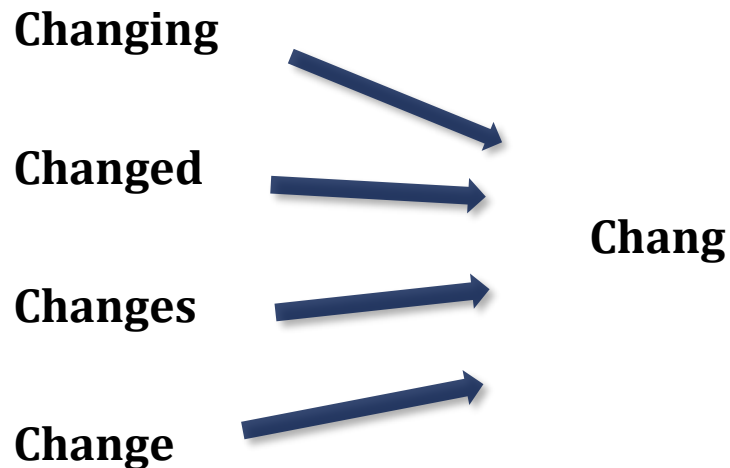
For example

“this is a stop word”  **“stop word”**

Pre-processing- Stemming

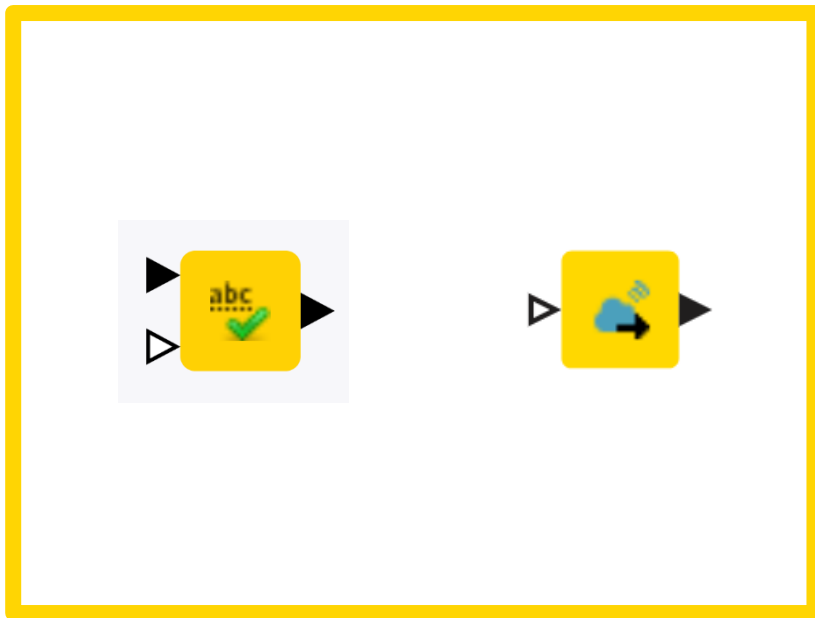
Stemming is the process of reducing inflected words to their word stem or root form.

For example



Pre-processing- Automated correction

KNIME NODES + SPELL CHECK API



```
"flaggedTokens": [  
  {  
    "offset": 5,  
    "token": "Az ure",  
    "type": "UnknownToken",  
    "suggestions": [  
      {  
        "suggestion": "Azure",  
        "score": "1"  
      }  
    ]  
  }  
],  
"_type": "SpellCheck"  
}
```

Answer classification

Dialog - 23:2 - String Manipulation

File

String Manipulation Flow Variables Memory Policy

Column List

ROWID

ROWINDEX

WEIGHT

S risposta_studente

S risposta_post_DATA_CLEANING

Category

All

Function

padLeft(str, size)

padLeft(str, size, chars)

padRight(str, size)

padRight(str, size, chars)

regexMatcher(str, regex)

regexReplace(str, regex, replaceStr)

removeChars(str)

removeChars(str, chars)

removeDiacritic(str)

removeDuplicates(str)

replace(str, search, replace)

replace(str, search, replace, modifiers)

replaceChars(str, chars, replace)

replaceChars(str, chars, replace, modifiers)

replaceUmlauts(str, omitE)

reverse(str)

string(x)

strip(str...)

stripEnd(str...)

stripStart(str...)

substr(str, start)

substr(str, start, length)

toBoolean(x)

toDouble(x)

toEmpty(str...)

toInt(x)

toLong(x)

toNull(str...)

Variable List

S condizione_codifica

1 regexMatcher(risposta_post_DATA_CLEANING,\$\${Scondizione_codifica}\$\$)





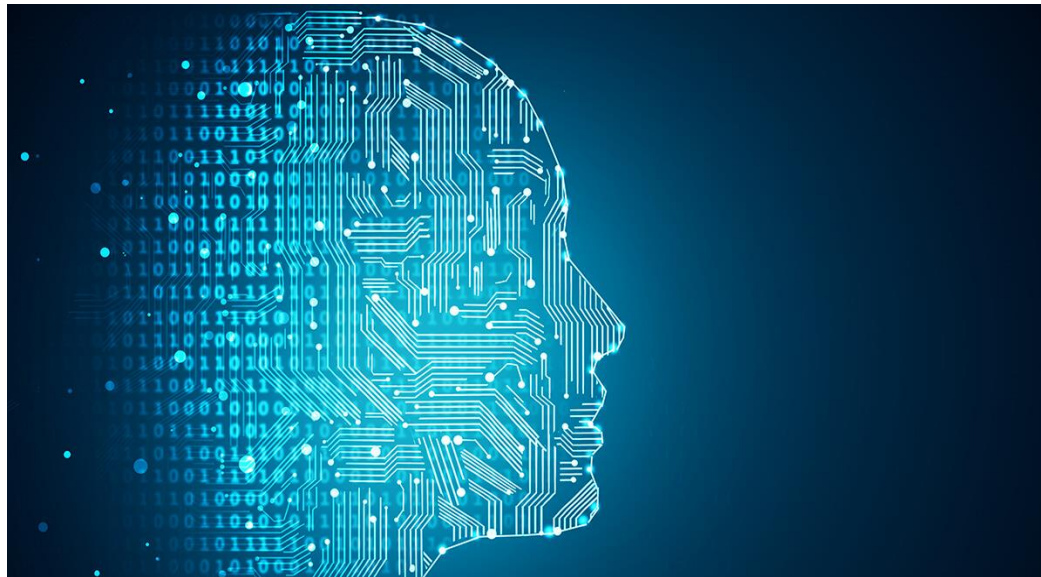
The report provides a clear picture of how the algorithm is recoding

For each item is calculated the frequency distribution of correct/incorrect answers

Each report is given to authors group to have feedback useful to modification of correction conditions

Reporting and feedback allow the correction team to update the correction conditions and the RegEx. This improve the precision and the accuracy of the correction conditions that evolve during the assesment period

The totally automatic correction procedure : a Machine Learning approach



ASAG – Automatic Short Answer Grading



The goal is the automatic evaluation of «short» answers in natural language



The trend for ASAG is Machine Learning

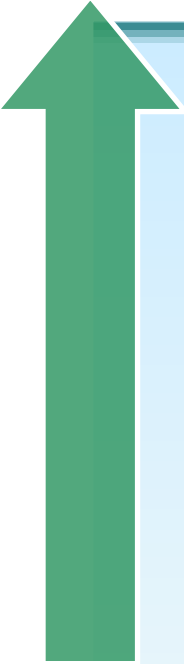


The «computer based» automatic grade improves the correction process



Faster feedback to both students and teachers

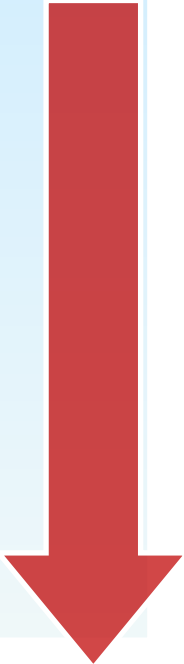
ASAG – METHODOLOGICAL APPROACH



METHODOLOGICAL APPROACH:

- Expert opinion
- Analyze the nature of dataset
- Data cleaning techniques
- Numerical statistics that are intended to reflect how important a word is to an answer (es. TF)
- Selection of best machine learning algorithm
- Model evaluation

**CLASSIFICATION
ERRORS
(FALSE NEGATIVE +
FALSE POSITIVE)**



Two kind of open answers



SHORT ANSWERS:

- The average length of answers usually does not exceed 10 words
- Words in the answer contained within the text of the item (e.g. reading comprehension questions)
- The lexical content of the answer is relevant



LONG ANSWERS:

- The average length of answers usually does exceed 10 words
- Words in the answer NOT contained within the text of the item (e.g. Request to explain how the answer to a mathematical problem was arrived at)
- The semantic content of the answer is relevant

From modeling to Scoring

Create a vector for each answer:

The size of the vector will be the number of distinct terms in the BoW

Classification algorithms:

The training answers contribute to the creation of the model that will be applied on the test answers

Bag of words

Document to vector

Partition

Machine learning

Model evaluation

Bag-of-words model:

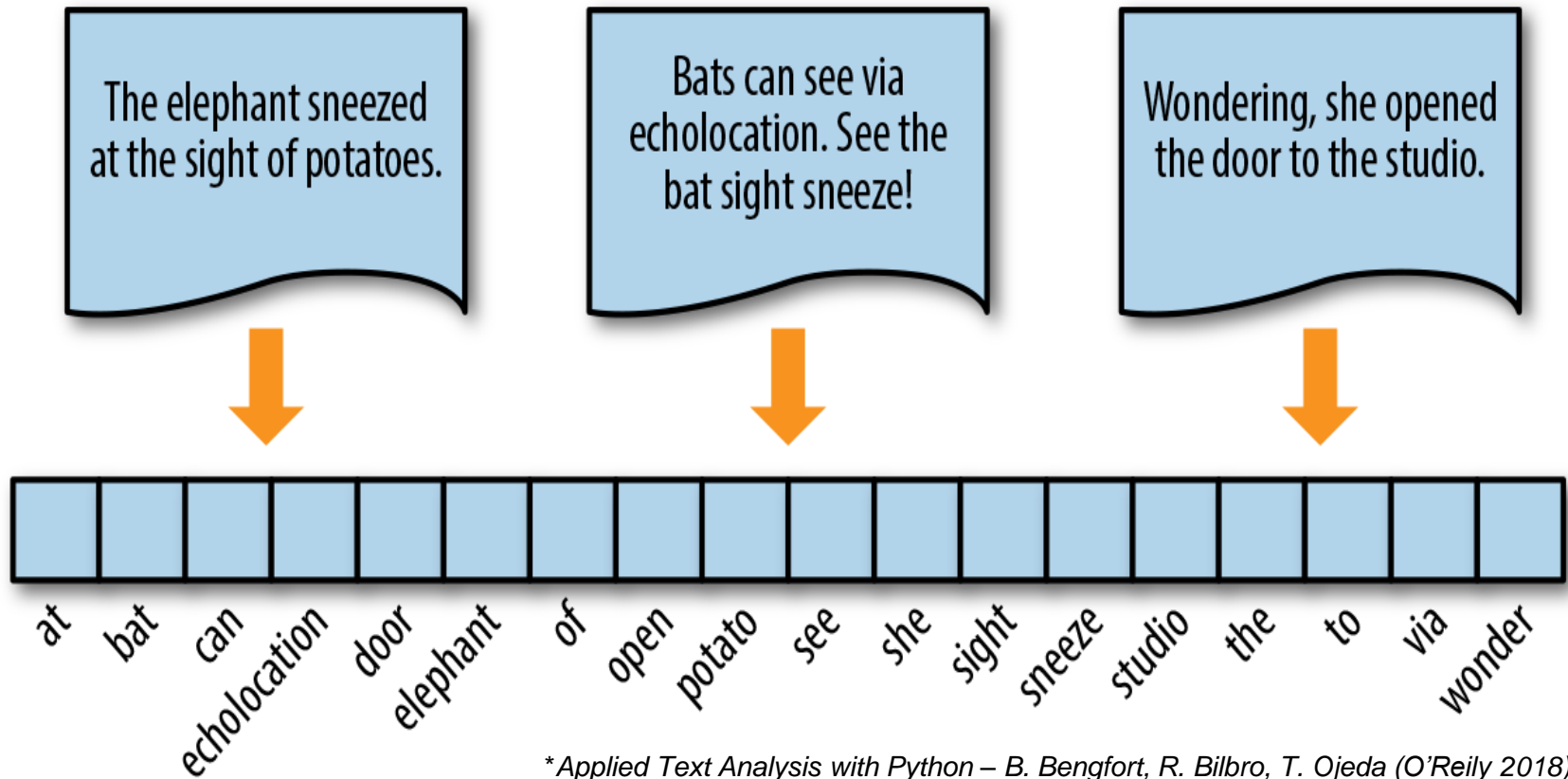
The BoW consists of two columns, one with the answer and one with the terms that occur in the corresponding answer

Partition of dataset:

The dataset is divided into training set and test set

Classification Model Evaluation:

Model evaluation metrics

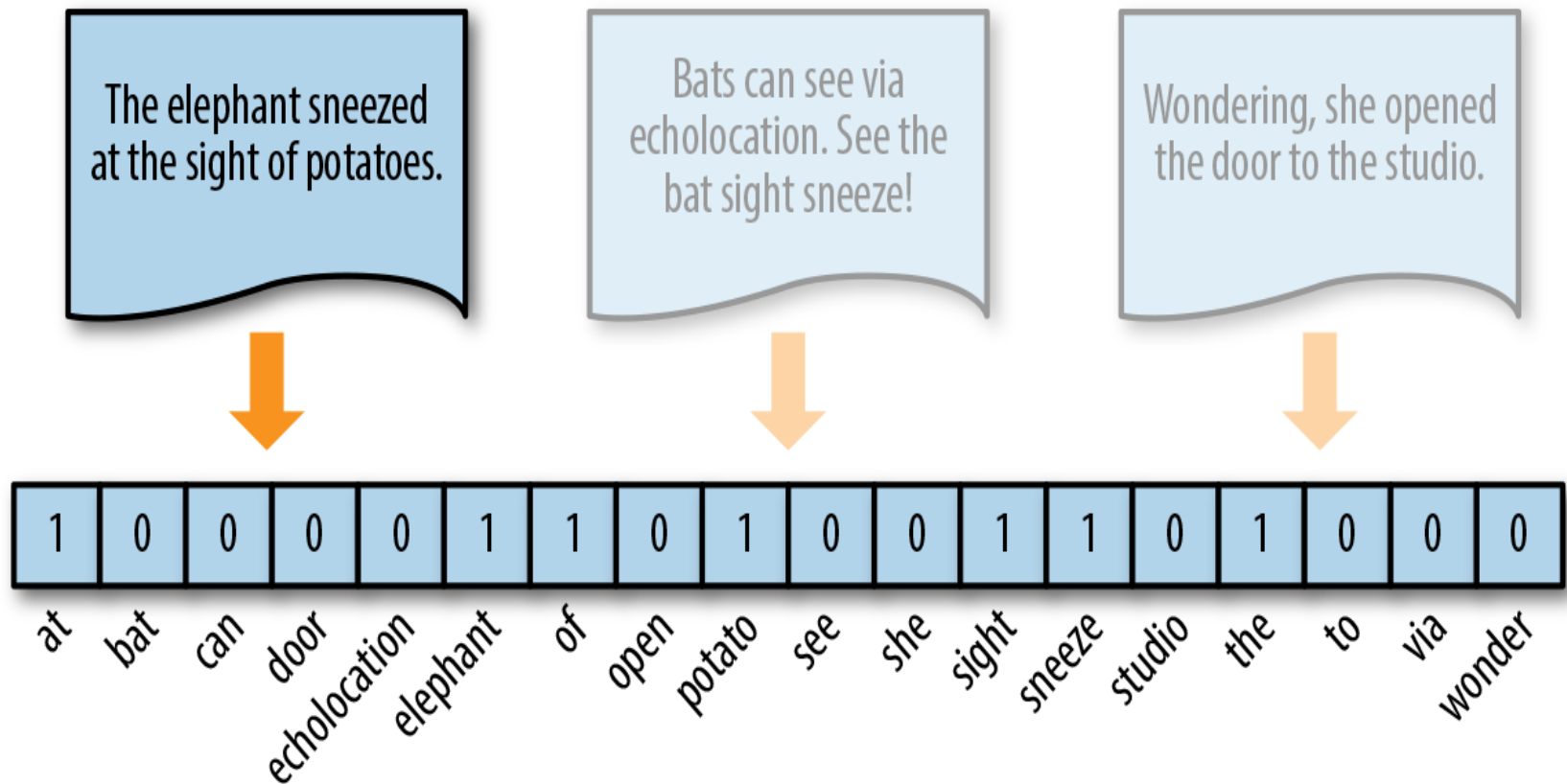


Text Vectorization - Vector encoding model

The main vector encoding models are:

- ✓ **Frequency Vectors:** fill in the vector with the frequency of each word as it appears in the document
- ✓ **One-Hot Encoding:** boolean vector encoding method that marks a particular vector index with a value of true (1) if the token exists in the document and false (0) if it does not
- ✓ **Term Frequency-Inverse Document Frequency:** consider the relative frequency or rareness of tokens in the document against their frequency in other documents
- ✓ **Distributed Representation:** encode the similarities between documents in the context of that same vector space

Text Vectorization -One-Hot Encoding



What is machine learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.**

There is a lot of data:
pictures, music, words,
videos etc.



The volume of data is so high that we will increasingly turn to automated systems that can **learn from the data and make better decisions** in the future, based on the examples that we provide.

How Machine Learning works

Input data is processed in order to obtain structured data, on which a machine learning algorithm is trained.

Input data → Processed data → Machine Learning Algorithms → Trained model

The trained model is applied on new data to make predictions

New data → New Processed data → Trained model → Prediction

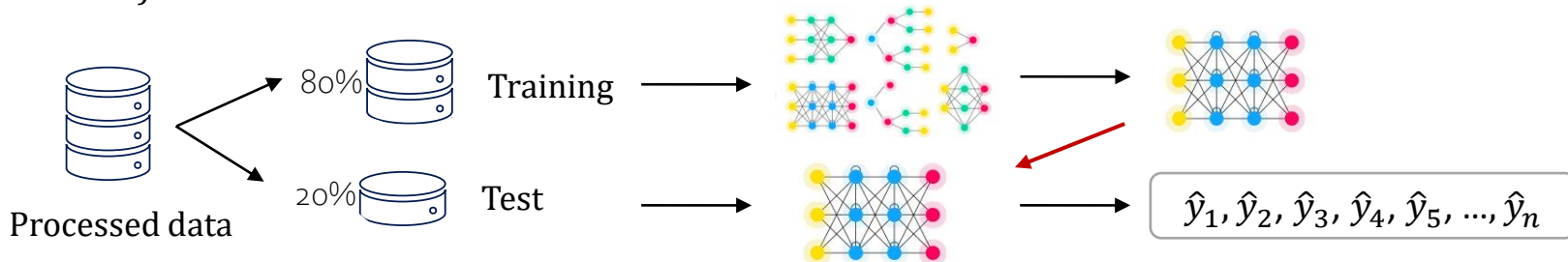
Supervised machine learning algorithms

(X and Y are given)

1. The processed data is divided into training and test.
2. Top performing parameters are determined on the training set, in order to build the best model for that data.
3. The trained model is used to perform predictions on the test set.

Multiple statistical metrics can be used to assess the performance of Machine Learning algorithms.

The final validated model is saved and used to perform predictions on new input data (until a new model is trained).



ASAG – Machine learning algorithm

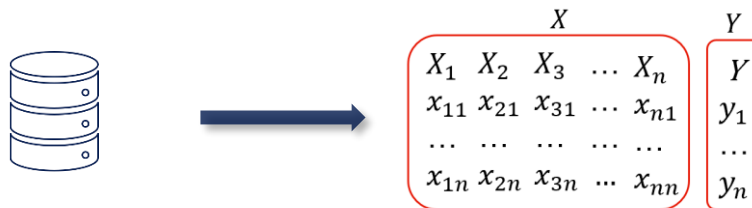
The algorithm is build upon the test' answers and the correspondent classification (TRUE/FALSE): the goal is to predict which classification should be assigned to a new answer

The most common Machine learning algorithms for classification problem are:

- **Support Vector Machine**
- **Random Forest**
- **Naive Bayes**
- **K-Nearest Neighbors**

The goal of using a Decision Tree is to create a training model that can predict the target variable by learning simple decision rules inferred from prior data.

Decision tree algorithms belongs to the family of **Supervised Learning algorithms**.



Problems that Decision Tree can solve:

- **Classification:** a classification tree will determine a set of logical if-then conditions to classify the target variable that is categorical.
- **Regression:** a regression tree is used when the target variable is numerical or continuous. A set of conditions based on the sum of squared errors are used to make the prediction.

How Decision tree works

Suppose we want to predict if the following Titanic passenger survived or died.









Mr. William Henry

Sex : Male

Age: 45 years

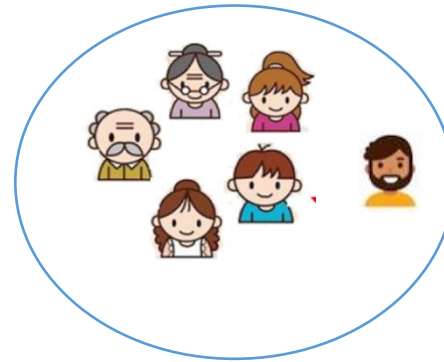
We have an initial dataset that we used to build the model. For which we know all the information:

	Sex	Age	Survived /died
	M	39	<u>died</u>
	F	10	survived
	M	20	survived
	M	72	<u>died</u>
	F	63	survived
	F	47	survived



How Decision tree works

We have an initial dataset:



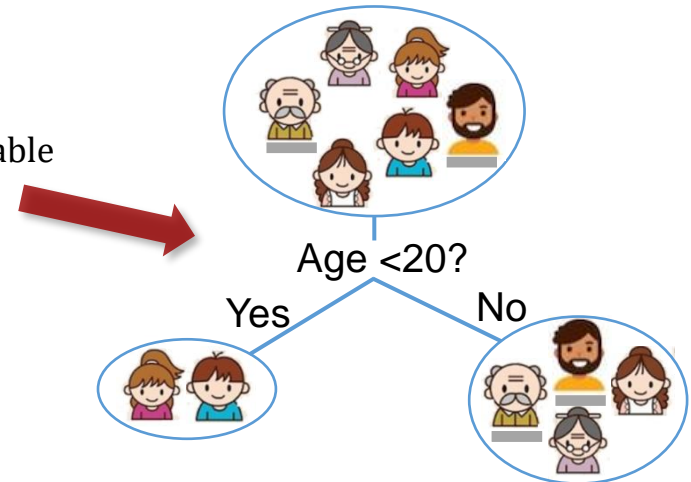
How Decision tree works

We have an initial dataset:

1. To give more information about the prediction of the target variable a decision rule is used to split the data into two subgroups:



The resultant sub-nodes are more homogeneous



How Decision tree works

We have an initial dataset:

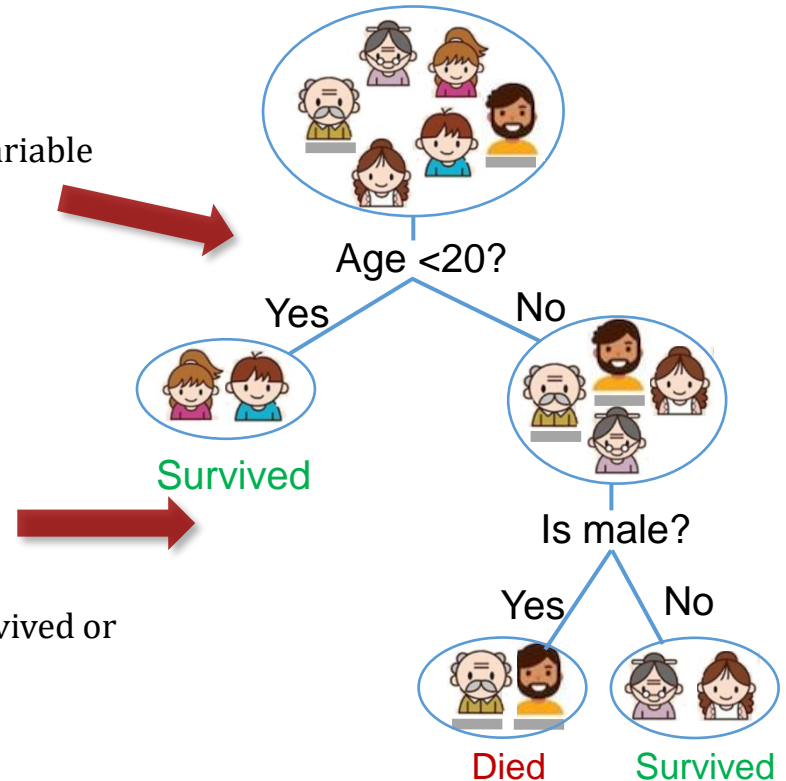
1. To give more information about the prediction of the target variable a decision rule is used to split the data into two subgroups:



The resultant sub-nodes are more homogeneous

2. We can also use the Sex variable to make another split:

Thanks to this grouping in predicting whether an individual survived or died we will make a smaller error.



How Decision tree works

In this way we can use this decision tree to predict our interest unit.



Mr. William

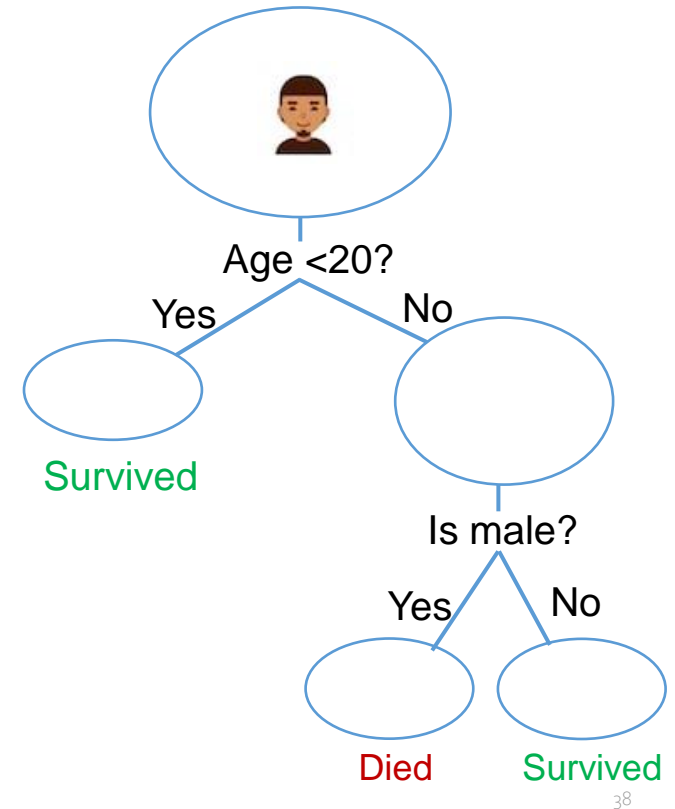
Henry

Sex : Male

Age: 45 years



Did William Henry survive?



38

How Decision tree works

In this way we can use this decision tree to predict our interest unit.



Mr. William

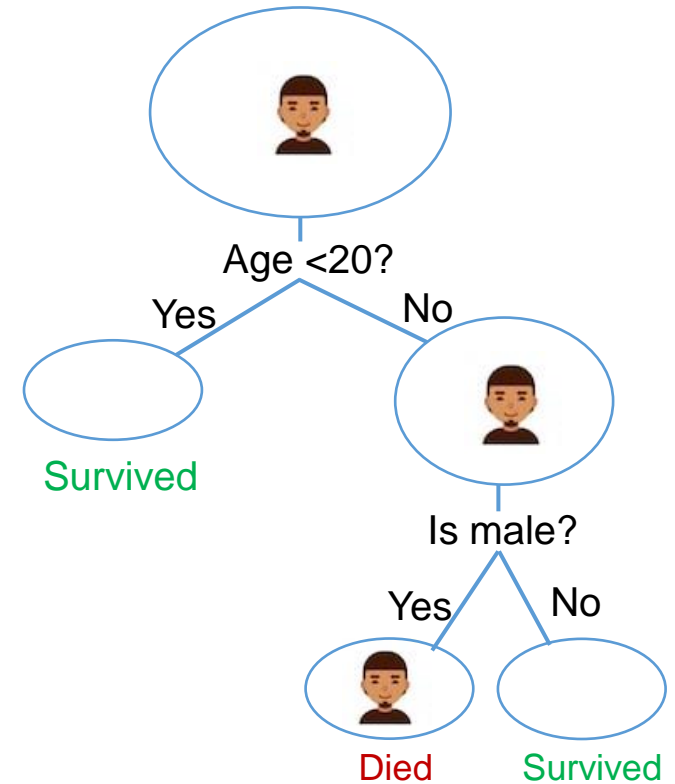
Henry

Sex : Male

Age: 45 years



Did William Henry survive? **NO**

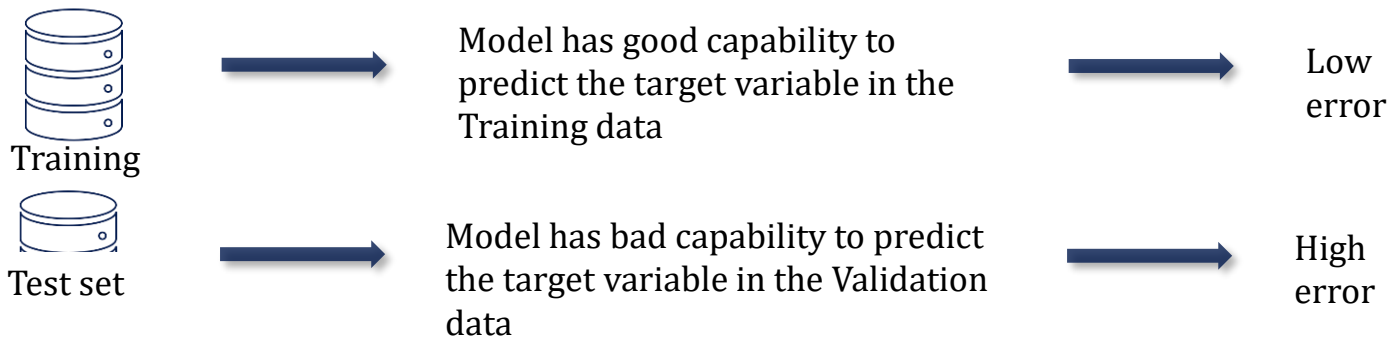


Overfitting problem

The parameters are optimized to obtain the model that best fits the data structure.

Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points.

Overfitting the model generally takes the form of making an overly complex model to explain the structure of the data.



Beyond a single tree

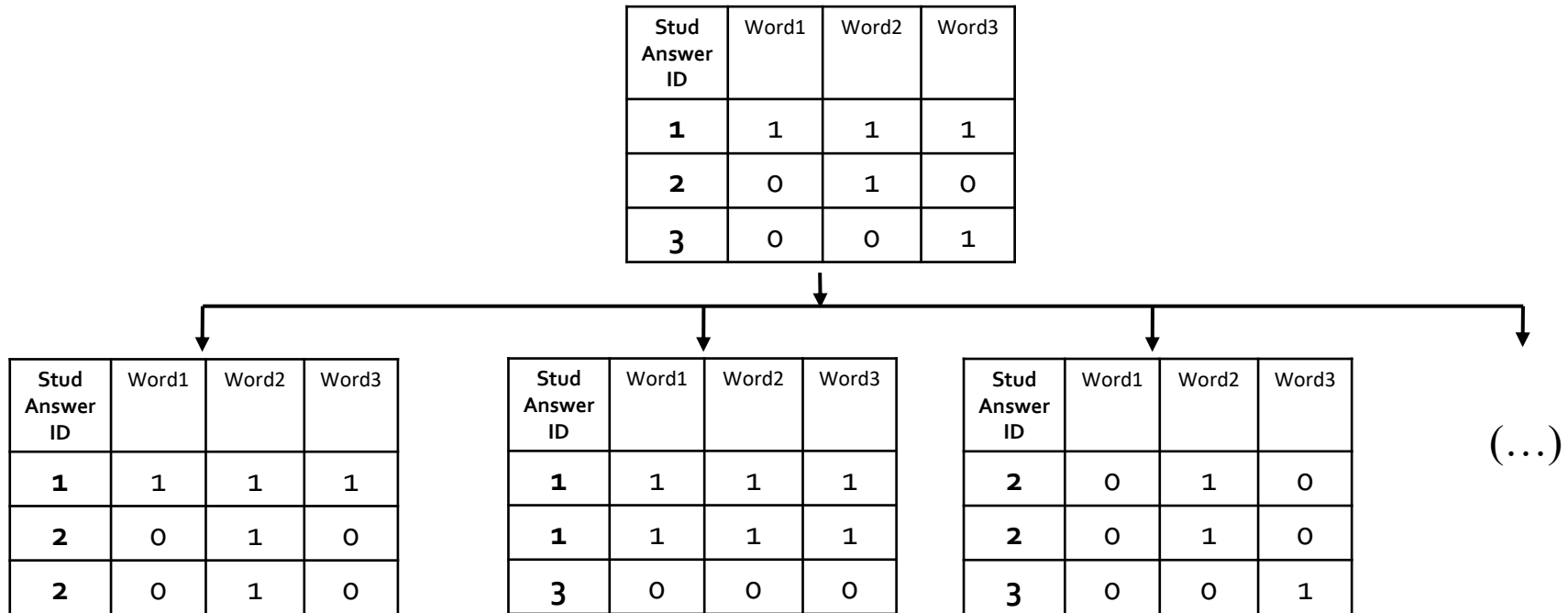
- Tree-based methods are simple and useful for interpretation
- They are not competitive with the best supervised learning approaches in terms of prediction accuracy



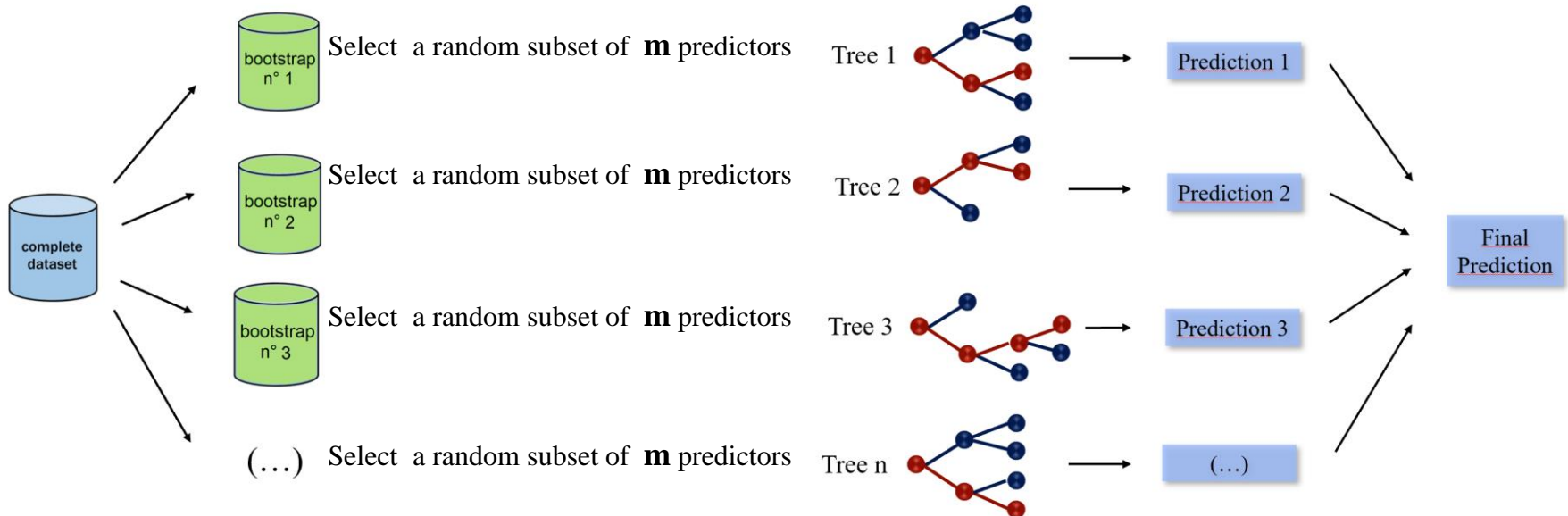
Random Forest

This method grow multiple trees which are then combined to yield a single prediction, reducing the variance and increasing the prediction accuracy

The bootstrap method involves iteratively, resampling a dataset with replacement. Obtaining **n-sub-samples** with equal sample size of the initial dataset.

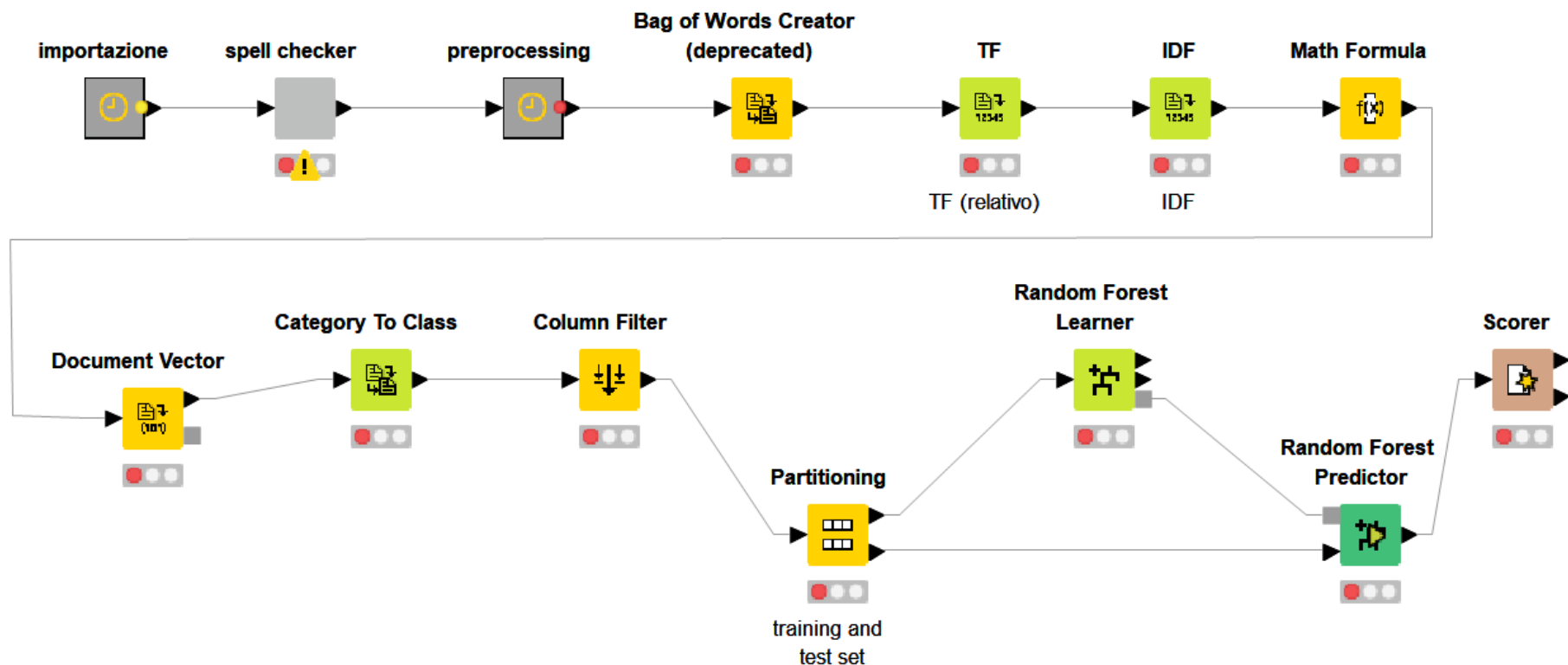


Random Forest



Decorrelated trees are created so the average of the resulting trees is less variable and hence more reliable

ASAG – KNIME Workflow Random Forest



Accuracy is one of the metrics that can be computed starting from a confusion matrix.

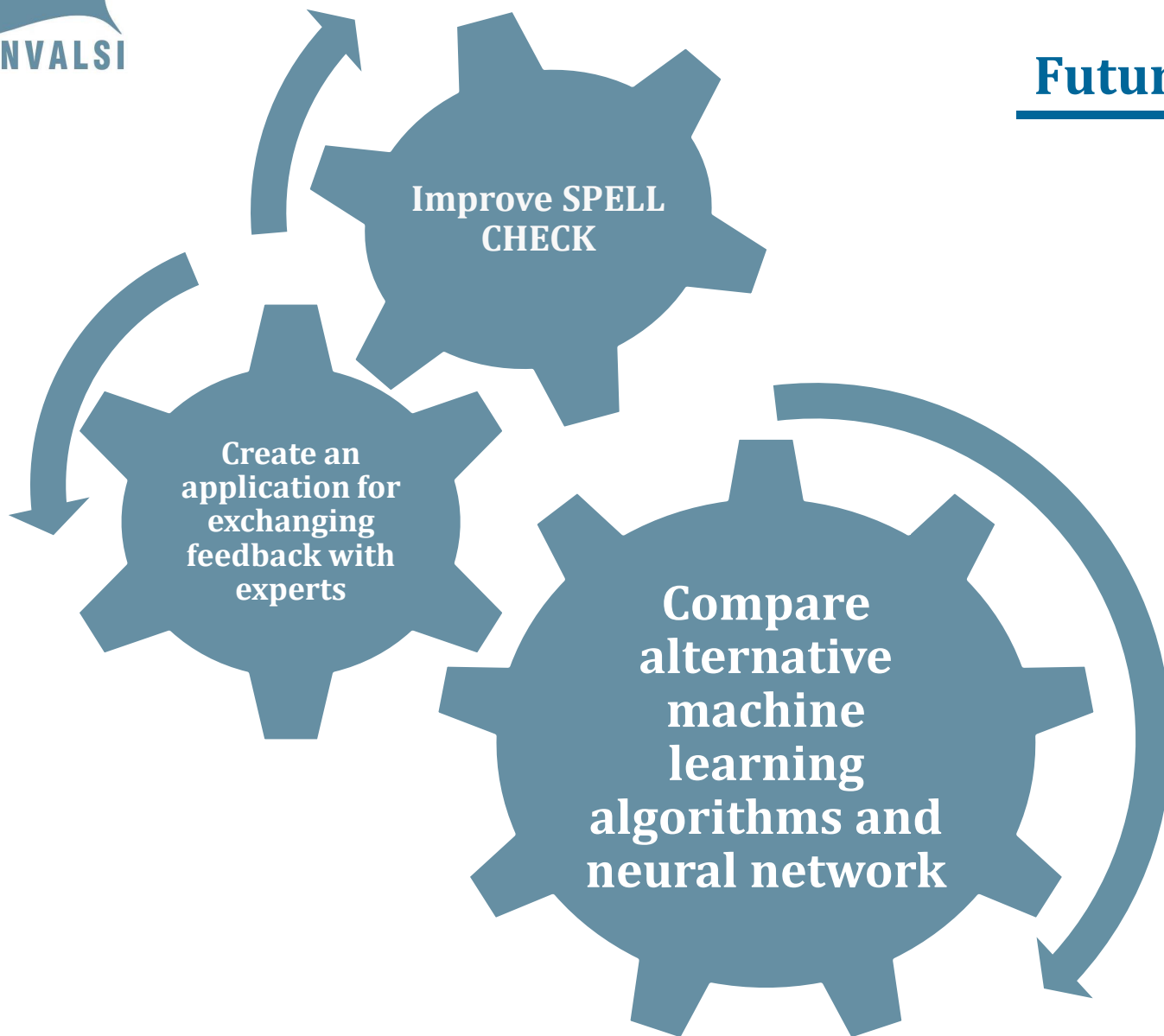
In classification problems accuracy is defined as the number of correct predictions made by the model over all considered cases.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})}$$

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

	Short answer	Long answer
	N=1234	N=3139
Accuracy	0.983	0.867
Kappa di Cohen	0.962	0.729
Error (%)	1.736	13.331

Future development



Thanks for your attention!



michele.marsili@invalsi.it



<https://invalsi-serviziostatistico.cineca.it/>



<https://www.invalsiopen.it/>



<https://www.facebook.com/Servizio-Statistico-INVALSI>