UNIVERSITÁ DEGLI STUDI DI MODENA E REGGIO EMILIA
Dipartimento di Ingegneria «Enzo Ferrari»
Corso di Laurea in Ingegneria Informatica

DATA PREPARATION E WORKFLOW MANAGEMENT CON KNIME

Relatore:
Prof.ssa Sonia bergamaschi
Correlatore:
Dott. Luca Zecchini

Laureanda:
Cristina Ventilati

ANNO ACCADEMICO 2020-2021

Data preparation

1

Raccolta ed esplorazione dei dati

Data discovery

2

Validazione dei dati

Data validation

3

Strutturazione dei dati

Data structuring

4

Arricchimento dei dati

Data enrichment

5

Filtraggio dei dati

Data filtering

6

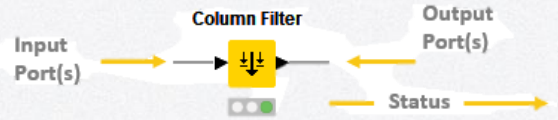
Pulizia dei dati

Data cleaning

Open for Innovation

KNIME

- Software gratuito e open source
- Nato all'università di Costanza nel 2006
- Progettato per processi di analisi dei dati e data mining
- Tutte le attività vengono rappresentate dai nodi che sono le unità base di computazione



Data preparation con KNIME

Articolo di riferimento:

"Data Preparation: A Survey of Commercial Tools", scritto da Mazhar Hameed e Felix Naumann

KNIME:

Si verifica se e come KNIME sia in grado di eseguire le varie funzionalità

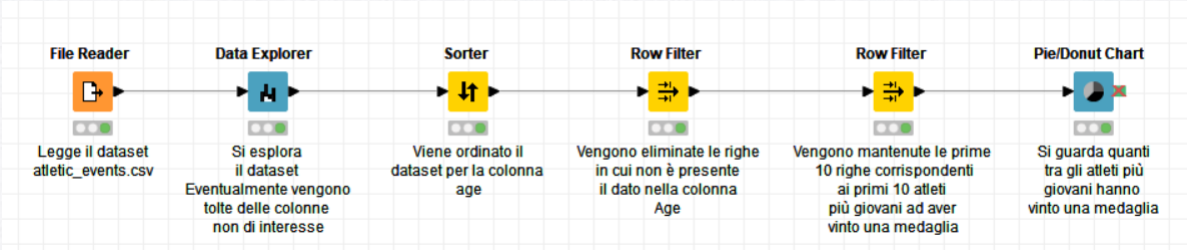
Categories	Available features
Data discovery	Locate missing values (nulls)
	Locate outliers
	Search by pattern
	Sort data
Data validation	Compare values (selection and join)
	Check data range
	Check permitted characters
	Check column uniqueness
	Find type-mismatched data
	Find data-mismatched datatypes
Data structuring	Change column data type
	Delete column
	Detect & change encoding
	Pivot / unpivot
	Rename column
	Split column
	Transform by example [13]
Data enrichment	Assign semantic data type
	Calculate column using expressions
	Discover & merge external data
	Duplicate column
	Generate primary key column
	Join & union
	Merge columns
	Normalize numeric values
Data filtering	Delete/keep filtered rows
	Delete empty and invalid rows
	Extract value parts
	Filter with regular expressions
Data cleaning	Change date & time format
	Change letter case
	Change number format
	Deduplicate data
	Delete by pattern
	Edit & replace cell data
	Fill empty cells
	Remove extra whitespace
	Remove diacritics
	Standardize strings by pattern
	Standardize values in clusters

Workflow management con KNIME

KNIME mette a disposizione apposite sezioni di nodi dedicati alla visualizzazione dei dati e al controllo di flusso.

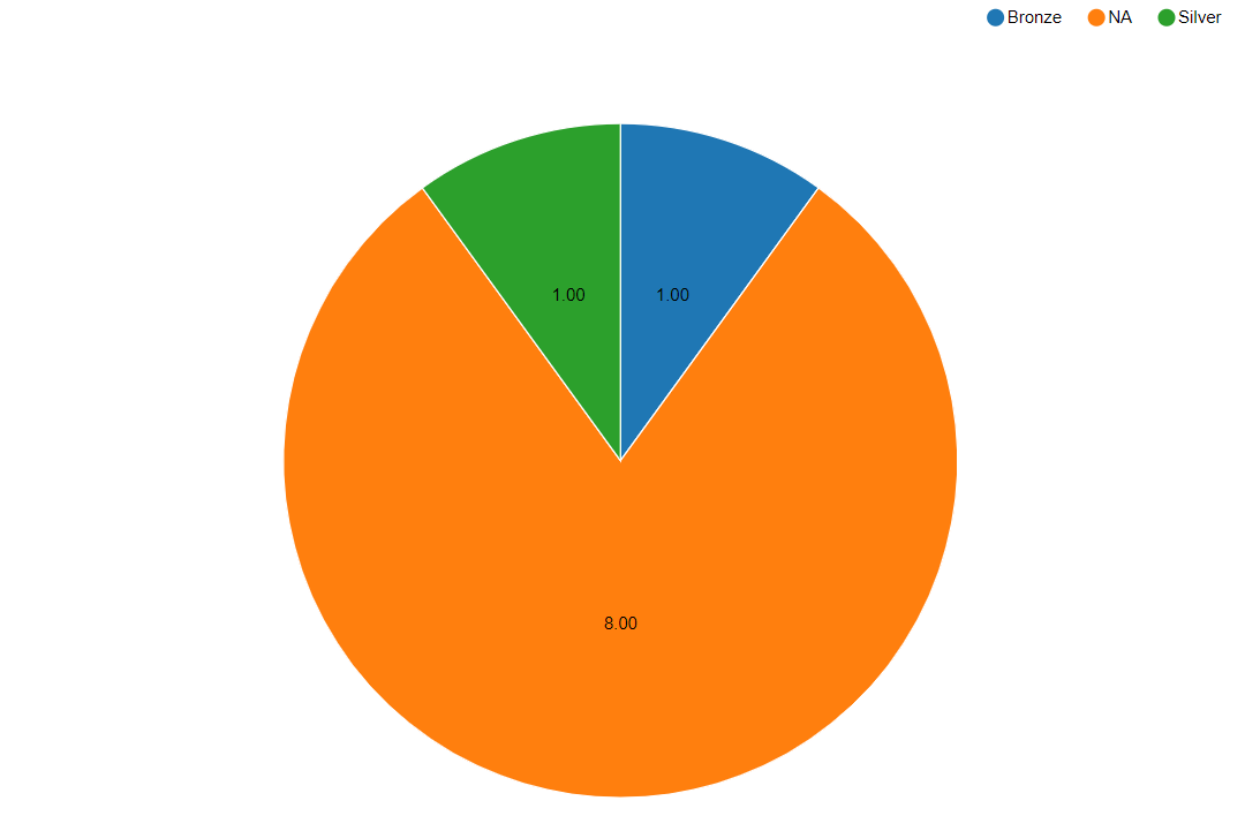
- **Data visualization**
  - Histogram
  - Scatter plot
  - Data Explorer
  - Pie/Donut Chart
  - Sunburst Chart
- **Workflow control**
  - Loop
  - Switches
  - Try-Catch

Chi sono i 10 atleti più giovani ad aver mai partecipato ai Giochi Olimpici?

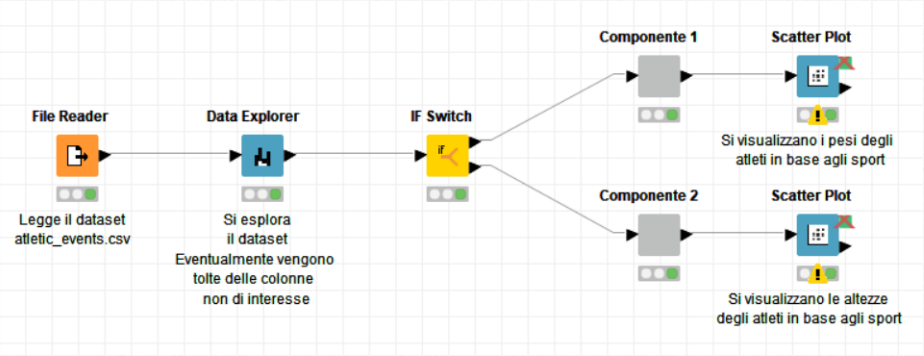


Row ID	Name	Sex	Age	Team	NOC	Games	Year	Season	City	Sport	Event
Row142882	Dimitrios Loundras	M	10	Ethnikos Gymnastikos Syllogos	GRE	1896 Summer	1896	Summer	Athina	Gymnastics	Gymnastics Men's Parallel Bars, Tea
Row43468	Magdalena Cecilia Colledge	F	11	Great Britain	GBR	1932 Winter	1932	Winter	Lake Placid	Figure Skating	Figure Skating Women's Singles
Row73461	Carlos Bienvenido Front Barrera	M	11	Spain	ESP	1992 Summer	1992	Summer	Barcelona	Rowing	Rowing Men's Coxed Eights
Row79024	Luigina Giavotti	F	11	Italy	ITA	1928 Summer	1928	Summer	Amsterdam	Gymnastics	Gymnastics Women's Team All-Arou
Row94058	Sonja Henie (-Topping, -Gardiner, -Onstad)	F	11	Norway	NOR	1924 Winter	1924	Winter	Chamonix	Figure Skating	Figure Skating Women's Singles
Row101378	Beatrice Hutiu	F	11	Romania	ROU	1968 Winter	1968	Winter	Grenoble	Figure Skating	Figure Skating Women's Singles
Row102916	Etsuko Inada	F	11	Japan	JPN	1936 Winter	1936	Winter	Garmisch-Pa...	Figure Skating	Figure Skating Women's Singles
Row140650	Liu Luyang	F	11	China	CHN	1988 Winter	1988	Winter	Calgary	Figure Skating	Figure Skating Mixed Ice Dancing
Row152798	Marcelle Matthews	F	11	South Africa	RSA	1960 Winter	1960	Winter	Squaw Valley	Figure Skating	Figure Skating Mixed Pairs
Row237141	Megan Olwen Devenish Taylor (-Mandeville-Ellis)	F	11	Great Britain	GBR	1932 Winter	1932	Winter	Lake Placid	Figure Skating	Figure Skating Women's Singles

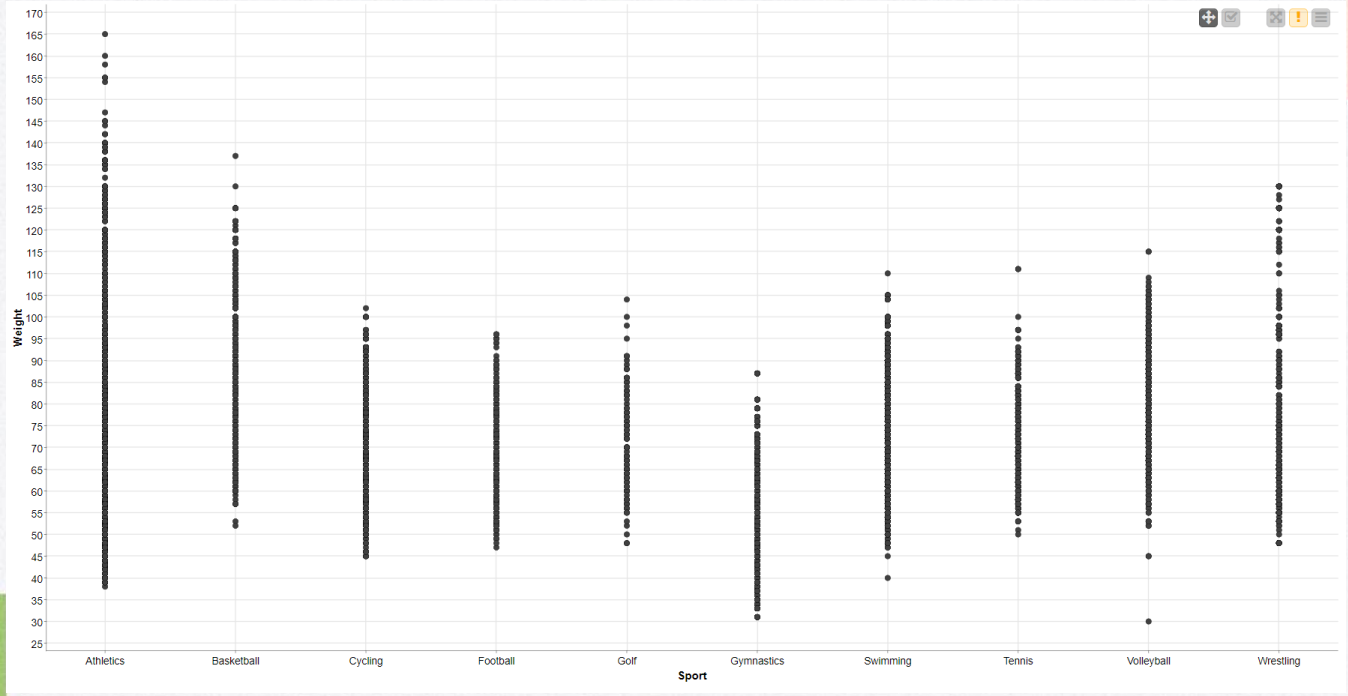
Grafico a torta dei vincitori più giovani ad aver partecipato alle olimpiadi



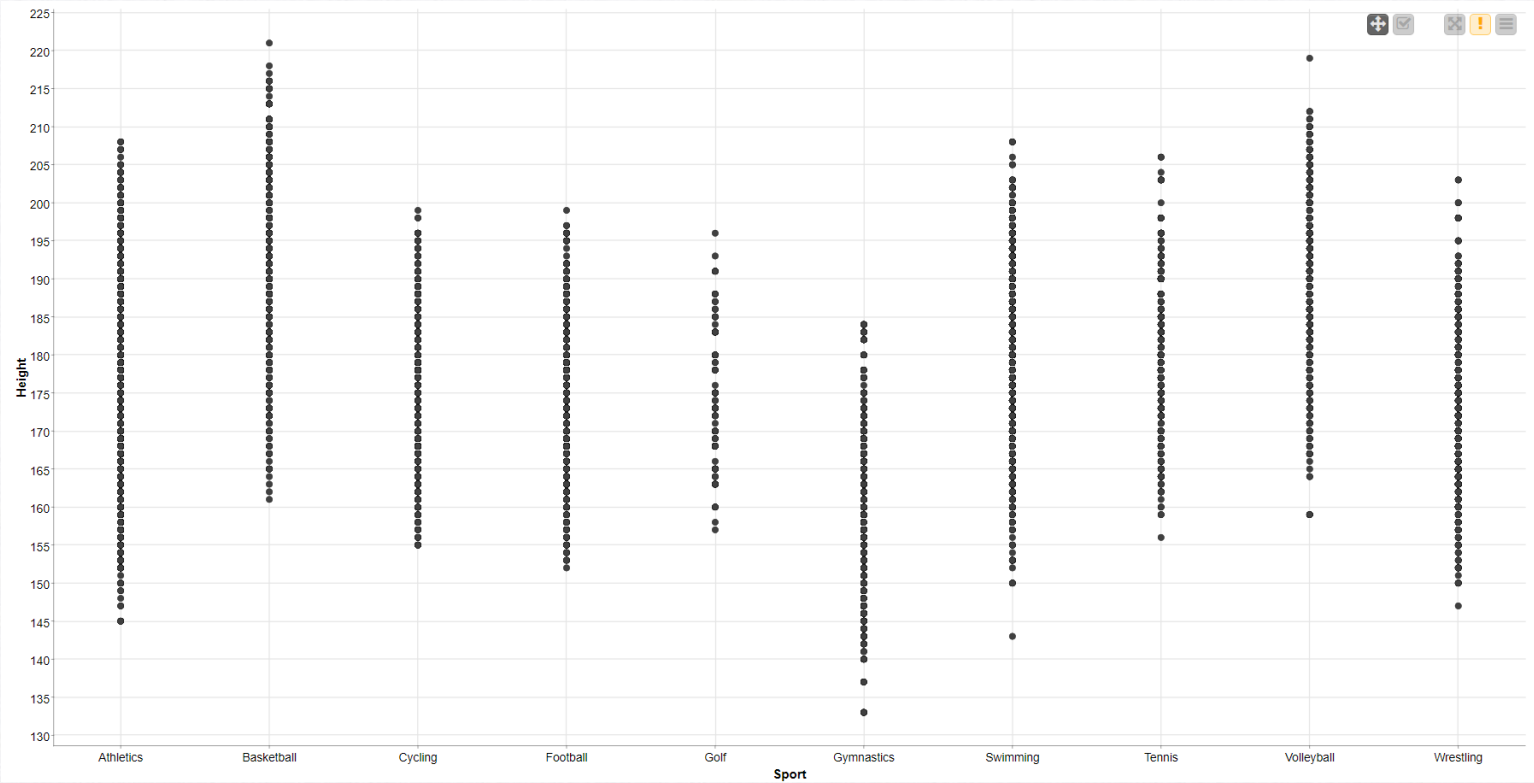
Analisi della struttura atletica (Altezza e peso) degli atleti per sport



Si riporta il grafico a dispersione per i pesi degli atleti



Si riporta il grafico a dispersione per le altezze degli atleti



Conclusioni

KNIME:

- Supporta processi ETL
- Implementa un’interfaccia user friendly
- Offre la possibilità di lavorare in team
- Presenta una parte di visualizzazione dei dati ma non è uno dei punti forti, essendo però un applicativo open source può essere integrato con uno stumento specifico come PowerBI o Tableau

GRAZIE PER L'ATTENZIONE