



StatSoft®

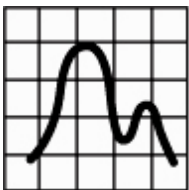
data analysis • data mining • quality control • web-based analytics

Settore dei Semiconduttori

e

STATISTICA

Case Study: Produzione di Wafer di Silicio



STATISTICA

**Soluzioni per Business Intelligence,
Data Mining, Quality Control, e
Web-based Analytics**

Tabella dei Contenuti

SETTORE DEI SEMICONDUTTORI E STATISTICA 3

CASE STUDY: PRODUZIONE DI WAFER DI SILICIO 3

Descrizione..... 3
 Comprensione del Processo Produttivo..... 3

Definizione del Problema..... 4

ANALISI DEI DATI CON STATISTICA 5

Selezione delle Caratteristiche 5

Box Plot..... 6

Scatterplot 6

Box Plot..... 7

Scatterplot 7

Alberi di Classificazione..... 7

Algoritmo *C&RT/CHAID Interattivi* 8
 Risultati – *C&RT Interattivi* 8
 Risultati – *CHAID Interattivi* 9

Altre Applicazioni di Data Mining 10

CONCLUSIONE 10

Riferimenti 10

Settore dei Semiconduttori e **STATISTICA**

Gli algoritmi di data mining hanno giocato un ruolo cruciale e di successo nell'ampio spettro di processi produttivi avanzati. La fase di gestione del prodotto ed il controllo di processo nella produzione di wafer di silicio hanno generato automaticamente grandi volumi di dati. Per questa ragione la tecnologia legata al data mining sta acquistando progressivamente maggiore importanza nella produzione di semiconduttori.

La sofisticazione e le complessità legate alla produzione di chip hanno sempre impedito di realizzare il sogno di ottenere un processo infallibile sul 100% del prodotto. Nonostante le ricette produttive attualmente in uso (e generalmente frutto della combinazione di strumenti utilizzati in 300 – 500 fasi) siano state attentamente progettate e revisionate allo scopo di massimizzare il prodotto, il prodotto stesso sarà sempre soggetto ad errori inevitabilmente introdotti da fattori sistematici (quali ad esempio l'impiego di strumenti difettosi o il ricorso a interazioni tra tool) così come da fattori random (ad esempio, polveri e particolato).

STATISTICA fornisce una suite di flessibili strumenti analitici che potranno essere usati in differenti applicazioni (come ad esempio la root-cause analysis, per l'identificazione di fattori sistematici quali strumenti difettosi o combinazioni di strumenti all'origine dei problemi produttivi). Modelli predittivi, quali quelli descritti in questa sede (*Albero Boosted a Gradiente Stocastico, C&RT Interattivi*, algoritmi *CHAID*, ecc.) possono essere utilizzati per analizzare le misurazioni registrate durante il processo produttivo per identificare come i fattori e le loro interazioni influiscano diversamente sulla qualità generale del prodotto.

Case Study: Produzione di Wafer di Silicio

Descrizione

Il case study rappresenta un esempio di utilizzo delle diverse applicazioni di data mining nel settore della produzione di wafer di silicio. Si noti che questo esempio illustra solo piccole porzioni delle funzionalità complete del kit di strumenti di *STATISTICA*.

Comprensione del Processo Produttivo

La produzione di lamine inizia dalla lavorazione di blocchi di cristalli di silicio, i quali sono soggetti ad una serie di passaggi produttivi (tra i 300 e i 500), al fine di ottenere chip microprocessori al termine dell'ultima fase. Un lotto di lamine (all'incirca 24 "tagli" del blocco di silicio) è chiamato Lotto. Nell'ultima fase, i chip vengono separati su lamina con un filo diamantato per formare singoli circuiti integrati. Durante questi elaborati processi, vengono raccolte tra le 1500 alle 5000 misurazioni su ogni chip.

File Dati

Il file dati *wafer_yield.sta* consiste di dati “Low Level” raccolti in uno stabilimento produttivo dedicato alla produzione di lamine. Questo file dati contiene 2858 variabili e 2062 casi. La maggior parte delle variabili contenute nel file dati contengono informazioni quali “Tools” (Strumentazione) e “Log Time” (Tempi di registrazione) riguardanti 1283 componenti, usati in differenti step (“Log-points”) del processo produttivo. Sono presenti inoltre un numero di altre misure di qualità (circa 40), delle quali la variabile dipendente più importante è il prodotto MULTIPROBE, la misura finale del prodotto per un lotto (o percentuale di chip funzionante). Per categorizzare la misurazione del prodotto finale in Alto/Basso in base al valore mediano di 59,025 è stata calcolata la nuova variabile LowHiYield (calcolata sulla base del prodotto MULTIPROBE).

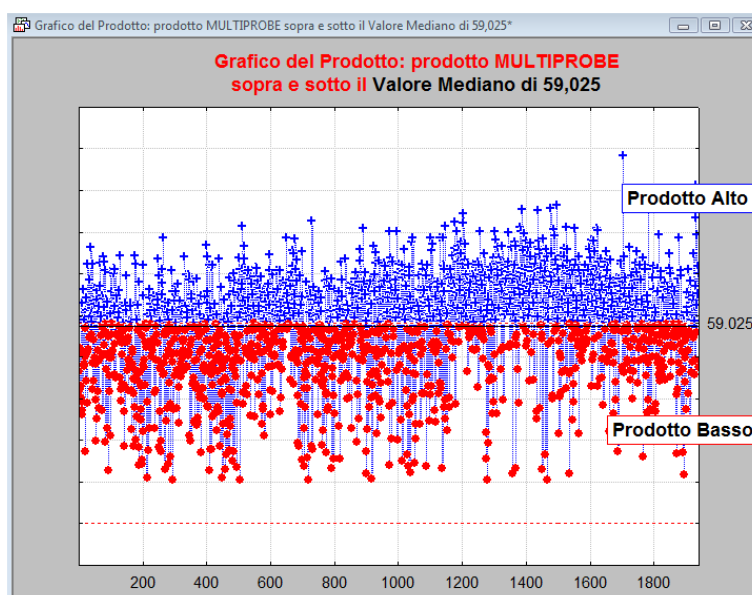
Informazioni sulle Variabili

- Predittori: EQUIPEMENT (“Strumentazione”) – 1406 variabili; LOG TIME (“Tempi di registrazione”) – 1406 variabili; OTHERS (“Altre”) – 46 variabili
- Variabile Dipendente/di Output (o Variabile d’interesse) – prodotto MULTIPROBE
- Nuova Variabile Dipendente – LowHiYield (“Prodotto Alto/Basso”) (basata sul valore mediano di 59,025)

Definizione del Problema

Il problema può essere definito sulla base di due obiettivi: 1) Identificazione degli strumenti e delle interazioni tra strumenti che possono portare all’ottenimento di un prodotto di “bassa” qualità; e 2) Identificazione dei log-points (“punti di rilevazione”) da utilizzare per esaminare qualsiasi fattore di confondimento o di correlazione esistente tra specifici strumenti utilizzati in particolari istanti di tempo.

STATISTICA include una selezione completa di metodi grafici da utilizzare sia per obiettivi di analisi e presentazione dei risultati. La seguente illustrazione è una presentazione visiva delle osservazioni del prodotto MULTIPROBE.



Questo tipo semplificato di carte di controllo della qualità può essere utilizzato per osservare visivamente la distribuzione del prodotto MULTIPROBE a seconda che i rispettivi valori cadano sopra o sotto i livelli standard di cutoff (in questo caso, il prodotto è stato categorizzato come alto o basso a seconda di dove il rispettivo valore si colloca rispetto al valore mediano di 59,025). Qualsiasi particolare tendenza del prodotto può solitamente essere individuata lungo i lotti consecutivi in brevissimo tempo dalla sola osservazione di questa carta.

Analisi dei Dati con STATISTICA

Selezione delle Caratteristiche

In questa dimostrazione viene impiegato lo strumento di Selezione delle Caratteristiche STATISTICA per identificare i migliori predittori (in questo caso, le strumentazioni) che discriminano chiaramente tra i prodotti di Alta/Bassa qualità (dipendente categoriale). Lo strumento di *Selezione di Caratteristiche* è estremamente utile per ridurre la dimensionalità del problema analitico, cioè per selezionare gli specifici predittori (su 1283 in questo caso) da considerare come possibili “candidati” ad essere eletti come cause del problema.

	Predittori migliori per var. d	
	Chi-quadro	p-value
Equipment NW00/5722	137,8842	0,000000
Equipment NW00/5922	137,5644	0,000000
Equipment NW00/5822	137,8248	0,000000
Equipment NW60/8751	84,1724	0,000000
Equipment NW30/9002	61,1891	0,000000
Equipment NW51/11223	37,8189	0,000001
Equipment NW80/11223	39,6548	0,000001
Equipment NW55/4672	31,9599	0,000002
Equipment NW80/8799	41,2171	0,000005
Equipment NW50/11223	31,9504	0,000006

LowHiYield (Categoriale)

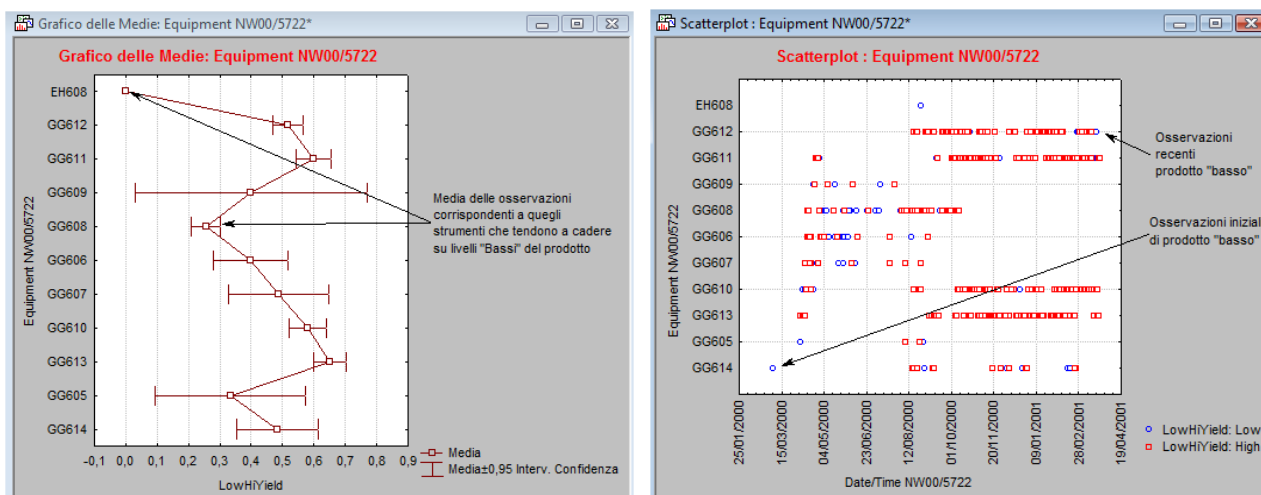
	Predittori migliori per var. d	
	test F	p-value
Equipment NW00/5822	16,07817	0,000000
Equipment NW00/5722	14,53163	0,000000
Equipment NW00/5922	14,45699	0,000000
Equipment NW60/8751	12,57838	0,000000
Equipment NW30/9002	6,58303	0,000000
Equipment NW50/11223	7,91513	0,000000
Equipment NW80/8799	4,96985	0,000001
Equipment NW60/11223	5,64243	0,000002
Equipment NW80/11223	5,51984	0,000003
Equipment NW50/3210	4,84602	0,000006

MULTIPROBE yield (Continua)

La selezione delle caratteristiche ha identificato otto strumentazioni (evidenziate in rosso) quali migliori predittori per la variabile “MULTIPROBE yield” e per la variabile dipendente da essa derivata “LowHiYield”. Questi risultati giustificano l’uso della variabile categoriale “LowHiYield” per meglio comprendere e definire il problema.

Analisi Esplorative: Il seguente passaggio logico consentirà di eseguire segmentazioni sulle informazioni delle strumentazioni identificate per individuare le componenti che con maggiore probabilità sono all’origine della produzione di prodotti di “bassa qualità”.

Variabile Dipendente (Categoriale) – LowHiYield



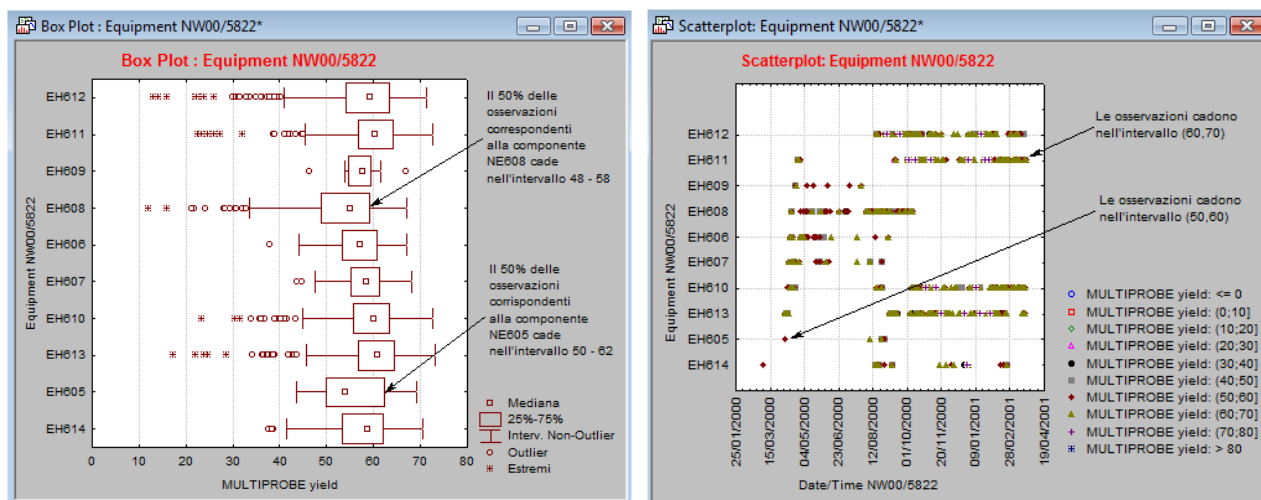
Box Plot

Il grafico riporta le medie della variabile dipendente (LowHiYield) segmentate in base alle componenti della strumentazione NW00/5722 (la funzionalità di selezione delle caratteristiche ha identificato questa strumentazione quale Migliore predittore della variabile LowHiYield). Dall'osservazione di questo grafico, possiamo affermare che il prodotto di bassa qualità è maggiormente associato alle componenti EH608 e GG608. Ad ogni marcatore di media è associato un insieme di baffi |---| che rappresentano le barre d'errore, ovvero sia gli intervalli di confidenza intorno alle media (misura principale del grado di variazione osservata per la particolare componente/strumentazione). Questi grafici sono in genere utilizzati per confrontare le medie marginali lungo i gruppi e verificare l'eventuale attendibilità delle rispettive misure medie.

Scatterplot

In questi grafici categorizzati (scatterplot) sono riportate le osservazioni Alte/Basse di prodotto corrispondenti ad ogni categoria (differenti componenti), per la particolare strumentazione. Questi grafici sono riportati sequenzialmente all'interno di un'unica visualizzazione, per consentire il confronto tra gli andamenti dei dati associati alle rispettive categorie (componenti). Lo scatterplot categorizzato illustra anche quali componenti hanno portato alla realizzazione di prodotti di bassa qualità ed in corrispondenza di quale istante temporale (log-time). In apparenza, le componenti identificate dal Grafico delle Medie quali maggiori responsabili dai problemi di bassa qualità del prodotto (EH608 e GG608) sono principalmente utilizzate nella prima fase della lavorazione. I problemi più recenti sembrano verificarsi durante l'impiego della componente GG612.

Variabile Dipendente (Continua) – MULTIPROBEyld



Box Plot

Nei box plot, per ogni gruppo di casi definiti dalla variabile categoriale sono riportati i rispettivi intervalli di variazione. Per ogni gruppo di osservazioni viene calcolata la tendenza centrale (ad es., media o mediana) e l'intervallo di variazione o statistiche di variabilità (ad es., quantili, errori standard, deviazioni standard, ecc.). Sono riportati su grafico anche i valori più influenti e anomali (si vedano le voci di Legenda Outlier ed Estremi).

I risultati per la variabile dipendente (di output) non trasformata MULTIPROBE ci dicono che la funzionalità di selezione delle caratteristiche ha individuato NW00/5822 quale Migliore Predittore. Il box plot mostra come in corrispondenza delle componenti EH608 e EH605 si ottengano risultati di bassa qualità in confronto alle altre componenti della stessa strumentazione. Il rettangolo ritratto intorno al valore mediano indica l'intervallo di variazione in cui cadrà presumibilmente il 50% delle osservazioni. Si noti inoltre che in corrispondenza della maggior parte delle componenti della medesima strumentazione vi è un alto numero di outlier.

Scatterplot

In questo caso, ogni osservazione del prodotto MULTIPROBE è stata categorizzata in 10 differenti categorie (che rappresentano i differenti intervalli di variazione) e riportata su scala temporale. Per rappresentare la distribuzione del prodotto sono stati riportati i differenti simboli e colori (si veda la legenda) relativi ai differenti intervalli di variazione.

Alberi di Classificazione

Gli alberi di classificazione vengono utilizzati per prevedere l'appartenenza dei casi o degli oggetti nelle classi di una variabile dipendente categoriale sulla base delle rispettive misurazioni su una o più variabili predittrici. L'analisi con alberi di classificazione rappresenta tradizionalmente una delle principali tecniche in uso nel data mining. Il modulo *Alberi di Classificazione* in *STATISTICA Data Miner* è un'implementazione completa di tutte le funzionalità e di tutte le tecniche necessarie per il calcolo degli alberi di classificazione binaria basati su suddivisioni univariate di variabili predittrici categoriali, variabili predittrici ordinali (ovvero sia misurate almeno su scala ordinale), o

di un misto dei di entrambi i tipi di predittori. Il modulo dispone inoltre di opzioni per il calcolo di alberi di classificazione basati su suddivisioni lineari combinate per specifiche variabili predittrici misurate su scala intervallare.

L'obiettivo degli alberi di classificazione è prevedere o spiegare le risposte contenute in una particolare variabile dipendente categoriale, ed in quanto tale, le tecniche raccolte in questo modulo hanno molto in comune con le tecniche utilizzate nella maggioranza delle tecniche più tradizionali quali ad esempio l'Analisi Discriminante, l'Analisi dei Gruppi, e la Regressione Logistica. La flessibilità degli alberi di classificazione fa di questo modulo un'alternativa analitica molto attraente, ciò nonostante non è raccomandabile ricorrere al suo utilizzo come unico strumento di ricerca.

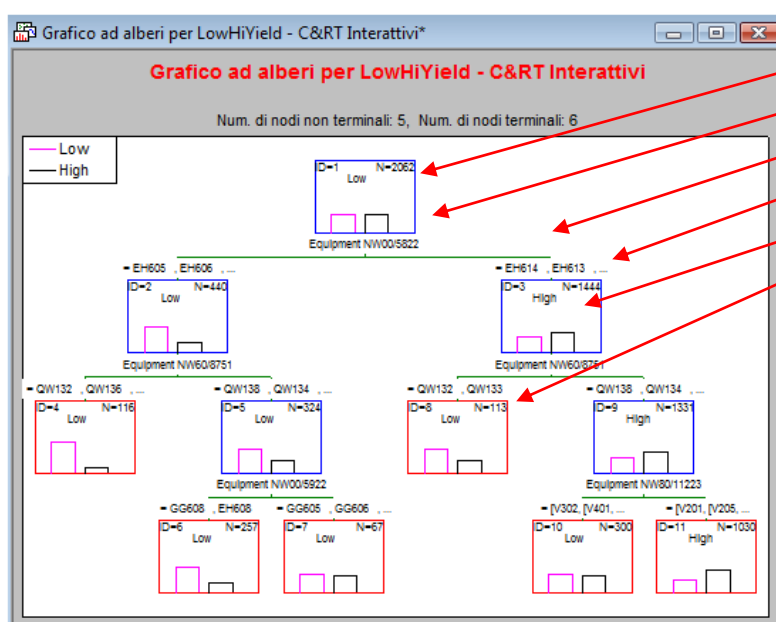
Lo studio e l'impiego degli alberi di classificazione non è molto diffuso nel campo dell'analisi probabilistica e statistica tradizionale (Ripley, 1996), ciò nonostante gli alberi di classificazione sono ampiamente utilizzati in diversi campi applicativi quali ad esempio la produzione (come in questo caso), e la psicologia (teoria decisionale). Gli alberi di classificazione sono infatti molto facili da interpretare data la loro particolare rappresentazione grafica.

Algoritmo C&RT/CHAID Interattivi

Il modulo *STATISTICA Alberi Interattivi (C&RT, CHAID)* costruisce alberi di classificazione e di regressione così come alberi *CHAID* sulla base di metodi automatici (algoritmici), regole e criteri definiti dall'utente specificati via interfaccia utente grafica (strumenti di brushing), e molti altro. L'obiettivo del modulo è fornire un ambiente altamente interattivo per la costruzione di alberi di classificazione e di regressione (basati sui metodi classici *C&RT* e *CHAID*) per consentire agli utenti di sperimentare l'impiego di diversi predittori e criteri di suddivisione in combinazione con quasi tutte le funzionalità di costruzione automatica degli alberi.

Nota: la capacità degli algoritmi *C&RT* e *CHAID Interattivi* di gestire i dati mancanti è una delle ragioni principali per cui preferire questi moduli in sede di analisi esplorativa dei dati.

Risultati - C&RT Interattivi



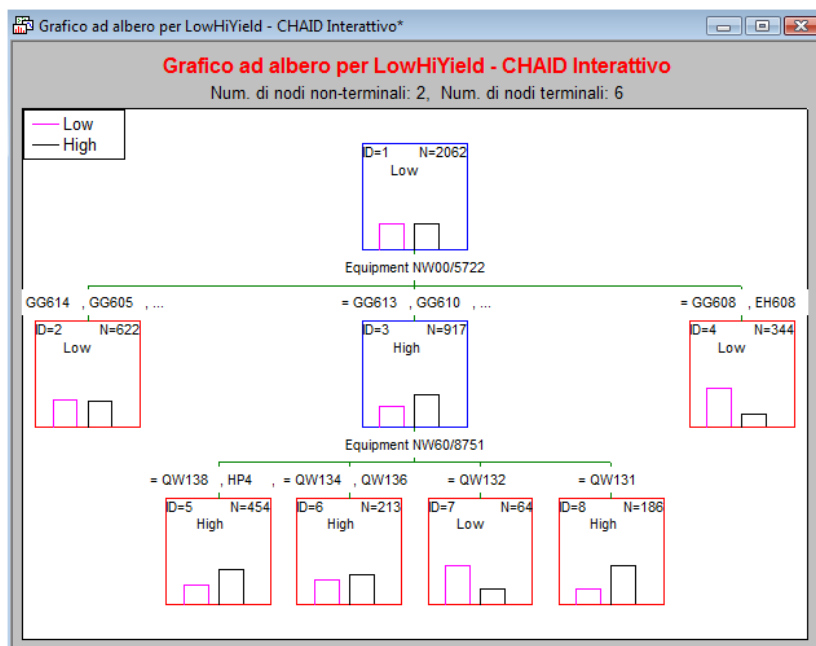
Nodo (Suddivisione) decisionale
Condizione di suddivisione
Nuovo nodo formatosi dal genitore
Numero di casi inviati al figlio
Istogramma dei casi in ogni classe del nodo
Nodo terminale (foglia)

L'Interpretazione di questi alberi è piuttosto semplice. L'algoritmo *C&RT* ha identificato interazioni tra la componenti le 43NW/3600, 746NW/6529, e 428NW/2225 quali principali fattori che hanno determinato l'ottenimento di un prodotto di "Alta" o di "Bassa" qualità.

Le regole generate da questi alberi (disponibili inoltre nella tabella della struttura ad albero) può consentire ai tecnici d'individuare le componenti (o le specifiche combinazioni di strumenti) che hanno causato problemi di bassa qualità nel prodotto. Come si può osservare dal grafico precedente, l'algoritmo *C&RT* ha distinto tra 5 possibili decisioni alternative (contenute nei 6 nodi terminali evidenziati in rosso) costruite sulla base di 5 condizioni "se allora" che consentono la previsione della categoria di prodotto. Seguendo il percorso che parte dal nodo radice (ID=1) fino al nodo terminale (ID=4), è possibile derivare una regola generale per l'ottenimento di un prodotto di "Bassa" qualità. Se le componenti NE 605, NE 606, ecc. della strumentazione 43NW/3600 vengono utilizzate insieme alle componenti NT132, NT136, ecc. della strumentazione 746NW/6529, allora vi saranno 116 casi (osservazioni) registrate la cui maggioranza cadrà all'interno della categoria "Low". Nello stesso modo, è possibile analizzare gli altri rami per trarne conclusioni ulteriori. La legenda che identifica quali barre degli istogrammi di nodo corrispondano alle due categorie di prodotto sono collocate nella parte in basso a destra del grafico.

Le analisi seguenti con altri algoritmi di data mining danno un supporto ulteriore a questi risultati.

Risultati - CHAID Interattivi



L'algoritmo *CHAID* ha identificato un'interazione tra NW00/5722 ed NW60/8751, che rispecchia quanto stabilito durante l'analisi *Selezione di Caratteristiche*. Come si può vedere dal grafico precedente, l'algoritmo *CHAID* ha distinto 6 possibili soluzioni alternative (contenute in altrettanti nodi terminali evidenziati in rosso) costruite sulla base di 6 condizioni "se allora" utilizzate per prevedere la categoria del prodotto. Un utente può esaminare le suddivisioni presenti in questo albero di classificazione esattamente come fatto con l'albero decisionale *C&RT*. Per esempio, in corrispondenza della prima suddivisione, l'algoritmo *CHAID* ha identificato che l'uso di specifiche componenti (GC608 ed EC608) della strumentazione NW00/5722 influisce negativamente il prodotto (le osservazioni contenute nel nodo terminale per questa regola decisionale sono per lo più associate a prodotti di bassa qualità). Tuttavia, i risultati *CHAID* sono talvolta più sensibili ai

particolari andamenti dei dati, mentre i risultati *C&RT* dovrebbero essere esplorati sempre nella prima fase.

Altre Applicazioni di Data Mining

Le tecniche di data mining possono essere utilizzate per molte altre applicazioni nei seguenti settori:

1. **Metodi automatizzati che possono consentire l'identificazione e la classificazione di gruppi difettosi di chip di memoria.** (Spatial Signature Analysis) – Il controllo della qualità nel settore dei semiconduttori è stato tradizionalmente basato sull'osservazione dei dati riassuntivi generali (ad esempio le misure di rapporto tra numero di chip buoni sul totale dei chip prodotti). L'uso di queste misure aggregate sarebbe accettabile se i chip difettosi si distribuivano casualmente su tutta la superficie dei wafer così come lungo tutti i lotti di produzione. In realtà, i difetti si presentano sempre in gruppi, oppure sono contraddistinti da comportamenti sistematici da poter utilizzare per tracciare i fattori che causano i problemi. L'identificazione e la classificazione di chip difettosi possono essere totalmente automatizzate attraverso il ricorso a tecniche avanzate di raggruppamento e ad altri algoritmi di data mining quale alternativa all'ispezione manuale di ogni singolo pezzo, come attualmente avviene nella maggior parte delle industrie. Ciò può consentire di abbattere i tempi e costi di monitoraggio e di ottenere nel contempo risultati molto più attendibili.
2. **Arrotondamento durante la lucidatura dei wafer:** Una delle fasi finali della produzione di chip coinvolge la lucidatura dei wafer prima dell'inserimento dei circuiti integrati. Durante questa fase, spesso accade che gli angoli dei wafer subiscano una sorta di arrotondamento dovuto all'azione delle spazzole lucidatrici. Di conseguenza viene ridotta l'area d'integrazione dei chip. I fattori parametrici (quali temperatura, pressione, tipo di spazzole per la pulitura, ecc.) all'origine del problema possono facilmente essere individuati attraverso il ricorso a specifiche tecniche di data mining, e messi in relazione tra loro attraverso avanzati modelli predittivi da impiegare per "correggere" specifiche impostazioni ed evitare in questo modo che si verifichi un qualsiasi deterioramento del prodotto.

Conclusione

L'esempio ha qui descritto una delle innumerevoli applicazioni in cui l'utilizzo del software *STATISTICA Data Miner* risulta vincente. Le funzioni analitiche e grafiche implementate in *STATISTICA* possono essere applicate per risolvere una varietà di problemi produttivi, ed in particolare, per discriminare tra i possibili fattori che determinano la buona riuscita di un prodotto (come abbiamo visto in questa sede). Gli strumenti avanzati di data mining sono estremamente utili per incrementare l'efficacia delle tecnologie ed i processi esistenti, per produrre prodotti di migliore qualità, e per ottimizzare la capacità ed i livelli produttivi.

Riferimenti

White, K., Mastrangelo, C., Modeling, Analysis, and Information Technologies for Semiconductor Manufacturing, from <http://www.sys.virginia.edu/research/semi.asp>

Kusiak, A., Decomposition in Data Mining: An Industrial Case, from

<http://www.icaen.uiowa.edu/~ankusiak/Journal-papers/Decomp.pdf>