PREDIZIONE DEL RISCHIO DI MALATTIE LAVORO-CORRELATE ATTRAVERSO ANALISI DI CLUSTERING E OTTIMIZZAZIONE GENETICA

by

Antonio di Noia

A thesis submitted
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information and Communication Engineering

at

Sapienza University of Rome

May 2016

XXVII

Author ……Dr. Antonio di Noia
Department of Information Engineering, Electronics and Telecommunications

Certified by …… Prof. Antonello Rizzi
Scientist
Thesis Supervisor

Ringraziamenti

Il lavoro presentato in questa tesi è stato condotto presso il DIET - Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni dell'Università "Sapienza" di Roma e presso il Dipartimento di Ricerca di Medicina, epidemiologia, igiene del lavoro e ambientale - Area Salute sul Lavoro dell'INAIL - Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro grazie ad un accordo di collaborazione scientifica.

Voglio esprimere il mio più profondo apprezzamento e ringraziamento al mio tutor Prof. Antonello Rizzi per la sua pazienza, disponibilità, continua presenza, per i suoi consigli e per i suoi incoraggiamenti nei momenti di difficoltà utili per il completamento di questa tesi.

Voglio ringraziare i componenti della mia sotto-commissione di tesi di laurea per i loro feedback e suggerimenti, un grazie particolare alla Prof.ssa Maria Gabriella Di Benedetto per la sua pazienza e disponibilità.

Un grazie particolare al mio collega e amico Ing. Paolo Montanari del Dipartimento di Ricerca dell'INAIL che mi ha fatto da co-tutor, collaborando e cooperando attivamente nelle varie fasi di questo lavoro di ricerca, consigliandomi e indirizzandomi nelle scelte in modo critico facendomi crescere professionalmente aiutandomi a completare questo lavoro di tesi.

Un grande grazie a mia moglie per la sua tolleranza e comprensione e per il suo sostegno morale nei miei momenti di scoraggiamento. Un grazie ai miei due figli per avermi compreso e tifato sempre per me incoraggiandomi con la loro presenza e con i loro giochi risollevandomi dai miei momenti di pessimismo.

Infine voglio esprimere la mia gratitudine a mia madre, mio padre e mio fratello per il loro pieno supporto e sostegno nell'arco di tutta la mia vita formativa ed educativa.

Ai miei figli

Sommario

Indice delle figure

Indice delle tabelle

Glossario

AG	=	Algoritmo Genetico
API	=	Application Programming Interface
ASL	=	Azienda Sanitaria Locale
ATECO	=	Codifica Istat dell'ATtività ECOnomica svolta da un'azienda.
DB	=	Database
DBMS	=	Database Management System
ER	=	Entità Relazioni
INAIL	=	Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro
ISCO	=	International Standard Classification of Occupations
ISPESL	=	Istituto Superiore per la Prevenzione e la Sicurezza del Lavoro
ISTAT	=	Istituto Nazionale di Statistica
LWD	=	Linearly Weighted Distance
MALPROF	=	Modello strutturato per lo studio delle Malattie Professionali.
MaProWeb	=	Applicativo web-based per l'inserimento delle segnalazioni di Malattie Professionali
M.D.L.	=	Medicina del Lavoro
NACE	=	Nomenclatura delle Attività Economiche nella Comunità Europea
ODBC	=	Open DataBase Connectivity
OOAD	=	Object-Oriented Analysis and Design
OO	=	Object Oriented
OOP	=	Object-Oriented Programming
OS	=	Operating System
PA	=	Pubblica Amministrazione
POMOS	=	POlo di MObilità Sostenibile - Centro di ricerca dell'Università "Sapienza" di Roma

SPARE	=	Something for PAttern REcognition - Librerie di classi (template) C++
STL	=	Standard Template Library
SVM	=	Support Vector Machine
UML	=	Unified Modeling Language. Linguaggio di Modellazione e specifica Unificato.

Abstract (compendio)

Il lavoro di questa tesi è stato condotto attraverso le tecniche di intelligenza computazionale per uno studio sulla predizione dei rischi per la salute nei posti di lavoro. Il dataset disponibile è stato popolato da parte delle Aziende Sanitarie Locali (ASL) nell'ambito di un programma per la realizzazione del Sistema Nazionale di Sorveglianza per le malattie professionali e gli infortuni mortali. Lo scopo principale di questo lavoro è la progettazione di un'applicazione software capace di evidenziare situazioni di maggior criticità per la manifestazione di Malattie Professionali che possa essere usata agevolmente da parte dei medici del lavoro come strumento di supporto nella loro attività di prevenzione e sorveglianza della salute dei lavoratori. Gli algoritmi proposti, utilizzano tecniche di clustering e l'ottimizzazione genetica per determinare in maniera automatica sia i pesi delle caratteristiche prese in considerazione nel calcolo della distanza interindividuale che il numero di cluster per la sintesi del classificatore finale. In particolare, si propone un nuovo approccio che consiste nel definire il classificatore generale come un insieme di classificatori specifici per ciascuna classe di patologia, ciascuno addestrato a riconoscere le condizioni di rischio che caratterizzano una singola patologia. I primi risultati sono incoraggianti e suggeriscono interessanti temi di ricerca per un ulteriore sviluppo del sistema.

1. Introduzione

1.1. Le origini dello studio delle malattie professionali

I primi studi della storia della medicina sulle malattie correlate al lavoro furono condotti dal medico Bernardino Ramazzini (Carpi, 4 ottobre 1633- Padova, 5 novembre 1714), professore di medicina pratica all'Università di Padova.

Egli scrisse e pubblicò, nell'ultimo decennio del suo secolo, per la prima volta un trattato, il "*De Morbis Artificum Diatriba"*; l'opera di Ramazzini è interessante soprattutto da un punto di vista storico- culturale, derivante anche dall'originalità dell'argomento; oggi, in realtà è considerato l'atto fondante dell'attuale Medicina del Lavoro.

L'associazione tra pericolo o fattore di rischio e danno o malattia evidenziata da Ramazzini nasce come sostanziale "valutazione del rischio" in senso epidemiologico da intuizioni e deduzioni logiche che pur fondate solidamente sulle migliori conoscenze cliniche e "sociologiche" presenti ai tempi in cui l'autore scrive, anticipano, in una certa misura, gli studi epidemiologici di tipo occupazionale.

Egli, per primo, riconobbe la necessità di realizzare ambienti di lavoro "sicuri", obiettivo che cercò di raggiungere proponendo i primi, rudimentali dispositivi di protezione dei lavoratori.

1.2. Contesto

Oggi è più opportuno parlare di *"malattia correlata al lavoro"* e non di *"malattia da lavoro"* per indicare la multifattorialità delle malattie contratte nel luogo di lavoro.

La *malattia professionale* (anche detta lavoro-correlata) è un evento dannoso alla persona che si manifesta in modo lento, graduale e progressivo, involontario e in occasione del lavoro. Nella malattia professionale, diversamente che nell'infortunio, l'influenza del lavoro nella genesi del danno lavorativo è specifica, poiché la malattia deve essere contratta proprio nell'esercizio ed a causa di quell'attività lavorativa o per l'esposizione a quella determinata noxa patogena.

A livello mondiale, in molti Stati, l'attenzione nei confronti della salute dei lavoratori è in aumento, sia in ambito pubblico che privato. Si riscontra un progressivo maggior interesse ed un incremento negli investimenti di risorse da parte di tutte le organizzazioni nazionali di prevenzione del rischio per la salute, per la salvaguardia e prevenzione sulle principali malattie/patologie derivanti dalle varie attività lavorative, assumendo spesso una rilevanza strategica per l'azienda. Secondo studi internazionali, nelle economie avanzate i costi economici connessi alle malattie professionali ammontano ad almeno il 3% del prodotto interno lordo.

L'Unione Europea persegue il miglioramento delle condizioni di vita e di lavoro dei lavoratori europei, in collaborazione con i governi dei Paesi membri della U.E. e con le agenzie europee che si occupano di prevenzione e sicurezza nei luoghi di lavoro (OSHA e Eurofound), con particolare attenzione alle politiche di prevenzione e riduzione delle malattie professionali sviluppando adeguate strategie e strumenti.

Nuovi fattori di rischio stanno emergendo per le patologie lavoro-correlate a causa dell'accresciuta complessità delle caratteristiche occupazionali ed ambientali. Inoltre, la durata della vita lavorativa, in alcuni casi, è stata prolungata facendo aumentare di conseguenza il rischio per esposizione. Infatti, una parte dei costi pubblici dedicati alla prevenzione della salute dei lavoratori può essere ridotto monitorando e controllando i pericoli nei luoghi di lavoro grazie allo studio

dell'insorgere delle patologie lavoro-correlate in relazione a specifiche attività professionali e a specifiche condizioni di lavoro, in modo da individuare le più idonee misure di prevenzione.

La valutazione dei rischi professionali, nonché la loro prevenzione, non solo è in grado di proteggere i diritti e gli interessi dei lavoratori riguardo alla loro salute, ma consente anche di migliorare la gestione della salute sul lavoro da parte dei datori di lavoro con conseguente guadagno economico.

1.3. Stato dell'arte

Diversi studi mostrano che l'applicazione di tecniche di intelligenza computazionale può portare a rilevare l'esistenza di regolarità nei dati, difficili da rilevare con altri approcci.

1. In (Hongbo Liu et al, 2009) è studiata una tecnica per l'identificazione e la classificazione di gruppi ad alto rischio da pneumoconiosi di lavoratori del carbone con l'utilizzo di una rete neurale artificiale basata sulle storie lavorative. E' stata sviluppata una rete neurale artificiale three-layer con 6 variabili di input, 15 neuroni nello hidden layer e 1 neurone di output, utilizzando i dati dell'esposizione professionale dei minatori di carbone.
2. In (T Taylor et al, 2009) è stata studiata l'utilità di modelli fuzzy nella valutazione del rischio biologico.
3. In (Razan et al, 2010) sono state applicate tecniche di clustering su dati medici per predire con una certa probabilità il rischio di malattie.
4. In (Chinmoi et al, 2012) viene presentato lo sviluppo e l'implementazione di un sistema di supporto decisionale per la cura della salute degli impiegati; è stato ideato al fine di ridurre il costo delle cure sanitarie dei dipendenti, migliorando la prevenzione e aumentando la consapevolezza della forza lavoro riguardo alla prevenzione delle malattie preservando la salute con piccoli accorgimenti giornalieri.
5. In (Zhaohui Huang Yu Daoheng Jianye Zhao, 2000) sono state utilizzate le reti neurali artificiali per la previsione di malattie professionali in un dato settore.
6. In (Shankaracharya et al, 2010) viene condotta un'analisi per delineare un insieme di interesse della vasta gamma di opzioni dei recenti sviluppi degli algoritmi di machine learning utilizzati per la realizzazione di strumenti di diagnosi per il diabete e delle loro potenzialità. In particolare l'attenzione è rivolta verso i metodi supervisionati e non supervisionati.
7. In (Michael P. Menden et al, 2013) attraverso l'uso di tecniche di machine learning, nel tentativo di integrare altri approcci, viene condotto uno studio di previsione della risposta di linee cellulari tumorali al trattamento farmacologico, sulla base sia di caratteristiche genomiche delle linee cellulari che delle proprietà chimiche dei farmaci considerati.
8. In (Mattia CF Prosperi et al, 2013) vengono utilizzate tecniche di machine learning per prevedere fenotipi di asma e eczema nell'ambito della diagnostica clinica. Utilizzando i metodi di apprendimento automatico si è cercato di predire i casi di asma, respiro affannoso ed eczema in un campione casuale di popolazione adoperando un grande insieme eterogeneo di attributi con l'obiettivo di individuare in quale misura queste informazioni possono essere combinate per ottenere una buona predizione per i casi clinici in esame.
9. In (Julio J. Valdes et al, 2007) mediante l'uso dell'intelligenza computazionale e di tecniche di visual data mining, viene presentato uno studio di analisi di dati di espressione genica mediante microarray di pazienti affetti e non affetti da sclerodermia al fine di trovare sottoinsiemi di attributi rilevanti con capacità di alta classificazione.

10. In (Joachim Schneider et al, 2007) è stato condotto uno studio con lo scopo di valutare la potenza diagnostica di un classificatore fuzzy per la rilevazione dei tumori al polmone rispetto ai pazienti di asbestosi ad alto rischio di sviluppare il cancro ai polmoni.
11. In (Ishtake S.H and Sanap S.A.) il principale obiettivo è stato lo sviluppo di un prototipo di un sistema intelligente di predizione delle malattie del cuore utilizzando tre tecniche di modellazione di data driven ed esattamente: alberi decisionali, tecniche di Naive Bayes e reti neurali.
12. In (Jyoti Soni et al, 2011) è stato progettato un sistema per la predizione di malattie cardiovascolari. La predizione viene eseguita attraverso un indagine sui dati storici del paziente o di un repository. Viene descritto un classificatore che si basa su di una regola di associazione pesata. Differenti pesi vengono assegnati a differenti attributi utilizzando l'esperienza dei medici del settore.
13. In (K.Srinivas et al, 2010) è stato svolto un lavoro con le più note tecniche di data mining con utilizzo delle reti neurali per la previsione di malattie cardiache dei lavoratori delle miniere di carbone.

In tutti questi lavori è possibile ritrovare studi di ricerca per l'applicazione di strumenti di computational intelligence per la salvaguardia della salute e benessere di lavoratori in un determinato settore lavorativo, in una particolare area o attività economica, o nell'ambito della diagnostica in generale, ma in nessuno di essi è stato riscontrato o rilevato uno studio compiuto su tutte le attività lavorative e professionali.

L'applicazione di tecniche di intelligenza artificiale, knowledge discovery e data mining alla ricerca e prevenzione delle malattie lavoro-correlate in generale, risulta essere uno studio nuovo ed un primo tentativo di utilizzare tecniche già note.

L'esplorazione ed il tentativo di reperire informazioni circa attività di ricerca nel settore delle malattie lavoro-correlate con utilizzo di tecniche di computational intelligence anche attraverso la consultazione di siti specifici e dedicati alla salute ed al benessere dei lavoratori di interesse internazionale quali: il World Health Organization (http://www.who.int/en/), l'Office of Scientific and Technical Information of U.S. Department of Energy (http://www.osti.gov), il CORDIS - Servizio Comunitario di Informazione in materia di Ricerca e Sviluppo (http://cordis.europa.eu/home_it.html), l'Occupational Health and Safety - American Magazine (https://ohsonline.com/Home.aspx), l'Occupational Safety & Health Administration - Agenzia Europea per la sicurezza e la salute sul lavoro (https://osha.europa.eu/it) non ha prodotto riscontri riguardo ad attività anche solo simili.

E' da ritenere, pertanto, che il presente studio può essere considerato innovativo ed iniziatico anche qualora si rinvenissero lavori simili al momento non rilevati nel qual caso potrebbero essere utilizzati sia per attività di confronto reciproco che per ulteriori sviluppi collaborativi ed integrativi a quello qui presentato.

Inoltre, l'idea di sviluppare un software ex novo attraverso un ambiente di sviluppo con un linguaggio di programmazione quale il Visual C++ per un sistema operativo specifico quale Windows non diffusi nello stato dell'arte della bibliografia rilevata ed in generale in ambito di lavori di ricerca, dove solitamente vengono adoperati strumenti già collaudati, arricchisce ulteriormente il presente lavoro di un ulteriore elemento di innovazione.

1.4. Scopo e obiettivo

Una sfida potenzialmente utile è quella di applicare tecniche di knowledge discovery e data mining su banche dati correlate per l'estrazione di informazioni utili finalizzate all'elaborazione con tecniche di intelligenza computazionale ed alla valutazione dei rischi sul lavoro con i metodi di classificazione dei rischi sanitari.

In riferimento alle malattie professionali (ma anche agli infortuni sul lavoro con relativo danno biologico), il tema di ricerca vuole riguardare uno studio particolareggiato nell'ambito della salute e sicurezza sul lavoro, al fine di:

- Progettare e realizzare un'applicazione software nuova su SO Windows di interesse medico-sanitario, attraverso lo sviluppo di prototipi di sistemi automatici, basati sulle tecniche dell'intelligenza computazionale attraverso lo studio ed analisi di diverse combinazioni di tali tecniche.
- Sperimentare l'impiego dei sistemi sviluppati su dati reali del settore delle malattie lavoro-correlate ai fini predittivi del rilevamento precoce e della prevenzione (supporto alla diagnostica clinica).

Tale progetto è stato anche inquadrato in una collaborazione con l'Area Ricerca dell'Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro - INAIL.

L'idea proposta è stata quella di sviluppare un sistema di classificazione basato sulla combinazione di Cluster Analysis e di un Algoritmo Genetico in grado di interfacciarsi con il database delle malattie occupazionali di livello nazionale, prelevando i dati necessari a tale elaborazione attraverso una preventiva fase di Data Cleaning.

1.5. Attività di ricerca

L'attività di ricerca svolta ha riguardato l'area dell'apprendimento automatico inerente il *Pattern Recognition*, i metodi di *apprendimento non supervisionato e supervisionato*, la *cluster analysis*, gli *algoritmi genetici.*

Le attività di ricerca e sviluppo possono essere sintetizzate dal seguente elenco di sub-task:

- Reperimento di dati reali di malattie lavoro-correlate inerenti lavoratori affetti da patologie e loro analisi. I dati utilizzati, si basano sulla conoscenza di informazioni raccolte attraverso la collaborazione dei servizi di prevenzione delle malattie lavoro-correlate.
- Individuazione delle caratteristiche per l'estrazione del dataset necessario per la definizione della base dati.
- Sviluppo di un algoritmo di machine learning per la prevenzione e predizione delle malattie lavoro-correlate.
- Sviluppo ed implementazione di tecniche di intelligenza computazionale quali: analisi di clustering, algoritmi genetici, indici di validazione, tecniche di classificazione.
- Implementazione di un applicativo software come strumento innovativo per la prevenzione/predizione delle malattie lavoro-correlate.

I principali algoritmi utilizzati riguardano: *Clustering* e *Regole di associazione*; in particolare, come algoritmo di clustering è stato utilizzato il *K-means, e sue varianti*.

Per la realizzazione sono state utilizzate tecniche di clustering ottimizzato, congiuntamente alla progettazione di un algoritmo genetico e di opportuni indici relativi di validazione di analisi di clustering, per la definizione della funzione obiettivo.

In una prima fase sono state completate attività di analisi e preprocessamento dei dati provenienti dal sistema MalProf dell'Inail per l'individuazione del sottoinsieme di interesse dei casi costituenti i dataset di input all'applicazione da sviluppare, per l'individuazione delle variabili più significative, eliminazione dei record che su tali variabili presentavano dati ambigui e/o incompleti, per la creazione di raggruppamenti ad hoc relativi alle variabili individuate, per la determinazione di un insieme di test per l'algoritmo di clusterizzazione.

In una seconda fase si è proceduto all'implementazione di algoritmi di data-mining basati su tecniche di clustering finalizzati alla definizione di opportune misure di dissimilarità parametriche. Lo sviluppo di una applicazione software in grado di interfacciarsi con i data base dell'INAIL ha consentito l'individuazione di un set ottimale dei parametri della metrica adottata per il clustering, caratterizzante la popolazione dei lavoratori, in grado di evidenziare la distribuzione delle patologie nei cluster in rapporto alla popolazione totale, nonché la caratterizzazione dei lavoratori che hanno sviluppato una determinata patologia. Il software è stato confrontato e validato anche con strumenti di analisi statistica dei dati di terze parti.

1.6. Fasi, strumenti e tecniche

Il lavoro di ricerca e studio si è sviluppato attraverso tre principali fasi. Una prima fase di individuazione, reperimento dei dati loro analisi e preprocessamento per la costituzione dei dataset da elaborare con parallela attività di individuazione degli strumenti e tecniche per l'elaborazione dei dataset estratti. Una seconda fase riguardante l'implementazione dello strumento software con cui elaborare i dataset. Una terza fase di elaborazione dei dati ed analisi dei risultati ottenuti.

In particolare per il *preprocessamento*, è stata condotta un'attività di ricerca nel database di origine dei dati più significativi, con conseguente loro estrazione e relativo cleaning. Dai dati ottenuti sono state individuate le features dei lavoratori e costruiti i dataset da sottoporre alle elaborazioni.

In parallelo all'attività di analisi e reperimento dei dati, è stata condotta un'altra attività anch'essa di base per lo sviluppo delle fasi successive. Essa ha riguardato l'individuazione degli strumenti di lavoro attraverso i quali realizzare il software per l'elaborazione dei dati contenuti nei dataset su cui sperimentare gli algoritmi di clustering prima e genetico poi.

Il dataset così ottenuto, è stato sottoposto ad una attività di sperimentazione di tecniche di *clustering* per la valutazione di varie funzioni di dissimilarità tra dati al fine di individuare la più idonea funzione a fini predittivi/preventivi delle malattie professionali; i risultati del clustering, sono stati inoltre utilizzati per eseguire analisi ulteriori sui dati in esame.

Il clustering è stato utilizzato in combinazione con un *algoritmo genetico* per l'ottimizzazione dei pesi delle features e l'individuazione del numero ottimale di clusters.

In tutte queste attività sono state utilizzate tecniche ed algoritmi già noti, combinandoli secondo uno schema innovativo per cercare di ottenere uno strumento nuovo, utile alla previsione e prevenzione delle malattie professionali, facilmente riutilizzabile ed estensibile ai casi di malattie generiche, anche derivanti da cause diverse da quelle professionali.

2. I Dati

2.1. Fonte dei dati

I dati oggetto delle elaborazioni, nel rispetto del codice in materia di protezione dei dati personali (codice della privacy), decreto legislativo 30 giugno 2003, n. 196, sono stati estratti dagli archivi messi a disposizione dal settore ricerca dell'INAIL nell'ambito della linea di ricerca P31L04: *"Progettazione e sperimentazione di sistemi intelligenti applicati alla prevenzione sul lavoro"* - piano di attività 2013 - 2015.

Tali archivi sono uno dei risultati ottenuti dal progetto MALPROF dell'INAIL ex ISPESL (Istituto Superiore per la Prevenzione e la Sicurezza del Lavoro), progetto che viene attivato nel 2007 attraverso una convenzione stipulata tra il Ministero della Salute e l'ex ISPESL, avente come obiettivo la raccolta e la registrazione delle segnalazioni di patologie correlate al lavoro da parte dei Servizi di prevenzione delle Aziende Sanitarie Locali (ASL) secondo un modello strutturato, denominato appunto MALPROF, allo scopo di analizzare la possibile esistenza di nessi causali tra l'attività lavorativa e la patologia riscontrata nel lavoratore, portando alla realizzazione di un database nazionale delle patologie correlate al lavoro residente presso l'INAIL.

Tale finalità si è concretizzata attraverso:

- implementazione e aggiornamento degli strumenti standardizzati per la registrazione delle patologie correlate al lavoro da inserire nel database nazionale dell'INAIL;

- programmazione di iniziative di aggiornamento professionale degli operatori, per migliorare le capacità di registrazione e analisi delle patologie correlate al lavoro ed il livello di omogeneità nella attribuzione dei nessi causali tra l'attività e la malattia stessa;

- rendere operativa l'architettura del Sistema di sorveglianza nazionale, realizzare la reportistica di restituzione delle informazioni residenti nel database nazionale dell'INAIL ed attivare modelli efficaci di comunicazione.

L'alimentazione della base dati per il sistema MALPROF, è stata realizzata per mezzo di un applicativo web based denominato *MaProWeb* utilizzato per l'inserimento delle segnalazioni di malattie professionali raccolte dai servizi di prevenzione delle ASL.

L'applicativo MaProWeb è il risultato della ricerca prevista dal piano di attività 2004 dell'ex ISPESL "Implementazione del sistema di sorveglianza nazionale sulle segnalazioni di malattia professionale che pervengono ai Servizi di Prevenzione delle ASL attraverso la realizzazione di un applicativo software web-based per l'inserimento delle informazioni prese dal modello concettuale, nel database centralizzato presso l'ISPESL".

L'applicativo di rilevazione web-based MaProWeb è stato implementato con la finalità di recuperare le conoscenze riguardanti le malattie professionali disponibili presso le ASL. Esso è installato su un server presso l'INAIL ex ISPESL e accessibile tramite INTERNET.

Attraverso l'applicativo MaProWeb, è stata svolta una attività che ha portato alla realizzazione di una base dati contenente informazioni raccolte per la prima volta in un unico DB a livello nazionale disponibile per le elaborazioni ed analisi attraverso sistemi di data warehouse interattivi e rappresentate in forma multidimensionale facilmente consultabili; nonché con tecniche di data mining ed altre tecniche che consentano la ricerca e l'estrazione di informazioni significative non evidenti. Il processo di alimentazione dati da parte degli utenti preposti attraverso l'applicativo MaProWeb è mostrato nella figura seguente:

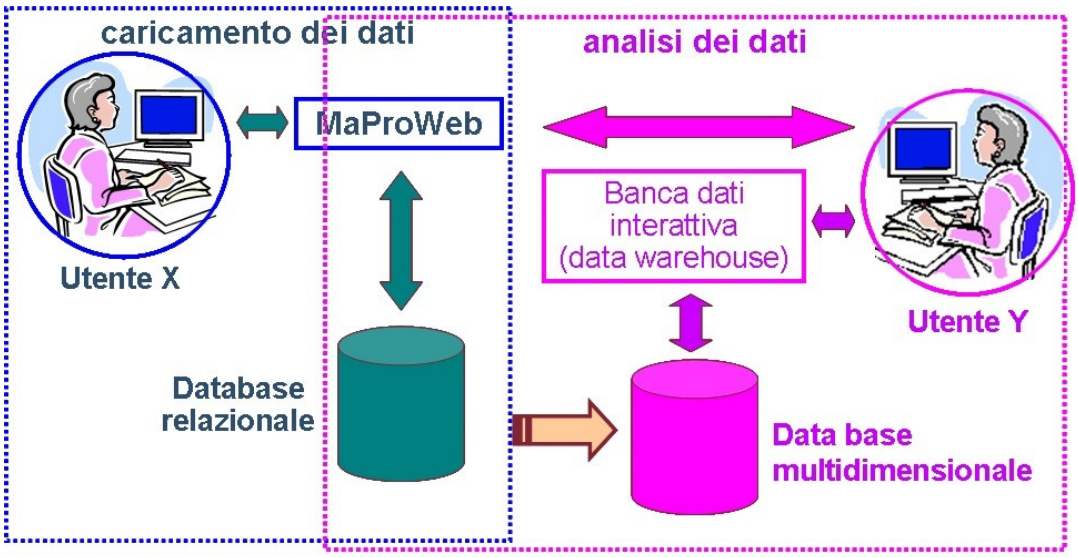

Figura 1: Alimentazione dati MaProWeb

Il database nazionale delle malattie professionali alimentato attraverso l'applicativo web based MaProWeb dagli operatori delle ASL, consente a questi ultimi, di inserire i dati riguardanti le malattie professionali, sia quelle denunciate presso l'INAIL che quelle segnalate dal lavoratore o rilevate dai medici ma non denunciate presso l'INAIL.

Il sistema MALPROF si inserisce tra i sistemi di sorveglianza epidemiologica e di ricerca delle malattie professionali e segue pertanto la logica di favorire il più possibile l'emersione delle cosiddette malattie professionali "perdute" registrando tutte le patologie segnalate come "correlate al lavoro", senza effettuare alcun tipo di filtro a priori sulle segnalazioni pervenute o acquisite e applicando criteri di attribuzione del nesso di causa tra esposizione professionale e malattia meno restrittivi rispetto a quelli seguiti dall'INAIL o dall'Autorità Giudiziaria.

2.2. La base dati

Per la realizzazione della base dati MaProWeb è stato utilizzato il DBMS SQLServer 2003, progettato per consentire la profilazione degli utenti. Il database contiene, oltre alle tabelle principali, svariate tabelle descrittive e tabelle operative funzionali, necessarie alla gestione del DB ed all'implementazione delle funzionalità dell'applicazione. Le tabelle principali, sono costituite dai dati inseriti dagli operatori territoriali delle ASL, e sono quelle analizzate in questo studio, e da cui sono stati estratti i dati per le nostre elaborazioni.

Le principali tabelle che compongono il DB MaProWeb sono:

- *Aziende:* contenente i dati di tutte le aziende anche quelle non più attive;
- *Malattie:* contenente le malattie rilevate dai servizi di prevenzione territoriali ASL;
- *Lavoratori:* contenente i soggetti per i quali è stata riscontrata una patologia;
- *Nessi:* contenente i dati che esprimono l'esistenza del nesso di causalità con l'attività professionale;
- *MansioniAttività:* contenente lo storico delle attività lavorative svolte da ciascun lavoratore.

Di seguito viene mostrato il diagramma ER delle principali tabelle della base dati MaProWeb:

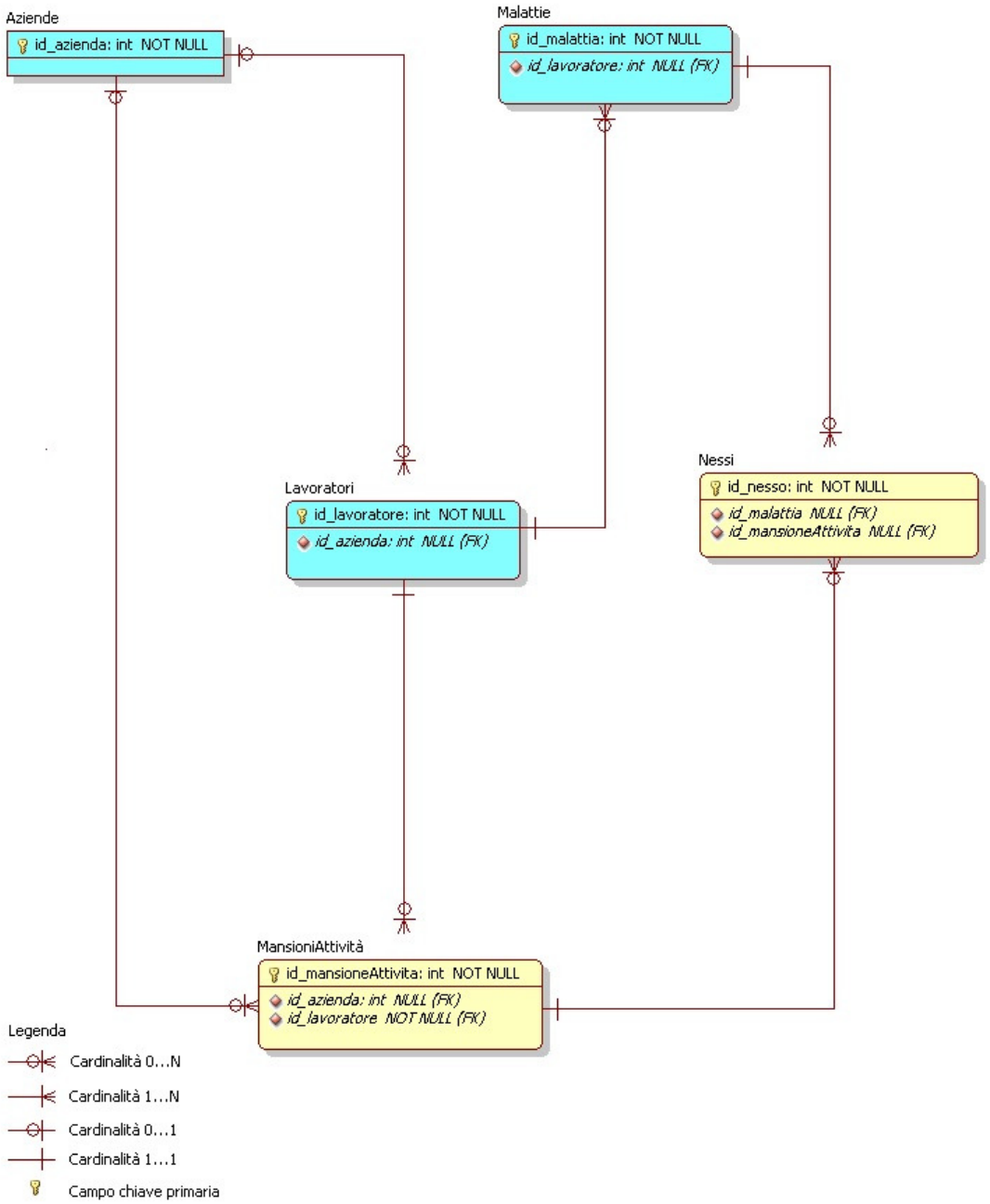

Figura 2: Diagramma Entità-Relazioni MaProWeb

Le tabelle *Aziende*, *Malattie* e *Lavoratori* sono quelle contenenti i dati principali ed oggetto della presente ricerca, le tabelle, *MansioniAttivita* e *Nessi* sono di relazione e legano le mansioni dei lavoratori con le malattie attraverso i nessi.

Di seguito, vengono dettagliate le suddette tabelle attraverso la presentazione dei loro tracciati record in cui si evidenziano i nomi dei campi, una breve descrizione per i campi più rilevanti, il loro relativo intervallo di valori ed il tipo di dato utilizzato all'interno del DBMS.

Tabella 1: Tabella *"Lavoratori"*

Nome campo	Tipo dato	Breve descrizione	Intervallo di valori
id_lavoratore (PK)	int	Non nullo. Chiave primaria intero progressivo.	Numeri interi
cognome	nvarchar(100)		
nome	nvarchar (100)		
id_cittadinanza	nvarchar (100)		
codice Fiscale	nvarchar (100)		
sesso	nvarchar (1)	Genere del lavoratore: M (maschio) F (Femmina).	2 possibili valori codificabili con 0, 1.
comune Nascita	nvarchar (100)		
id_statoNascita	nvarchar (100)	Tipo testo: stato di nascita criptato.	
data Nascita	nvarchar (100)		
id_qualificaAssicurativa	nvarchar (2)		
id_qualificaProfessionale	nvarchar (3)		
Telefono	nvarchar (100)		
Decesso	nvarchar (1)	Indica se il lavoratore è attualmente in vita (N) o no (S).	2 possibili valori codificabile con 0, 1
data Decesso	nvarchar (100)		
id_asl	nvarchar (4)	Codice Asl di residenza del lavoratore. Può essere Null	
codzonaASL	nvarchar (6)		
indirizzo Residenza	nvarchar (100)		
capResidenza	nvarchar (100)		
id_provinciaResidenza	nvarchar (100)		
id_comuneResidenza	nvarchar (100)		
indirizzo Domicilio	nvarchar (100)		
capDomicilio	nvarchar (100)		
id_provinciaDomicilio	nvarchar (100)		
id_comuneDomicilio	nvarchar (100)		
id_azienda (FK)	int	Codice dell'azienda relativo all'ultima attività lavorativa.	
inizio Data	datetime	Può essere Null	
id_mansionePrimaria	nvarchar (1)	Codice istat della mansione primaria al momento della segnalazione. Può non essere presente.	
id_mansioneSpecifica	nvarchar (4)	Codice Istat della mansione specifica in base alla primaria, al momento della segnalazione. Può non essere presente.	
id_ateco	nvarchar (5)	Attività economica dell'azienda in cui il lavoratore presta (o ha prestato) la sua attività al momento della segnalazione. Può non essere presente.	
Code	nvarchar (55)	Chiave anonima univoca associata al lavoratore. In tale codice sono contenuti i seguenti dati in chiaro: anno di nascita (posizione 33 - 36); sesso (posizione 37).	Estraibile l'anno di nascita ed età calcolabile. Si considera l'intervallo di valori [15 - 75].

I dati anagrafici dei lavoratori, per motivi di privacy, sono stati inseriti utilizzando un algoritmo di criptatura e vengono decriptati solo per chi è autorizzato alla loro lettura, ossia in base all'identificativo di zona ASL. Sono presenti i riferimenti all' ultima azienda presso cui il lavoratore ha svolto, o svolge, la sua attività lavorativa e sono presenti i riferimenti alla mansione del lavoratore al momento della sua denuncia. Infine è presente un campo denominato *Code* che ha lo scopo di identificare univocamente il lavoratore affinché, utenti di zone ASL differenti o con ruoli diversi, potessero consultare i dati relativi alle malattie professionali senza intaccare la privacy dei lavoratori affetti dalle malattie. Infatti, nel caso in cui l'utente sia di tipo "Amministratore", in luogo del cognome e del nome del lavoratore sarà visualizzato il suo codice unico. Dal codice code è possibile leggere in chiaro solo l'anno di nascita ed il sesso del lavoratore.

Tabella 2: Tabella "*Aziende"*

Nome campo	Tipo dato	Breve descrizione	Intervallo di valori
id_azienda (PK)	Int	Valore non nullo. Chiave primaria di tipo intero progressivo	Numeri Interi
id_ateco	nvarchar (5)	Codice ateco dell'attività economica dell'azienda.	Può essere nullo
partitiva	nvarchar (16)		
ragionesociale	nvarchar (80)		
indirizzo	nvarchar (50)		
CAP	nvarchar (5)		
id_provincia	nvarchar (3)		
id_comune	nvarchar (6)		
id_reg	nvarchar (2)		
telefono	nvarchar (25)		
fax	nvarchar (25)		
riferimento	nvarchar (25)		
id_asl	nvarchar (4)	Codice Asl rispetto alla sede aziendale. Può essere nullo.	
codzonaAsl	nvarchar (6)		
impiegati	Float		
operai	Float		
id_statovisura	nvarchar (1)		
id_naturagiuridica	nvarchar (2)		
frazione	nvarchar (30)		

Nella tabella *"Aziende"* sono presenti tutte le aziende relative a tutti i lavoratori, anche quelle che non sono più in attività, ma sono comunque indispensabili e utili per ricavare le attività economiche di cui si sono occupate al fine di poter avere una anamnesi lavorativa il più completa possibile.

Tabella 3: Tabella *"Malattie"*

Nome campo	Tipo dato	Breve descrizione	Intervallo di valori
Id_malattia (PK)	int	Chiave primaria. Intero progressivo	Interi non nulli.
Id_lavoratore (FK)	int	Chiave esterna del lavoratore.	Anche valori nulli
dataRegistrazione	datetime	Data di inserimento nel db da SW.	Non nullo
dataSegnalazioneAG	datetime	Data in cui la ASL segnala all'Autorità Giudiziaria (potrebbe essere l'esito di un indagine, oppure la trasmissione della denuncia tale e quale). In certi casi l'AG potrebbe essere la stessa ASL	
Note	nvarchar (255)	Alcune note inserite dal medico competente	
Id_fonteInformativa	nvarchar (1)	La fonte informativa della malattia.	
altraFonte	nvarchar (50)	Se non prevista nelle tabelle descrittive.	
denunciante	nvarchar (50)		
id_qualitaInfo	nvarchar (1)	Qualità dell'anamnesi.	
id_qualitaDiagnosi	nvarchar (1)	Qualità della diagnosi	
nrReferto	nvarchar (10)		
dataDiagnosi	datetime		
dataCertificato	datetime	Data di rilascio del certificato medico.	Non nullo
diagnosi	nvarchar (255)	Descrizione testuale della diagnosi. Da inserire se non presente il campo successivo.	
id_diagnosiICDIX	nvarchar (5)	Codice internazionale ICDIX della malattia. Da inserire se non è presente il campo precedente.	
id_patologiaConcomitante1	nvarchar (5)		
id_patologiaConcomitante2	nvarchar (5)		
id_tipoNessoGlobale	nvarchar (1)	Rappresenta il nesso globale tra l'attività lavorativa e la malattia del lavoratore in considerazione anche eventualmente dei nessi parziali.	
id_storicoICDIX	nvarchar (5)		
id_asl	nvarchar (4)		
codzonaASL	nvarchar (6)		
id_gruppoEta	nvarchar (2)		
cod_gravitamal	nvarchar (2)		
id_condizioneProfessionale	nvarchar (1)	Condizione professionale al momento della denuncia.	

La tabella *"Malattie"*, unitamente a quella dei lavoratori, è senz'altro la più interessante della base dati MaProWeb. In essa, sono contenute tutte le malattie legate alle attività lavorative "denunciate" dai lavoratori ed inserite dagli operatori ASL, che possano avere o non avere un nesso causale con la professione svolta da un dato lavoratore. Ogni lavoratore può essere ritrovato per una o più malattie.

Tabella 4: Tabella *"mansioniAttivita"*

Nome campo	Tipo dato	Breve descrizione	Intervallo di valori
id_mansioneAttivita (PK)	Int	Chiave primaria di tipo intero progressivo	Valore non nullo Numeri Interi
id_lavoratore (FK)	Int	Chiave esterna del lavoratore.	Può assumere valori nulli
id_azienda (FK)	Int	Chiave esterna dell'azienda	
dataDa	datetime	Data di inzio mansione	
dataA	datetime	Data di fine mansione	
durataAnni	Int	Durata in anni della mansione svolta	
id_mansionePrimaria	nvarchar (1)	Codice istat della mansione primaria.	
id_mansioneSpecifica	nvarchar (4)	Codice istat della mansione specifica in base alla primaria.	
id_ateco	nvarchar (5)	Attività economica dell'azienda in cui il lavoratore ha svolto la sua attività.	
Id_qualificaProfessionale	nvarchar (3)	Codice della qualifica assicurativa del lavoratore (ad esempio: operaio specializzato, operaio comune, apprendista artigiano, dipendente a tempo determinato, etc.).	
Id_qualificaAssicurativa	nvarchar (2)	Codice della qualifica professionale Istat del lavoratore (ad esempio: barista, bidello, insegnante etc.).	

La tabella delle mansioni e delle attività è una tabella di legame tra le aziende ed i lavoratori. Infatti attraverso di essa è possibile ricostruire l'anamnesi lavorativa di ciascun lavoratore.

Tabella 5: Tabella *"Nessi"*

Nome campo	Tipo dato	Breve descrizione	Intervallo di valori
id_Nesso (PK)	int	Chiave primaria di tipo intero progressivo	Valore non nullo Numeri Interi
id_tipoNesso	nvarchar (1)	Uno dei 4 valori definiti nella tabella successiva.	
id_malattia (FK)	bigint	Chiave esterna della malattia	
id_mansioneAttivita (FK)	bigint	Chiave esterna della mansione o attività	
note	nvarchar (255)	Campo testo in cui il medico ha la possibilità di inserire alcuni commenti	

La tabella dei nessi, è anch'essa una tabella di unione, in questo caso tra le malattie e le mansioni professionali di cui si sono occupati i lavoratori nel corso della loro storia lavorativa. In essa, viene espressa la valutazione da parte di un medico competente addetto alla prevenzione e sicurezza sul lavoro, riguardo al nesso causale tra una patologia ed un'attività lavorativa. Il campo *id_tipoNesso*, contiene il valore che esprime la valutazione del medico del lavoro in una scala di 4 valori.

Tra le tante tabelle descrittive ed operative costituenti il DB di MaProWeb, ritroviamo anche quelle utilizzate per scopi di decodifica. Di seguito citiamo alcune tra le più rilevanti di queste tabelle.

Tabella 6: Tabella di decodifica nessi

id_tipoNesso	Descrizione
1	ALTAMENTE PROBABILE
2	PROBABILE
3	IMPROBABILE
4	ALTAMENTE IMPROBABILE

In questa tabella ritroviamo le descrizioni delle valutazioni possibili utilizzabili da parte dei medici di prevenzione per l'associazione del nesso causale i cui valori vengono inseriti nella tabella *Nessi*.

Tabella 7: Tabella di decodifica fonti informative

id_fonteInformativa	Descrizione
A	SERVIZI di COMPETENZA ASL
B	MEDICI COMPETENTI D'AZIENDA
C	ISTITUTI UNIVERSITARI. M.D.L.
D	OSPEDALI
E	MEDICI DI BASE
F	MEDICI SPECIALISTI
G	PATRONATI
H	INAIL
I	ISPETTORATO DEL LAVORO
L	AUTORITA' GIUDIZIARIA
M	ALTRO

Nella tabella sono indicati tutti i possibili valori del campo *id_fonteInformativa* presente nella tabella *Malattie*, e rappresentano tutte le possibili fonti di informazione riguardo alle malattie professionali inserite nella tabella, ossia chi ha prodotto una documentazione riguardo alla malattia di un dato lavoratore.

Tabella 8: Tabella di decodifica per la qualità delle informazioni lavoro/malattia

id_qualitaInfo	Descrizione
1	INADEGUATA
2	INCOMPLETA
3	SUFFICIENTE
4	COMPLETA

In tale tabella sono indicati tutti i possibili valori del campo *id_qualitaInfo* presente nella tabella *Malattie*, e rappresentano i giudizi riguardo alla completezza globale delle informazioni riguardo alla qualità dell'anamnesi lavorativa.

Tabella 9: Tabella di decodifica per la qualità della diagnosi

id_qualitaDiagnosi	Descrizione
A	AFFIDABILE
D	DUBBIA
N	QUADRO NON DIAGNOSTICO

In questa tabella sono indicati tutti i possibili valori del campo *id_qualitaDiagnosi* presente nella tabella *Malattie*, e rappresentano i possibili valori riguardo alla qualità della diagnosi.

Tabella 10: Tabella di decodifica per le condizioni professionali

id_condizioneProfessionale	Descrizione
0	OCCUPATO
1	IN CERCA DI OCCUPAZIONE
4	PENSIONATO
5	ALTRE CONDIZIONI PROFESSIONALI
6	DECEDUTO
7	DISOCCUPATO
8	PREASSUNZIONE

In questa tabella sono indicati tutti i possibili valori del campo *id_condizioneProfessionale* presente nella tabella *Malattie*, e rappresentano i possibili valori dello stato dell'attività lavorativa del lavoratore al momento della denuncia.

2.3. I Dataset di analisi

In collaborazione con il settore ricerca dell'INAIL si è proceduto ad un'analisi approfondita dei dati ed all'estrazione dei dataset necessari allo scopo del presente studio di ricerca. In questo paragrafo viene descritta in dettaglio quella che è stata la fase di preprocessamento dei dati. Tale fase ha condotto alla creazione dei dataset di sviluppo, contenenti esclusivamente gli elementi utili, denominati *pattern,* costituiti esclusivamente dalle variabili individuate per l'analisi e costituenti la base su cui si poggia il presente studio. Successivamente è iniziata la realizzazione delle altre due fasi, quella di cluster analysis e degli algoritmi genetici attraverso la progettazione ed implementazione del software dell'applicativo per l'analisi delle malattie lavoro-correlate. Sono stati elaborati essenzialmente due principali dataset di sviluppo, ciascuno strutturato ad hoc per ognuna delle due fasi considerate propedeutiche tra loro.

L'attività di preprocessamento dei dati si è sviluppata attraverso l'analisi dei database ottenuti dal sistema MalProf dell'Inail allo scopo di:

- individuare il sottoinsieme di interesse dei casi costituenti i dataset di input all'applicazione da sviluppare;
- individuare le variabili più significative;
- eliminare quei record che sulle variabili più significative hanno presentato dati ambigui e/o incompleti.

Nella base dati esaminata, sono stati ricercati ed analizzati i dati riguardanti le malattie professionali segnalate dai lavoratori nel decennio 1999 - 2009. Allo scopo di ridurre i tempi di analisi e di elaborazione dei dati estrapolati, facilitando le procedure di calcolo, massimizzando ed accelerando le fasi di test sui dati, è stato considerato il sottoinsieme di dati riguardanti la Regione Lombardia in quanto contenente un numero significativo di casi delle malattie professionali contenute nel DB nazionale.

Sono state individuate le caratteristiche dei lavoratori più significative allo scopo di individuare le variabili da utilizzare nelle elaborazioni, e successivamente, sono stati eliminati i record che su tali variabili presentavano dati ambigui e/o incompleti. Inoltre, è stato condotto uno studio ad hoc, per la definizione delle metriche più appropriate per quelle variabili non numerabili ricavate dai campi *id_mansioneSpecifica* ed *id_ateco*.

Dalle tabelle principali e di decodifica di MaProWeb, è possibile osservare che i dati contenuti nel DB, presentano alcune particolarità:

- eterogeneità;
- diversa semantica;
- dati nulli o non significativi;
- dati da calcolare o da estrapolare.

I dati sensibili dei lavoratori, per il rispetto della legge sulla privacy, sono stati inseriti nel DB MaProWeb utilizzando un algoritmo di criptatura e pertanto non sono direttamente accessibili. Una delle attività condotte è consistita, pertanto, nell'analisi dei dati allo scopo di ricercare ed individuare quelli accessibili senza restrizioni.

Nella tabella *lavoratori*, uno dei dati individuati per l'estrazione, è stato l'anno di nascita del lavoratore estrapolato dal dato memorizzato nel campo *code*, una stringa alfanumerica, la cui codifica contiene l'anno di nascita in chiaro del lavoratore. L'estrazione di tale dato è stata condotta elaborando la stringa del campo *code* attraverso l'utilizzo del linguaggio di interrogazione delle basi di dati SQL. Dall'estrapolazione del dato *"anno di nascita"* del lavoratore, si è riscontrato che, in alcuni casi, il dato era palesemente errato, come ad esempio *anno = 1062*, oppure *anno = XXXX*, e così via. Per tali casi, si è dovuto valutare una modalità di trattamento dei dati errati o mancanti, valutando se fosse stato possibile procedere al loro reintegro o sostituzione. Avendo riscontrato che, nella gran parte dei casi, non era possibile reintegrare il dato, si è deciso di procedere con l'esclusione dell'intero record. Il numero totale di record per la tabella *lavoratori* per la sola regione Lombardia inizialmente, era pari a 39804; al termine di tale attività i record con età non valida ritrovati sono stati pari a 3087 ed i record con un età valida compresa tra i 18 e i 75 anni sono stati pari a 36717.

La tabella *mansioniAttivita* di partenza del DB nazionale, conteneva 96436 record, procedendo alla sua analisi per l'estrazione dei dati relativi ai lavoratori della Lombardia, si sono riscontrate incompletezze e inesattezze per i valori di alcuni campi. In particolare il campo *id_lavoratore* ha presentato un valore pari a *null* per 147 record e per gli stessi record, anche i valori di individuazione di inserimento territoriale del campo *usr_codZonaAsl* si sono rilevati nulli, pertanto non è stato possibile recuperare i valori mancanti neanche attraverso le zone territoriali di competenza che hanno inserito i dati. Il numero di record della tabella *mansioniAttivita* corrispondenti ai lavoratori della Lombardia è stato rilevato pari a 33964. Il campo *id_mansioneSpecifica* si è rilevato con un valore pari a *null* per 5552 record, ed inoltre, ha presentato valori non classificati come ad esempio *7l43* (sette-elle-quattro-tre), oppure semplicemente *77* (sette-sette).

In questa fase di analisi solo sulle tabelle *lavoratori* e *mansioniAttivita*, sono state fatte alcune assunzioni per la determinazione dei dati da estrarre:

- si sono considerati solo i lavoratori che al momento della denuncia della malattia avevano tra 18 e 75 anni (quindi si considerati i campi *Malattie.dataCertificato* e *Lavoratori.code(anno di nascita)*);
- si sono considerati i lavoratori con un solo periodo lavorativo;
- si sono considerati i casi con una sola malattia per un solo periodo lavorativo, quindi sono stati esclusi tutti i casi con più denunce di malattie per uno stesso periodo lavorativo;
- si sono considerati i casi in cui la mansione specifica dei lavoratori della tabella *mansioniAttivita* era non nulla e contenente valori significativi.

Al termine delle elaborazioni, i record utilizzabili della tabella *lavoratori*, sono risultati essere pari a 13767.

In una fase successiva di analisi sono state prese in considerazione anche le tabelle *malattie* ed *lk_mansioniSpecifiche* (tabella descrittiva delle mansioni). Nella tabella *mansioniAttivita* i campi *dataDa* e *dataA* hanno presentato dei valori nulli oppure in contrasto tra loro od in contrasto con quelli presenti nel campo della data del certificato *dataCertificato* della tabella *malattie*. Come specificato nella descrizione della tabella *mansioniAttivita*, i campi *dataDa* e *dataA* rappresentano rispettivamente la data di inizio mansione e la data di fine mansione di un lavoratore, di seguito i casi in contrasto riscontrati:

1. *dataCertificato* < *dataDA*;
2. *dataCertificato* = *dataDA*
3. *dataCertificato* < *dataA;* oppure *dataCertificato* > *dataDA*;
4. *dataA* non presente (ancora in attività) e *dataCertificato* presente;
5. *dataCertificato* = *dataA;*
6. *dataCertificato* > *dataA*;
7. *dataA* < *dataDA;*
8. *dataA* = *dataDA;*
9. *dataDA* = *null;*

Le durate delle attività presenti nella tabella *mansioniAttivita*, vengono rappresentate in anni e sono state ricalcolate in mesi allo scopo di avere una maggiore precisione nella fase delle nostre elaborazioni.
Altri casi particolari sono stati riscontrati sulle date di inizio e fine mansione rispetto all'età del lavoratore, come evidenziato di seguito:

1. *dataDA* < *dataNascita;*
2. *dataDA* = *dataNascita;*
3. *dataDA - dataNascita = età in [1, 17].*

in questi casi è possibile osservare che la data di inizio mansione è antecedente all'anno di nascita del lavoratore o uguale e che la differenza tra l'anno di inizio mansione e quello di nascita del lavoratore è compresa in un intervallo tra 1 e 17.

Sono stati riscontrati molti altri casi particolari tra cui ad esempio, lavoratori di età pari a 80 anni e che avevano la *dataA* nulla come se fossero ancora in attività; o ancora lavoratori che avevano un'età di inizio attività maggiore o uguale a 60 anni; ed altri casi simili. Dalle ulteriori analisi ed elaborazioni eseguite sulla base dati ed introducendo la condizione che i record ricadenti nelle situazioni precedenti vengano esclusi il numero di record ottenuto è stato pari a 3558.

Dalle ulteriori analisi si è constatato che i casi in cui l'età di inizio attività del lavoratore era di 15 anni, erano abbastanza frequenti, pertanto si è ampliato l'intervallo di età dei lavoratori alla data del certificato medico portandolo da [18, 75] a [15, 75] anni; in tal caso la differenza *dataDA - annoNascita* è stata compresa nell'intervallo [1, 14], e quindi sono stati esclusi tutti i record ricadenti in tale intervallo. Si è analizzato inoltre, il valore del campo *id_diagnosiICDIX* della tabella *malattie* per il quale si è constatata la presenza di valori nulli, in corrispondenza dei quali sono stati esclusi i relativi record. Inoltre per la durata dell'attività sono state fatte le seguenti assunzioni per la data di fine attività:

se *dataA* >= *dataCert*	allora	*dataA* = *dataCert*
se *dataA* = *null*	allora	*dataA* = *dataCert*
se *dataA* < *dataDA*	allora	*dataA* = *dataCert*
se *dataA* <= *dataCert*	allora	*dataA* = *dataA*

Al termine delle elaborazioni, il numero dei record considerati è stato pari a 4462. Nella tabella *mansioniAttivita*, il campo *id_ateco* dell'attività economica dell'azienda in cui il lavoratore ha svolto la sua attività, si è rilevato contenere alcuni valori non validi come: valori *nulli*, casi di stringa vuota o altri casi non significativi; analogamente per il campo *sesso* della tabella *lavoratori.* I record contenenti tali dati non sono stati presi in considerazione. In particolare, per il campo *id_ateco*, prima di eliminare i record contenenti i dati non validi, si è provato a recuperarli sostituendoli con delle stringhe di 5 byte contenenti tutti *'0'*, solo dopo aver valutato i risultati ottenuti dai test eseguiti sul dataset con i nuovi valori, si è giunti alla decisione di non considerare i record con i dati non validi. Il sottoinsieme finale dei dati ottenuto è stato pari a 3427 record.

Riassumendo, il sottoinsieme finale è stato individuato considerando:

- solo i casi della regione italiana Lombardia di malattie professionali archiviate nel decennio 1999 - 2009;
- solo i casi di lavoratori con una sola malattia ed un solo periodo lavorativo (allo scopo di semplificare le strutture dei pattern);
- solo i casi contenenti dati non ambigui, consistenti e completi.

La distribuzione delle malattie, sulle quali sono state eseguite le elaborazioni, è mostrata nella tabella seguente:

Tabella 11: Distribuzione delle malattie contenute nel DB MaProWeb

Malattia	N. records	N. Cumulativo records	Frequenza	Frequenza Cumulativa
Sordità	1493	1493	0,436	0,436
Malattie del rachide	334	1827	0,097	0,533
Malattie muscoloscheletriche (escluso rachide)	288	2115	0,084	0,617
Tumori pleura e peritoneo	232	2347	0,068	0,685
Sindrome tunnel carpale	199	2546	0,058	0,743
Malattie della pelle	176	2722	0,051	0,794
Disturbi dell'orecchio (escluso sordità)	137	2859	0,040	0,834
Malattie mentali	98	2957	0,029	0,863
Malattie delle vie respiratorie	76	3033	0,022	0,885
Altre malattie	394	3427	0,115	1

Sul dataset di 3427 record è stato applicato un ulteriore filtro, eliminando i record delle malattie professionali con una frequenza inferiore al 5%, ottenendo così, un dataset composto di 2722 record costituenti comunque circa l'80% dei 3427 casi.

La tabella *lavoratori*, è stata la tabella maggiormente considerata unitamente a quella delle *malattie* e delle attività professionali *mansioniAttivita*. E' stata prodotta una nuova tabella dei

lavoratori contenente i nuovi dati calcolati ed i dati già disponibili nella tabella originale. Nella fase di pre-processamento e di pre-elaborazione dei dati, è stato individuato il seguente gruppo di variabili:

1. Il *genere* (campo *sesso*)
2. La *mansione specifica* (campo *id_mansioneSpecifica*)
3. L'*ateco* (campo *id_ateco*)
4. L'*età* (età dei lavoratori alla data del certificato medico tra i 15 e 75 anni)

Tali variabili, sono state utilizzate per condurre una serie di test attraverso una prima versione prototipale del software. Dall'analisi dei risultati ottenuti da tali test sono state individuate ed introdotte ulteriori due variabili:

5. la *durata mansione*, calcolata in mesi;
6. l' *età inizio*, età del lavoratore all'inizio della mansione.

Si è ottenuto così il nuovo insieme di variabili con le quali sono state eseguite le elaborazioni. Nella tabella seguente è illustrato il nuovo set di variabili con la relativa codifica.

Tabella 12: Le variabili considerate

Codifica	Variabile	Tipo di dato
x1	*Età.* L'età del lavoratore alla data del certificato medico (anni)	numerico
x2	*Durata mansione*. La durata della mansione svolta dal lavoratore (mesi)	numerico
x3	*Età inizio*. L'età del lavoratore all'inizio della mansione (anni)	numerico
x4	*Genere*. Maschio o femmina (M/F)	categorico
x5	*Mansione specifica*. La mansione lavorativa svolta dal lavoratore (codifica attraverso una coppia di caratteri del codice Istat)	categorico
x6	*Ateco*. L'attività economica dell'azienda (codifica attraverso una coppia di caratteri del codice Ateco Istat)	categorico

Le ultime tre variabili *x4*, *x5* e *x6* sono direttamente disponibili nel DB mentre le prime tre *x1*, *x2* e *x3* sono state calcolate attraverso opportune query ed elaborazioni.

Come si vedrà in seguito, per la fase di elaborazione inerente gli algoritmi genetici sono stati sviluppati dei nuovi dataset partendo dall'ultimo dataset ottenuto di 2722 record, in particolare sono stati elaborati tre dataset distinti definiti come *Training Set*, *Validation Set* e *Test Set*.

2.4. Il Dataset nazionale

Per quanto riguarda il dataset dei dati nazionali, l'attività di pre-processamento è stata analoga a quella per il dataset della regione Lombardia.

Il DB nazionale, dopo la fase più consistente di cleaning dei dati, ha presentato un numero di record pari a 11128. E' stato ulteriormente sottoposto ad una attività di cleaning per i seguenti casi:

1. Data del certificato < Data inizio attività
2. Codice malattia esistente nei gruppi di malattie.

Il numero di record ottenuto è stato pari a 11064. Un'ulteriore riduzione del dataset, è stata ottenuta, considerando i record corrispondenti alle stesse 6 malattie considerate per il dataset regionale della Lombardia elencate nella Tabella 11, ossia: *sordità, malattie del rachide, malattie muscolo scheletriche, tumori pleura, sindrome tunnel carpale e malattie della pelle* ottenendo così un dataset di 9676 record rappresentante circa l' 87% dei casi. Nella tabella successiva è mostrata la distribuzione di tali malattie.

Tabella 13: Distribuzione delle 6 malattie considerate nel DB MaProWeb nazionale

Malattia	N. record	N. Cumulativo record
Sordità	2973	2973
Malattie del rachide	1816	4789
Malattie muscoloscheletriche (escluso rachide)	2661	7450
Tumori pleura e peritoneo	601	8051
Sindrome tunnel carpale	1313	9364
Malattie della pelle	312	9676
Altre malattie	1388	11064

2.5. Descrizione delle variabili

Le variabili individuate mostrate nella Tabella 12, sono di due tipi: quantitative e nominali. Le variabili quantitative sono le *età* (*x1* e *x3*), la *durata della mansione* (*x2*) ed il *genere* (*x4*); le variabili nominali sono il *codice mansione specifica* e il codice dell'attività economica aziendale *ateco* (*x5* e *x6*).

Le variabili *età x1* e *x3* sono rappresentate in anni. I loro valori sono il risultato di una approssimazione per eccesso delle frazioni di mesi e giorni dei valori calcolati a partire dal valore dell'anno contenuto nel db iniziale; ed appartengono all'insieme $\{x \in \aleph \mid x \in [15,75]\}$.

La variabile *durata della mansione x2* rappresenta i mesi lavorati ed è tale che $x2 \in \aleph$, i valori di *x2* sono un'approssimazione per difetto delle frazioni dei giorni lavorati.

La variabile *genere x4* è dicotomica e può assumere solo i due valori corrispondenti a *maschio* o *femmina* che possono essere rappresentati o con due lettere dell'alfabeto, rispettivamente *M* e *F,* oppure con due numeri come *0* e *1*.

La codifica della mansione lavorativa *mansione specifica (x5)* e del codice dell'attività economica aziendale *ateco (x6)* sono di tipo testo e sono rispettivamente il codice ISTAT della classificazione delle professioni basato sulla versione italiana del sistema di classificazione internazionale ISCO (International Standard Classification of Occupations), ed il codice di classificazione internazionale delle attività economiche adottata dall'ISTAT basato sulla versione italiana del sistema di classificazione NACE (Nomenclature des Activités Économiques dans la Communauté Européenne) rispettivamente di 4 e 5 byte delle classificazioni ISTAT 1991, (le nuove classificazioni per le professioni CP2011 arrivano fino a 5 byte). Per una migliore successiva comprensione nel calcolo delle distanze per queste due variabili, di seguito vengono descritto in maggior dettaglio le codifiche di *x5* e *x6.*

I dati riguardanti la classificazione delle professioni contenuti nel DB MalProf sono nella classificazione ISTAT 1991 che recepisce tutte le novità introdotte nella versione della International Standard Classification of Occupations ISCO88. Nel corso degli anni, presso l'ISTAT si sono succeduti vari aggiornamenti e a partire dal 2011 è stata adottata la nuova classificazione delle professioni CP2011, frutto di un lavoro di aggiornamento delle precedenti versioni e di adattamento alle novità introdotte dalla International Standard Classification of Occupations – ISCO08. Pertanto presso l'INAIL è stato avviato un lavoro di adeguamento delle classificazioni delle professioni alla nuova classificazione CP2011, che al momento di inizio del presente lavoro, non era stato ancora terminato e pertanto, è stata considerata la classificazione 1991 comunque non pregiudizievole nelle elaborazioni e nei calcoli eseguiti con la cluster analysis e nella successiva applicazione dell'algoritmo genetico. Infatti, la logica su cui si basa la CP2011 è la stessa utilizzata dalla precedente classificazione CP1991 pur avendo introdotto un livello gerarchico in più, passando da 4 a 5.

La classificazione fornisce uno strumento per ricondurre tutte le professioni esistenti nel mercato del lavoro all'interno di un numero limitato di raggruppamenti professionali, da utilizzare per comunicare, diffondere e scambiare dati statistici e amministrativi sulle professioni, comparabili a livello internazionale; tale strumento non deve invece essere inteso come uno strumento di regolamentazione delle professioni.

L'oggetto della classificazione, la professione, è definito come un insieme di attività lavorative concretamente svolte da un individuo, che richiamano conoscenze, competenze, identità e statuti propri, il criterio fondante, è quello del livello e del campo di applicazione delle competenze richieste per eseguire in modo appropriato i compiti associati alla professione.

I cambiamenti introdotti dalla Classificazione 1991 recependo le novità introdotte nella nuova versione della International Standard Classification of Occupations - ISCO88, introduce così

una nuova logica classificatoria, utilizzata ancora oggi, basata sul livello e sul campo di applicazione delle competenze richiamate dall'esecuzione di quei compiti. La classificazione, in altre parole, si basa sul tipo di lavoro svolto, identificato attraverso le seguenti componenti: il livello di autonomia/responsabilità nei processi decisionali, la funzione espletata e l'area di specializzazione. I primi due, in particolare, orientano la dimensione gerarchica della classificazione, separando quelle che detengono un maggior livello di autonomia e di responsabilità nell'esercizio del lavoro e richiedono un più alto livello di qualificazione (in alto nella scala), da quelle che rispecchiano livelli inferiori (in basso nella scala).

Il nuovo assetto classificatorio 1991, attraverso il criterio della competenza delinea un sistema articolato su 4 livelli di aggregazione gerarchici:

- il primo livello, di massima sintesi, composto da 9 *Grandi gruppi professionali*;
- il secondo livello, comprensivo di 35 *gruppi professionali*;
- il terzo livello, con 119 classi professionali;
- il quarto livello, formato da 599 *categorie*

All'interno della struttura appena delineata trovano collocazione 6.319 voci professionali suddivise come risulta dalla successiva Tabella 14. Rispetto alla precedente classificazione il mutamento dei criteri classificatori ha determinato, da una parte, un notevole restringimento del numero dei raggruppamenti presenti nei livelli intermedi e, dall'altra, la creazione di un nuovo livello di dettaglio, il quarto digit, in grado di distinguere maggiormente le professioni tra di loro.

Tabella 14: Suddivisione per gruppi, classi, professioni elementari e voci.

GRANDI GRUPPI	PROFESSIONI			
	Gruppi	**Classi**	**Professioni elementari**	**Voci**
l. Legislatori, dirigenti e imprenditori	2	6	34	183
2. Professioni intellettuali, scientifiche e di elevata specializzazione	6	20	105	774
3. Professioni intermedie (tecnici)	4	13	90	859
4. Professioni esecutive relative all'amministrazione e gestione	2	6	44	215
5. Professioni relative alle vendite ed ai servizi per le famiglie	5	15	63	557
6. Artigiani, operai specializzati e agricoltori	5	21	117	1743
7. Conduttori di impianti, operatori di macchinari fissi e mobili e operai di montaggio industriale	4	21	102	1438
8. Personale non qualificato	6	16	43	464
9. Forze armate	1	1	1	86

La logica utilizzata per aggregare professioni diverse all'interno di un medesimo raggruppamento si basa sul concetto di competenza, visto nella sua duplice dimensione del *livello e del campo delle competenze* richieste per l'esercizio della professione.

Il *livello di competenza* è definito in funzione della complessità, dell'estensione dei compiti svolti, del livello di responsabilità e di autonomia decisionale che caratterizza la professione; il *campo di competenza* coglie, invece, le differenze nei domini settoriali, negli ambiti disciplinari delle conoscenze applicate, nelle attrezzature utilizzate, nei materiali lavorati, nel tipo di bene prodotto o servizio erogato nell'ambito della professione.

I grandi gruppi, i gruppi, le classi e le professioni elementari sono contraddistinti da apposita numerazione decimale, che vale anche come numerazione convenzionale agli effetti della codificazione.

Nella classificazione le singole voci professionali sono dapprima raggruppate in professioni elementari; più professioni elementari costituiscono le classi; più classi compongono i gruppi ed infine più gruppi formano i grandi gruppi professionali.

Le singole voci elementari stanno a definire, talvolta con sinonimi, le molteplici attività individuali che possono considerarsi in linea di massima omogenee.

I codici adottati sono costituiti da quattro cifre, di cui la prima indica il grande gruppo, la seconda il gruppo, la terza la classe e la quarta la professione elementare.

Ad esempio, per la mansione lavorativa *mansione specifica* dovendo provvedere alla codifica della voce professionale “bioinformatico” si adotterà il codice 2.1.1.4 in cui la prima cifra “2” designa il grande gruppo (Professioni intellettuali, scientifiche e di elevata specializzazione), la seconda cifra “1” indica, nell'ambito del grande gruppo, il gruppo (Specialisti in scienze

matematiche, fisiche, naturali ed assimilati), la terza “1” indica la classe (Specialisti in scienze matematiche, fisiche e naturali) e la quarta “4” la professione elementare (Informatici e telematici), infine si considera la singola voce elementare “bioinformatico”, facente parte di un insieme di singole voci elementari di possibili attività individuali in linea di massima omogenee, come di seguito rappresentato:

2.1.1.4 = Informatici e telematici.

Grande gruppo: 2 - Professioni intellettuali, scientifiche e di elevata specializzazione

Gruppo: 1 - Specialisti in scienze matematiche, fisiche, naturali ed assimilati

Classe: 1 - Specialisti in scienze matematiche, fisiche e naturali

Professione elementare: 4 – Informatici e telematici

singole voci elementari: analista di procedure

analista di programmi

analista di sistemi

analista programmatore edp

bioelettronico (esperto biochips e biocomputers)

bioinformatico

cibernetico

ingegnere software

progettista di sistemi vocali

progettista sistemi elaborazioni voci ed immagini

programmatore spaziale

specialista in scienze dell'informazione

Come per il codice di classificazione delle professioni anche per la classificazione delle attività economiche ATECO, i dati contenuti nel DB MalProf sono nella classificazione ISTAT 1991, ed anche in questo caso, è da precisare che al momento di inizio di tale lavoro, l'aggiornamento alla nuova versione ATECO 2007 non era stato ancora terminato e pertanto, è stata considerata la classificazione 1991, sicuramente completa su tutto il DB MalProf e comunque non pregiudizievole nelle elaborazioni e nei calcoli eseguiti nella cluster analysis e nella successiva applicazione dell'algoritmo genetico. Infatti, la logica su cui si basa l'ATECO 2007 è la stessa utilizzata dalla precedente classificazione 1991.

La classificazione delle attività economiche ATECO (ATtività ECOnomiche) è una tipologia di classificazione per le rilevazioni statistiche nazionali di carattere economico.

È la traduzione italiana della Nomenclatura delle Attività Economiche nella Comunità Europea (NACE) creata dall'Eurostat, adattata dall'ISTAT alle caratteristiche specifiche del sistema economico italiano. Attualmente è in uso la versione ATECO 2007, entrata in vigore dal 1º gennaio

2008, che sostituisce la precedente ATECO 2002 aggiornamento della ATECO 1991. La codifica NACE è a quattro cifre, al fine di soddisfare le esigenze di un'informazione più dettagliata a livello nazionale, per le classificazioni economiche italiane, si può avere un ulteriore dettaglio attraverso l'aggiunta di una quinta cifra.

Il codice della classificazione Ateco 1991 presenta una struttura in cui le varie attività economiche sono raggruppate, dal generale al particolare, in sezioni, sottosezioni, divisioni, gruppi, classi e categorie.

Tabella 15: Struttura della classificazione Ateco

LIVELLI	**DENOMINAZIONE**	**TIPO DI CODICE**
1° livello	Sezioni	1 lettera maiuscola
intermedio	*Sottosezioni*	*2 lettere maiuscole*
2° livello	Divisioni	2 cifre
3° livello	Gruppi	3 cifre
4° livello	Classi	4 cifre
5° livello	Categorie	5 cifre

Due sezioni, "Estrazione di minerali" e "Attività manifatturiere", sono articolate in sottosezioni. Le sezioni e le sottosezioni a loro volta si articolano in divisioni, gruppi e classi; la maggior parte delle classi sono articolate in categorie le quali costituiscono le componenti elementari della classificazione.

Tabella 16: Classificazione Ateco 1991

SEZIONI e SOTTOSEZIONI	Sezioni	Sottosezion.	Divisioni	Gruppi	Classi	Categorie
A Agricoltura, caccia e silvicoltura	**1**	**-**	**2**	**6**	**14**	**35**
B Pesca, piscicoltura e servizi connessi	**1**	**-**	**1**	**1**	**3**	**5**
C Estrazione di minerali	**1**	**2**	**5**	**13**	**18**	**24**
CA Estrazione di minerali energetici	-	1	3	6	8	8
CB Estrazione di minerali non energetici	-	1	2	7	10	16
D Attività manifatturiere	**1**	**14**	**23**	**103**	**245**	**354**
DA Industrie alimentari, delle bevande e del tabacco	-	1	2	10	35	48
DB Industrie tessili e dell'abbigliamento	-	1	2	10	30	42
DC Industrie conciarie, fabbricazione di prodotti in cuoio, pelle e similari	-	1	1	3	3	5
DD Industria del legno e dei prodotti in legno	-	1	1	5	6	9
DE Fabbricazione della pasta-carta, della carta e del cartone, dei prodotti di carta; stampa ed editoria	-	1	2	5	20	20
DF Fabbricazione di coke, raffinerie di petrolio, trattamento dei combustibili nucleari	-	1	1	3	3	6
DG Fabbricazione di prodotti chimici e di fibre sintetiche e artificiali	-	1	1	7	20	26
DH Fabbricazione di articoli in gomma e materie plastiche	-	1	1	2	7	7
DI Fabbricazione di prodotti della lavorazione di minerali non metalliferi	-	1	1	8	25	29
DJ Metallurgia, fabbricazione di prodotti in metallo	-	1	2	12	33	46
DK Fabbricazione di macchine ed apparecchi meccanici	-	1	1	7	20	36
DL Fabbricazione di macchine elettriche e di apparecchiature elettriche, elettroniche ed ottiche	-	1	4	15	17	34
DM Fabbricazione di mezzi di trasporto	-	1	2	8	11	20
DN Altre industrie manifatturiere	-	1	2	8	15	26
E Produzione e distribuzione di energia elettrica, gas e acqua	**1**	**-**	**2**	**4**	**4**	**6**
F Costruzioni	**1**	**-**	**1**	**5**	**17**	**18**
G Commercio all'ingrosso e al dettaglio; riparazione di autoveicoli, motocicli e di beni personali e per la casa	**1**	**-**	**3**	**19**	**77**	**193**
H Alberghi e ristoranti	**1**	**-**	**1**	**5**	**9**	**22**
I Trasporti, magazzinaggio e comunicazioni	**1**	**-**	**5**	**14**	**23**	**30**
J Attività finanziarie	**1**	**-**	**3**	**5**	**12**	**18**
K Attività immobiliari, noleggio, informatica, ricerca, servizi alle imprese	**1**	**-**	**5**	**23**	**37**	**68**
L Amministrazione Pubblica	**1**	**-**	**1**	**3**	**10**	**20**
M Istruzione	**1**	**-**	**1**	**4**	**6**	**11**
N Sanità e assistenza sociale	**1**	**-**	**1**	**3**	**7**	**20**
O Altri servizi pubblici, sociali e personali	**1**	**-**	**4**	**12**	**28**	**48**
P Attività svolte da famiglie e convivenze	**1**	**-**	**1**	**1**	**1**	**1**
Q Organizzazioni ed organismi extraterritoriali	**1**	**-**	**1**	**1**	**1**	**1**
TOTALE	**17**	**16**	**60**	**222**	**512**	**874**

Si tratta di una classificazione alfa-numerica che si sviluppa su cinque livelli di dettaglio. Le *sezioni* e le *sottosezioni* sono contraddistinte da un codice alfabetico costituito, rispettivamente, da

una e da due lettere maiuscole, denominato codice di tabulazione ed utilizzato principalmente nella fase di diffusione dei dati statistici ed indicano il macro-settore di attività economica. Le *divisioni* (2 cifre), i *gruppi* (3 cifre), le *classi* (4 cifre) e le *categorie* (5 cifre) di attività economica sono contraddistinti da un codice numerico, indipendente dal codice alfabetico di tabulazione. Solamente il codice numerico vale come numerazione convenzionale in fase di codificazione. I numeri rappresentano, con diversi gradi di dettaglio, le articolazioni e le disaggregazioni dei settori stessi.

Ciascuna attività economica viene codificata generalmente con un numero di cinque cifre, l'ultima delle quali è separata da un punto dalle due precedenti, a loro volta separate da un punto dalle prime due.

Ciascun codice numerico incorpora i precedenti.

Un esempio:

Considerata l'attività economica DA 15.11.1

- *sezione:* D: Attività manifatturiere;
  - *sottosezione:* DA: Industrie alimentari, delle bevande e del tabacco;
    - *divisione:* 15: Industrie alimentari e delle bevande;
      - *gruppo:* 15.1: Produzione, lavorazione e conservazione di carne e di prodotti a base di carne;
        - *classe:* 15.11: Produzione, lavorazione e conservazione di carne, esclusi i volatili;
          - *categoria:* 15.11.1: Produzione di carne, non di volatili, e di prodotti della macellazione;
          - *categoria:* 15.11.2: Conservazione di carne, non di volatili, mediante congelamento e surgelazione.

2.6. Dati strutturati e vettori sezionati

Si definiscono *dati strutturati di prima specie*, gli aggregati di dati di natura omogenea la cui informazione strutturale risiede nell'ordinamento, più precisamente:

Si dice *dato strutturato di prima specie* un insieme ordinato di numeri reali in cui l'informazione strutturale risiede nell'ordine in cui compaiono.

Si definiscono *dati strutturati di seconda specie* gli aggregati di dati di natura non omogenea la cui informazione strutturale risiede nel differente significato dei diversi campi, più precisamente:

Si dice *dato strutturato di seconda specie* un insieme di campi di natura eterogenea (il cui ordinamento non reca un'informazione strutturale essenziale), contenenti dati non strutturati o dati strutturati di prima specie; l'informazione strutturale di tali dati risiede nel diverso significato associato ai diversi campi.

Ogni record (pattern) del dataset, può essere considerato un vettore i cui elementi sono le variabili introdotte precedentemente ed elencate nella Tabella 12, ed essendo di natura eterogenea, ogni pattern è un dato strutturato di seconda specie.

Si consideri la seguente definizione:

dicesi *vettore sezionato* una sequenza ordinata di M n-uple di valori appartenenti ad un insieme $V \subseteq \{\Re, \Im, T\}$ dove $\Re, \Im$ rappresentano rispettivamente gli insiemi dei numeri reali e immaginari e T è un insieme qualunque non ordinale costituito da un numero finito di elementi. Formalmente un elemento dell'insieme V è:

$$v = (v_0, v_1, ..., v_{M-1}) \qquad \text{dove } v_i \in V_{(i)}^{n(i)} \text{ con } i \in [0, M-1]$$

Essendo i vettori sezionati costituiti da elementi che sono a loro volta vettori di elementi di vario tipo; per definire in maniera completa un vettore sezionato è necessario definire anche il suo insieme di informazioni che definisce completamente il suo formato; così si parla di *struttura* di vettore sezionato. Tale insieme di informazioni è costituito da:

a) il *valore* di M (numero di sezioni);

b) *nome* di ciascuna sezione M a scopo identificativo (etichetta);

c) *tipo* di ciascuna sezione M (valore di V);

d) *dimensione* di ciascuna sezione M (valore di n di ciascuna n-upla);

Nel caso del punto c) se fossero presenti una o più sezioni di tipo T per ciascuna di esse deve essere data la definizione degli insiemi finiti coinvolti come elenco degli elementi.

Vediamo un esempio per chiarire meglio i concetti di vettore sezionato e di struttura:

a) *valore* di M = 4, (4 n-uple);

b) siano M1, M2, M3, M4 i *nomi* delle 4 sezioni

c) il *tipo* di ciascuna sezione:

M1, M2 di tipo $\Re$; M3 di tipo $\Im$; ed M4 di tipo T

d) e siano:

$n(0) = 3$;

$n(1) = 4$;

$n(2) = 2$;

$n(3) = 5$;

i valori delle dimensioni di ciascuna sezione.

Quindi avremo:

$v_0 \in \Re^3$; $v_1 \in \Re^4$; $v_2 \in \Im^2$; $v_3 \in T^5$; dove T è l'insieme dei caratteri dell'alfabeto occidentale minuscoli e maiuscoli.

$v_0 = (a_1, a_2, a_3,)$ *dove* $a_j \in \Re$; $\forall j = 1,2,3$

$v_1 = (b_1, b_2, b_3, b_4)$ *dove* $b_j \in \Re$; $\forall j = 1,2,3,4$

$v_2 = (c_1, c_2)$ *dove* $c_j \in \Im$; $\forall j = 1,2$

$v_3 = (A,b,D,F,g)$ *dove* $A,b,D,F,g \in \{x \in T \mid T$ *è l'insieme dei caratteri dell'alfabeto occidentale minuscoli e maiuscoli}*

Quindi un'istanza di vettore sezionato è:

$v = (vo,v1,v2,v3) = ((a1,a2,a3),(b1,b2,b3,b4),(c1,c2),(A,b,D,F,g))$

Si definisce la *Linearly Weighted Dissimilarity Measure (LWDM)* come misura di dissimilarità tra vettori sezionati nel modo seguente:

$$\mathrm{LWDM}(u,v) = \sum_{i=0}^{M-1} p_i D_i(u_i, v_i)$$

dove u e v, in base alla notazione precedente, sono due vettori sezionati di uguale struttura dove $p_i \in \Re$; $p_i \in [0,1]$ e dove D_i è una misura di dissimilarità nello spazio $V_{(i)}^{n(i)}$, e se ciascuna D_i è una metrica nello spazio $V_{(i)}^{n(i)}$ e se $p_i \neq 0,\ \forall i \in [0, M-1]$ anche LWDM risulta essere una metrica. In tal caso la misura prende semplicemente il nome *Linearly Weighted Distance (LWD)*, inoltre, se le D_i sono misure il cui valore è normalizzato tra 0 e 1, anche il valore della LWDM (o LWD) lo è.

I coefficienti p_i sono detti *pesi* delle diverse sezioni e permettono di attribuire maggiore o minore importanza a ciascuna sezione nel calcolo della dissimilarità tra due vettori e possono essere considerati in due modi: con il vincolo tale che $\sum_{i=0}^{M-1} p_i = 1$; oppure senza vincolo in maniera tale che i valori dei pesi possano variare liberamente per tutte le variabili nell'intervallo *[0,1]*.

2.7. La definizione delle metriche per le distanze

Allo scopo di poter realizzare il clustering dei dati in esame, attraverso il ben noto algoritmo *k-means* è stato necessario individuare e definire opportune metriche per il calcolo delle distanze tra le variabili omologhe costituenti i pattern dei dataset, in modo tale da definire una funzione di dissimilarità ad hoc.

Avendo a disposizione variabili di due tipologie, quantitative e nominali, sono state definite due metriche diverse. In riferimento alla Tabella 12, per quanto riguarda le variabili quantitative, come le *età (x1 e x3)*, la *durata mansione (x2)*, ed il *genere (x4)*, è stato possibile adottare una misura per la distanza tra i dati di tipo euclideo; per le altre due variabili nominali di tipo categorico, la *mansione specifica (x5)* (campo *id_mansioneSpecifica* della tabella *lavoratori*) e il codice dell'attività economica aziendale *ateco (x6)* (campo *id_ateco* della tabella *mansioniAttivita*), è stato necessario definire e studiare ad hoc una metrica appropriata per il calcolo delle distanze tra i dati. Per tale metrica sono stati effettuati diversi cicli di test utilizzando un'apposita tabella all'interno del dataset definita per lo scopo, apportando alla metrica, di volta in volta, le opportune correzioni ed aggiustamenti.

Sia *S* il dataset considerato e sia $a \in S$ un suo generico elemento definito e strutturato nel seguente modo:

a = {

[0]	ETA' (*x1*)	quantitativa
[1]	DURATA MANSIONE (*x2*)	quantitativa
[2]	ETA' INIZIO (*x3*)	quantitativa
[3]	GENERE (*x4*)	quantitativa
[4]	MANSIONE SPECIFICA (*x5*)	nominale
[5]	ATECO (*x6*)	nominale

}

Il record *a* del dataset in questione è definito come un vettore di $n = 6$ elementi ciascuno dei quali corrispondente ad una variabile da esaminare.

Tutte le distanze sono state calcolate in modo normalizzato nell'intervallo unitario *[0,1]*.

La distanza euclidea tra due variabili omologhe di tipo quantitativo è stata calcolata come segue:

Sia *S* il dataset considerato;

siano *a* e $b \in S$ due elementi qualsiasi del dataset considerato;

le variabili *x1*, *x2* e $x3 \in \{x \in \aleph \mid x \in [n_1, n_2] \subset \aleph\}$ sono numeri naturali;

sia $range = (n_2 - n_1) \in \aleph$

$$d(xi_a, xi_b) = \frac{|xi_a - xi_b|}{range_{xi}} \qquad \text{dove } i \in \{1,2,3\}; \qquad 0 \leq d(xi_a, xi_b) \leq 1$$

La variabile *genere x4* è dicotomica e può assumere solo due valori corrispondenti a *maschio* o *femmina* i quali possono essere rappresentati o con due lettere dell'alfabeto rispettivamente *M* e *F* oppure con due numeri come *0* e *1*. La distanza viene calcolata come segue:

$d(x4_a, x4_b) = 0$ $\quad se\ x4_a = x4_b$ (entrambi maschi o entrambi femmine)
$d(x4_a, x4_b) = 1$ $\quad se\ x4_a \neq x4_b$ (uno maschio l'altro femmina)

Per le variabili *mansione specifica* e *ateco x5* e *x6* di tipo nominale, associate ai sistemi di classificazione, è stata definita una metrica ad hoc che rispetti il criterio che la loro distanza sia valutata attraverso valori appartenenti all'intervallo unitario *[0,1]*.

$0 \leq d(xi_a, xi_b) \leq 1$ $\quad$ dove $i \in \{5,6\}$; $\quad$ a, b$\in S$

La definizione finale è stata raggiunta mediante passi successivi testati opportunamente.

La variabile *mansione specifica x5,* è di tipo testo codificata da 4 byte, in un primo approccio la distanza tra le variabili di due record del dataset è stata definita come segue:

4 byte uguali	→ d = 0
primi 3 byte uguali	→ d = $\frac{1}{4}$
primi 2 byte uguali	→ d = $\frac{1}{2}$
primo byte uguale	→ d = $\frac{3}{4}$
primo byte differente	→ d = 1

Su tale definizione di distanza sono stati eseguiti i relativi test, raccogliendo i primi risultati di cluster analysis riguardo alla significatività delle aggregazioni tra gli elementi e dei tempi di elaborazione. In un successivo approccio, allo scopo di ottenere raggruppamenti più significativi, è stata considerata la distanza definita su un troncamento della codifica della *mansione specifica* considerando solo i primi 3 byte (livelli di grande gruppo, gruppo e classe) e successivamente solo i primi 2 byte (livelli di grande gruppo e gruppo).

Confrontando i risultati ottenuti dai test effettuati sui codici con codifica intera a 4 byte e sui codici a codifica parziale considerando solo i primi 3 byte e i primi 2 byte, si è potuto riscontrare che sia per l'aggregazione tra gli elementi, sia per il calcolo delle distanze sono state rilevate differenze importanti; nel caso della codifica a 2 byte, le aggregazioni tra gli elementi sono risultate essere più omogenee, dato che per la variabile *mansione specifica* si considera la codifica fino al livello di gruppo ed infine, sono stati ottenuti notevoli guadagni nei tempi di elaborazione.

La definizione di distanza per la variabile *mansione specifica (x5)* che è stata infine considerata è la seguente:

2 byte uguali	→ d = 0
primo byte uguale	→ d = $\frac{1}{2}$
primo byte diverso	→ d = 1

Di seguito un esempio di calcolo della distanza tra due codici di codifica della variabile *mansione specifica* mettendo a confronto i due metodi appena illustrati.

Considerate le seguenti 4 variabili di tipo *mansione specifica* dei quattro record *a, b, c* e *d* del dataset *S*

$x5_a$ = 2.1.1.4 – Informatici e Telematici

$x5_b$ = 2.1.1.3 – Matematici e Statistici

$x5_c$ = 6.1.2.2. – Muratori in cemento armato

$x5_d$ = 6.2.3.7. – Meccanici collaudatori

Il valore delle distanze in base alla definizione precedente è:

$d(x5_a, x5_b) = d(2.1, 2.1) = 0$

$d(x5_c, x5_d) = d(6.1, 6.2) = \frac{1}{2} = 0{,}5$

$d(x5_a, x5_c) = d(2.1, 6.1) = 1$

In questo modo "Informatici e Telematici" e "Matematici e Statistici" quasi sicuramente verranno aggregati nello stesso gruppo, i "Muratori in cemento armato" ed i "Meccanici collaudatori" verranno aggregati al 50%, infine, quasi sicuramente gli "Informatici e Telematici" non saranno aggregati con i "Muratori in cemento armato".

Con il primo approccio avremmo avuto:

$d(x5_a, x5_b) = d(2.1.1.4, 2.1.1.3) = \frac{1}{4} = 0{,}25$

con possibilità di aggregazione al 75% e con necessità di più cicli elaborativi.

La variabile *ateco x6,* è di tipo testo codificata da 5 byte rappresentati da 2 coppie ed 1 byte singolo. In un primo approccio la distanza tra le variabili di due record del dataset è stata definita come segue:

5 byte uguali	→ d = 0
primi 4 byte uguali	→ d = $\frac{1}{3}$
primi 2 byte uguali	→ d = $\frac{2}{3}$
primo 2 byte differenti	→ d = 1

Per la realizzazione di tale definizione, poiché nel DB MaProWeb sono presenti anche valori di lunghezza inferiore ai 5 byte corrispondenti ai livelli di tipo *divisione* (2 cifre), *gruppo* (3 cifre) o *classe* (4 cifre) e non di *categoria* (5 cifre), il calcolo della distanza tra due variabili di tali valori, è stato eseguito aggiungendo tanti *'0'* in coda al codice necessari al raggiungimento dei 5 byte.

Su tale definizione di distanza sono stati eseguiti i relativi test di cluster analysis, raccogliendo i primi risultati delle aggregazioni tra gli elementi e dei tempi di elaborazione. In un successivo approccio, allo scopo di ottenere raggruppamenti più significativi, è stata considerata la distanza definita su un troncamento della codifica della variabile *ateco* considerando solo i primi 2 byte (livello delle *divisioni*).

Confrontando i risultati ottenuti dai test effettuati sui codici con codifica intera a 5 byte e sui codici a codifica parziale considerando solo la prima coppia di byte, si è potuto riscontrare che sia per l'aggregazione tra gli elementi, sia per il calcolo delle distanze non sono state rilevate differenze importanti; anzi, nel caso della codifica a 2 byte, le aggregazioni tra gli elementi sono risultate essere più omogenee, dato che per la variabile *ateco* si considera la codifica al livello di divisione. Infine, sono stati ottenuti notevoli guadagni nei tempi di elaborazione.

La definizione di distanza per la variabile *ateco (x6)* sulla prima coppia di byte, è stata data in due versioni l'una successiva all'altra e testate in momenti successivi; vengono rappresentate di seguito:

prima versione:

primi 2 byte uguali	→ d = 0
primo byte uguale	→ d = $\frac{1}{2}$
primo byte differente	→ d = 1

seconda versione:

primi 2 byte uguali	→ d = 0
altrimenti	→ d = 1

Di seguito un esempio di calcolo della distanza tra due codici di codifica della variabile *ateco* mettendo a confronto le versioni di distanza appena illustrate.

Considerate le seguenti 5 variabili di tipo *ateco* dei cinque record *a, b, c, d* ed *e* del dataset *S*

$x6_a$ = 01.13.1 - Colture viticole e aziende vitivinicole

- la coltivazione di uva da vinificazione e di uva da tavola
- la produzione di vino di uva di produzione propria

$x6_b$ = 01.13.2 - Colture olivicole

- la coltivazione di olive per la produzione di olio e per il consumo diretto

$x6_c$ = 05.01.1 - Esercizio della pesca in acque marine e lagunari

- la pesca alturiera, costiera
- la raccolta di crostacei e molluschi marini
- la caccia ad animali acquatici: tartarughe, ascidie, tunicati
- la raccolta di prodotti marini: ostriche perlifere, spugne, ricci di mare, coralli e alghe
- la cattura di balene

$x6_d$ = 01.12.1 - Coltivazione di ortaggi

$x6_e$ = 14.30.2 - Estrazione di zolfo e pirite

Il valore delle distanze in base al primo approccio è:

$d(x6_a, x6_a) = d(01.13.1, 01.13.1) = 0$

$d(x6_a, x6_b) = d(01.13.1, 01.13.2) = \frac{1}{3} = 0{,}33$

$d(x6_a, x5_d) = d(01.13.1, 01.12.1) = \frac{2}{3} = 0{,}67$

$d(x6_a, x6_c) = d(01.13.1, 05.01.1) = 1$

Il valore delle distanze in base alla prima versione della definizione del secondo approccio è:

$d(x6_a, x6_b) = d(01, 01) = 0$

$d(x6_a, x6_c) = d(01, 05) = \frac{1}{2}$

$d(x6_a, x6_e) = d(01, 14) = 1$

Il valore delle distanze in base alla seconda versione della definizione del secondo approccio è:

$d(x6_a, x6_b) = d(01, 01) = 0$

$d(x6_a, x6_c) = d(01, 05) = 1$

Con l'ultimo approccio si avranno aggregazioni nell'ambito delle divisioni e quindi più significative, inoltre l'elaborazione risulta essere molto più rapida avendo introdotto criteri di ottimizzazione per il confronto dei byte.

Quanto si è appena detto riguarda la definizione delle metriche ed il calcolo delle distanze tra le singole variabili omologhe costituenti i record dei dataset in esame, ma gli elementi da considerare sono i record presi nella loro interezza comprendente tutte le variabili descritte. Quindi, per il calcolo della distanza tra due record è necessario definire una misura che consideri l'intero record. E' possibile applicare la teoria dei vettori sezionati ed ai fini della cluster analysis è possibile definire una misura di dissimilarità come la Linearly Weighted Distance (LWD).

Infatti, riprendendo le definizioni del precedente paragrafo sui dati strutturati e vettori sezionati, un record di un dataset può essere considerato come un dato strutturato di seconda specie.

E' pertanto possibile applicare le precedenti definizioni e teoria ai record dei nostri dataset; infatti:

a) *valore* di M = 6, (6 n-uple di dimensione 1 pari al numero delle variabili individuate);

b) il *nome* per ciascuna sezione, ossia la sua etichetta, è il nome di ciascuna delle variabili precedentemente elencate in Tabella 12;

c) il *tipo* di ciascuna sezione, è uguale a quello delle variabili;

d) la *dimensione* di ciascuna sezione sarà uguale *1*.

Pertanto la misura di dissimilarità è:

$$\mathrm{LWD}(u,v) = \delta(u,v) = \sum_{i=1}^{M=6} p_i \delta_i(u_i, v_i) \qquad (2.1)$$

dove *u* e *v*, sono due record dei dataset creati e in esame;

$p_i \in \Re;\ p_i \in [0,1]\ \ \forall i \in [1,6]$ è il peso della *i-esima variabile*

$\delta_i(u_i, v_i)$ è la distanza tra i pattern *u* e *v* relativa alla *i-esima* variabile.

Quindi, in riferimento alla Tabella 12, si riepilogano di seguito le relative distanze tra le variabili di due pattern introdotte precedentemente, e definite in modo diverso in base al tipo di caratteristica in esame, di tipo continuo o di tipo categorico (nominale / discreto).

- Per le *età (x1 e x3)* e la *durata mansione (x2)*, δ_i è la distanza Euclidea normalizzata all'intervallo [0,1];
- Per il *genere (x4)* e il codice dell'attività economica aziendale *ateco (x6)*, nel caso di concordanza $\delta_i = 0$, altrimenti $\delta_i = 1$ (confronto semplice);
- Per la *mansione specifica (x5)* (attività professionale del lavoratore), in caso di concordanza di entrambi i caratteri $\delta_i = 0$, in caso di concordanza solo del primo carattere $\delta_i = ½$, altrimenti $\delta_i = 1$.

3. Gli algoritmi

3.1 L'algoritmo di clustering

Il *clustering* è una tecnica nata in ambito statistico, e consente attraverso un'analisi dei dati di segmentarli e raggrupparli secondo dei pattern (elementi). Gli algoritmi di clustering sono da sempre di notevole interesse per innumerevoli settori: biologia molecolare, progettazione componenti elettronici, text mining, bioinformatica, biomedicina e medicina, marketing per la segmentazione della clientela, Social Network Analysis (SNA), Data Mining Analysis, Web, filtraggio immagini, etc.

Il clustering o analisi dei gruppi (dal termine inglese *cluster analysis* introdotto da Robert Tryon nel 1939) è un insieme di tecniche di analisi multivariata dei dati volte alla selezione e raggruppamento di elementi omogenei in un insieme di dati. Le tecniche di *clustering* si basano su misure relative alla somiglianza tra gli elementi. In molti approcci questa similarità, o meglio, dissimilarità, è concepita in termini di distanza in uno spazio multidimensionale. La bontà delle analisi ottenute dagli algoritmi di *clustering* dipende molto dalla scelta della *metrica*, e quindi da come è calcolata la distanza. Gli algoritmi di *clustering* raggruppano gli elementi sulla base della loro distanza reciproca, e quindi l'appartenenza o meno ad un *insieme* dipende da quanto l'elemento preso in esame è distante dall'insieme stesso.

Le tecniche di *clustering* si possono basare su due principali metodologie:

- Dal basso verso l'alto (*metodi aggregativi o bottom-up*):

In tal caso si prevede che inizialmente tutti gli elementi siano considerati *cluster* a sé, e poi l'algoritmo provvede ad unire i *cluster* più vicini. L'algoritmo continua ad unire elementi al *cluster* fino ad ottenere un numero prefissato di *cluster*, oppure fino a che la distanza minima tra i *cluster* non supera un certo valore, o ancora in relazione ad un determinato criterio statistico prefissato.

- Dall'alto verso il basso (*metodi divisivi o top-down*):

All'inizio tutti gli elementi sono un unico *cluster*, e poi l'algoritmo inizia a dividere il *cluster* in tanti *cluster* di dimensioni inferiori. Il criterio che guida la divisione è naturalmente quello di ottenere gruppi sempre più omogenei. L'algoritmo procede fino a che non viene soddisfatta una regola di arresto generalmente legata al raggiungimento di un numero prefissato di *cluster*.

Nel 1999 Jain classificò gli algoritmi di clustering (algoritmi non supervisionati) in due grandi famiglie:

• algoritmi gerarchici di clustering, che fanno uso di rappresentazioni grafiche (dendrogrammi)

• algoritmi partition clustering, che assegnano i dati o partizionano i dati, secondo criteri (come metriche, densità, similarità etc), inserendo essi in cluster differenti.

Sono state fatte diverse classificazioni degli algoritmi di clustering, oggi si dovrebbero classificare come:

Gerarchico, *Partizionale*, *Densità*, *Modello*, *Grid-Based*; la tendenza, comunque, è spesso di avere un misto di tecniche tra essi, per migliorare la scalabilità e l'efficienza, tale da ridurli a sole due o tre categorie principali.

Infatti tenendo conto del tipo di algoritmo utilizzato per dividere lo spazio, un'altra suddivisione delle tecniche di clustering è la seguente:

- *Clustering Partizionale (detto anche non gerarchico, o k-clustering)*, in cui per definire l'appartenenza ad un gruppo viene utilizzata una distanza da un punto rappresentativo del cluster (centroide, medioide ecc...), avendo prefissato il numero di gruppi della partizione risultato. Si tratta di derivazioni del più noto algoritmo di clustering, quello detto delle *k-means,* introdotto da Mac Queen nel 1967.
- *Clustering Gerarchico*, in cui viene costruita una gerarchia di partizioni caratterizzate da un numero (de)crescente di gruppi, visualizzabile mediante una rappresentazione ad albero (dendrogramma), in cui sono rappresentati i passi di accorpamento/divisione dei gruppi.

Ognuno di questi algoritmi presenta una particolarità, tra questi algoritmi il *k-means* risulta essere molto efficiente con elementi di tipo numerico, meno efficiente con dati non numerici, inoltre l'algoritmo, in generale, converge molto velocemente e pertanto può essere applicato più volte e fra le soluzioni prodotte scegliere quella più soddisfacente. Tale algoritmo non determina automaticamente il numero di cluster ma questo deve essere fissato a priori (esistono metodi per determinare in modo automatico il numero di cluster), infine questo tipo di algoritmi, mira ad identificare i gruppi naturali presenti nel dataset, consentendo di suddividere gruppi di oggetti sulla base dei loro attributi e sono ottimi per grandi dataset. Gli algoritmi di *clustering gerarchico* danno come risultato una serie di partizioni innestate e mirano ad evidenziare le relazioni tra i vari pattern del dataset. Inoltre, determinano automaticamente il numero di cluster ma sono improponibili per dataset di grandi dimensioni.

Poiché nel nostro caso, l'analisi dei dati viene eseguita su un dataset su base nazionale in continua crescita e quindi di dimensioni non fisse e note a priori, ed inoltre, uno degli scopi della ricerca è quello di identificare gli elementi del dataset che possano costituire gruppi di elementi simili tra loro, si è scelto di utilizzare gli algoritmi di *clustering partizionale* ed in particolare il *k-means*.

Utilizzando l'algoritmo di clustering del *K-means*, sono state eseguite varie elaborazioni allo scopo di ottenere uno strumento ottimizzato ed utilizzabile per essere impiegato in combinazione con l'algoritmo genetico.

3.2 Gli indici di validazione

Gli indici di validazione danno una misura della compattezza e dispersione dei clusters di una partizione.

Sono stati utilizzati in una versione del software per la determinazione della partizione ottima e successivamente nella valutazione della fitness dell'algoritmo genetico.

Sono stati considerati i seguenti indici di validazione:

- Indice di *Davies-Bouldin*
- Indice di *Kalinski-Harabasz*
- Indice di *Maulik-Bandyopadhyay* detto indice *I*

Per questi tre indici la partizione ottima è ottenuta, per *Davies-Bouldin* in corrispondenza del minimo valore dell'indice; per *Kalinski-Harabasz* e per *Maulik-Bandyopadhyay* in corrispondenza del massino valore dell'indice.

3.2.1 Indice di Davies Bouldin

$$DB(K) = \frac{1}{K}\sum_{k=1}^{K} R_k \tag{3.1}$$

con $K \in N$ numero di cluster

dove

$R_k = \max_{j=1,...,K,\, j\neq k}\{R_{jk}\}$ per $k \in [1,...,K]$ indice per il cluster k

separazione dei clusters j e k $R_{jk} = \frac{e_j + e_k}{d(c_j, c_k)}$ dove d è la distanza tra il centroide del cluster j e del cluster k ed e è la compattezza del *j-esimo* e *k-esimo* cluster rispettivamente.

$e_k = \sqrt{\sum_{j=1}^{n_k}(d(x_j, c_k))^2}$ compattezza del cluster k. Dove n_k è la cardinalità del cluster k.

La partizione ottima è quella in corrispondenza del valore minimo di R_k.

3.2.2 Indice di Kalinski-Harabasz

$$CH(K) = \frac{[traceB/(K-1)]}{[traceW/(N-K)]} \tag{3.2}$$

con K numero di clusters e N cardinalità del dataset.

la trace(B) (ossia la traccia sulla matrice B di dispersione inter-cluster).

$traceB = \sum_{k=1}^{K} n_k (d(c_k, c))^2$ dove n_k è la cardinalità del cluster *k-esimo* e d calcola la distanza tra il centroide del cluster *k-esimo* ed il centroide del dataset (c)

la trace(W) (ossia la traccia sulla matrice W di dispersione intra-cluster).

$traceW = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(d(x_i, c_k))^2$ dove d è la distanza intra-cluster di ciascun elemento ed il centroide del *k-esimo* cluster.

La partizione ottima è quella in corrispondenza del valore massimo di *CH(K).*

3.2.3 Indice di Maulik-Bandyopadhyay (indice I)

$$I(K) = \left(\frac{1}{K} \frac{E(1)}{E(K)} D(k) \right)^2 \tag{3.3}$$

con *K* numero di clusters

dove gli *E(K)* sono le distanze intra-cluster e *D(K)* è la distanza inter-cluster.

$E(K) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d(x_i, c_k)$ dove n_k è la cardinalità del cluster *k-esimo* e *d* calcola la distanza tra il centroide del cluster *k-esimo* e ciascun suo elemento.

$D(K) = \max_{i,j=1,\ldots,K, j \neq i} d(c_i, c_j)$

La partizione ottima è quella in corrispondenza del valore massimo di *I(K)*.

3.3 Algoritmo Genetico

Un algoritmo genetico è una particolare meta-euristica di ottimizzazione, ispirata dal principio della selezione naturale ed evoluzione biologica teorizzato nel 1859 da Charles Darwin.

Gli algoritmi genetici valutano le soluzioni di partenza e, ricombinandole tra loro ed introducendo un certo grado di randomicità, ne creano di nuove al fine di far evolvere la popolazione verso soluzioni accettabili, seppur sub-ottime.

Queste tecniche vengono di norma utilizzate per tentare di risolvere problemi di ottimizzazione per i quali non si conoscono altri algoritmi efficienti di complessità lineare o polinomiale. Date le tecniche evolutive alla base degli AG, questi ultimi, pur non potendo garantire il raggiungimento della soluzione ottima, comunque assicurano, generazione dopo generazione, il miglioramento delle soluzioni.

Gli algoritmi genetici rientrano nello studio dell'intelligenza artificiale e più in particolare nella branca della *computazione evolutiva*, vengono studiati e sviluppati all'interno del campo dell'intelligenza artificiale e delle tecniche di soft computing ma trovano applicazione in un'ampia varietà di problemi afferenti a diversi contesti quali l'elettronica, la biologia e l'economia.

La vera prima creazione di un algoritmo genetico è storicamente attribuita a John Henry Holland che, nel 1975, nel libro *Adaptation in Natural and Artificial Systems* pubblicò una serie di teorie e di tecniche tuttora di fondamentale importanza per lo studio e lo sviluppo della materia. Agli studi di Holland si deve infatti sia il teorema che assicura la convergenza degli algoritmi genetici verso soluzioni ottimali sia il cosiddetto teorema degli schemi, conosciuto anche come "Teorema fondamentale degli algoritmi genetici".

Enormi contributi si devono anche a John Koza che nel 1992 inventò la programmazione genetica ossia l'applicazione degli algoritmi genetici alla produzione di software in grado di evolvere e diventando capace di compiti che in origine non era in grado di svolgere.

Nel 1995 Stewart Wilson re-inventò i sistemi a classificatori dell'intelligenza artificiale ri-denominandoli come XCS e rendendoli capaci di apprendere attraverso le tecniche degli algoritmi genetici mentre nel 1998 Herrera e Lozano presentarono un'ampia rassegna di operatori genetici. Gli operatori di Herrera e Lozano sono applicabili a soluzioni codificate mediante numeri reali ed

hanno reso il campo dei numeri reali un'appropriata e consolidata forma di rappresentazione per gli algoritmi genetici in domini continui.

3.3.1 Il Funzionamento

E' necessario premettere che gli AG ereditano e riadattano dalla biologia alcune terminologie che vengono qui preventivamente presentate:

- *Cromosoma*: una delle soluzioni ad un problema considerato. Generalmente è codificata con un vettore di bit, di caratteri o numeri reali.
- *Popolazione*: insieme di soluzioni relative al problema considerato.
- *Gene*: parte di un cromosoma. Generalmente consiste in una o più parti del vettore di bit, caratteri o numeri reali che codificano il cromosoma.
- *Fitness*: grado di valutazione associato ad una soluzione. La valutazione avviene in base ad una funzione appositamente progettata detta *funzione di fitness*.
- *Crossover*: generazione di una nuova soluzione mescolando delle soluzioni esistenti.
- *Mutazione*: alterazione casuale di una soluzione.

Un tipico algoritmo genetico, nel corso della sua esecuzione, provvede a fare evolvere delle soluzioni secondo il seguente schema di base:

1. Generazione casuale della prima popolazione di soluzioni (cromosomi).
2. Applicazione della funzione di *fitness* alle soluzioni (cromosomi) appartenenti all'attuale popolazione.
3. Selezione delle soluzioni considerate migliori in base al risultato della funzione di fitness e della logica di selezione scelta.
4. Procedimento di crossover per generare delle soluzioni ibride a partire dalle soluzioni scelte al punto *3*.
5. Creazione di una nuova popolazione a partire dalle soluzioni identificate al punto *4*.
6. Riesecuzione della procedura a partire dal punto *2* ed utilizzando la nuova popolazione creata al punto *5*.

La *condizione di terminazione* dell'algoritmo può essere data da:

- Un'alta percentuale degli individui di una generazione ha la stessa fitness dell'individuo migliore.
- Sono state generate *n* generazioni, con *n* fissato.

L'iterazione dei passi presentati permette l'evoluzione verso una soluzione ottimizzata del problema considerato.

Poiché questo algoritmo di base soffre del fatto che alcune soluzioni ottime potrebbero essere perse durante il corso dell'evoluzione e del fatto che l'evoluzione potrebbe ricadere e stagnare in "ottimi locali" spesso viene integrato con la tecnica dell' *"elitismo"* e con quella delle *"mutazioni casuali"*. La prima consiste in un ulteriore passo precedente al punto *3* che copia nelle nuove popolazioni anche gli individui migliori della popolazione precedente in base ad una percentuale

prefissata; la seconda invece, successiva al punto *4*, introduce (in una percentuale prefissata) nelle soluzioni individuate delle occasionali mutazioni casuali in modo da permettere l'uscita da eventuali ricadute in ottimi locali.

Le soluzioni (cromosomi) al problema considerato, sia quelle casuali di partenza sia quelle derivate da evoluzione, devono essere codificate con qualche tecnica. Le codifiche più diffuse sono:

- *Codifica vettoriale binaria*: è la più diffusa, consiste in un vettore di *n* campi binari dove i valori 1 o 0 identificano delle caratteristiche elementari della soluzione. I vantaggi di questa tecnica risiedono nel fatto di essere semplice da implementare e da gestire durante l'intera evoluzione. Gli svantaggi consistono nelle difficoltà intrinseche della conversione delle soluzioni in questa codifica e dalle scarse possibilità rappresentative.
- *Codifica vettoriale reale*: come la codifica vettoriale binaria ma vengono utilizzati dei numeri reali. Il vantaggio è quello di introdurre una maggiore espressività e versatilità nella codifica, a scapito di un'aumentata complessità.
- *Codifica vettoriale diretta*: consiste in una codifica vettoriale dove ogni campo contiene direttamente i valori relativi al problema. Il vantaggio è quello di una facile codifica, lo svantaggio risiede nella difficile gestione dell'algoritmo e nella difficile progettazione della funzione di fitness e dei processi di crossover e mutazione.
- *Codifica ad albero*: Ogni cromosoma è un albero di alcuni oggetti come ad esempio funzioni e comandi di un linguaggio di programmazione. In questo caso, per la sua particolare semantica e sintassi, viene spesso utilizzato il linguaggio di programmazione Lisp che semplifica notevolmente le operazioni di codifica.

3.3.2 La funzione di fitness

La *funzione di fitness* è quella che permette di associare ad ogni soluzione uno o più parametri legati al modo in cui quest'ultima risolve il problema considerato. Generalmente è associata alle prestazioni computazionali e quindi alle prestazioni temporali della soluzione.

La *qualità* di un individuo (cioè quanto è buona la soluzione per il problema) è misurata mediante una *funzione di fitness*. In un certo senso, la funzione di fitness indica l'adattabilità all'ambiente: gli individui che meglio si adattano ("fit") hanno piú probabilità di riprodursi e di trasmettere i propri geni alle generazioni future.

Un AG è una procedura di ricerca iterativa il cui scopo è l'ottimizzazione della funzione di fitness. Gli AG sono procedure di massimizzazione, quindi valori di fitness più alti sono associati ad individui migliori (problemi di minimizzazione vengono di solito riformulati). A volte la fitness di un cromosoma è misurata in maniera implicita, valutando la qualità della corrispondente soluzione rispetto al problema. Ad esempio, se abbiamo a disposizione un insieme di esempi, la fitness può essere calcolata in funzione dell'errore della soluzione (differenza tra la soluzione attuale e l'uscita desiderata).

3.3.3 La logica di selezione

A causa di complessi fenomeni di interazione non lineare (epistaticità), non è dato per scontato né che da due soluzioni promettenti ne nasca una terza più promettente né che da due soluzioni con valori di fitness basso ne venga generata una terza con valore di fitness più basso. Per ovviare a questi problemi, durante la scelta delle soluzioni candidate all'evoluzione, oltre che sul parametro ottenuto dalla funzione di fitness ci si basa anche su particolari tecniche di "selezione". Le più comuni sono:

- *Selezione a roulette*: la probabilità che una soluzione venga scelta per farla evolvere è direttamente proporzionale al valore restituito dalla funzione di fitness. Questa tecnica presenta dei problemi nel caso in cui ci siano delle grosse differenze di valori perché le soluzioni peggiori verrebbero selezionate troppo raramente.
- *Selezione per categoria*: simile alla selezione per roulette ma la valutazione è effettuata in maniera proporzionale alla somma del valore della funzione di fitness per ogni coppia possibile di soluzioni. Il problema presentato da questa tecnica di scelta è rappresentato dalla lentezza di convergenza nel caso in cui ci siano delle differenze troppo piccole tra coppie di soluzioni candidate.
- *Selezione a torneo*: le soluzioni vengono raggruppate e si procede a valutarle con un algoritmo come quello presentato di seguito:
    a. Scegliere in maniera casuale *N* individui appartenenti alla popolazione.
    b. Scegliere l'individuo migliore attraverso il confronto delle loro fitness e si imposta la sua probabilità di scelta a *p*.
    c. Scegliere il secondo individuo migliore e impostare la probabilità di scelta a *p*(1-p)*
    d. Scegliere il terzo individuo migliore e impostare la sua probabilità di scelta a $p*(1-p)$.
    e. ...proseguire fino ad esaurire le soluzioni scelte.
- *Selezione di Boltzmann*: le soluzioni vengono scelte con un grado di probabilità che, agli inizi dell'algoritmo, favorisce l'esplorazione e che poi tende a stabilizzarsi.

3.3.4 Il Crossover

Per semplicità, si farà riferimento alle codifiche vettoriali binarie ma il procedimento per le codifiche con elementi reali o ad albero è simile, per quest'ultimo caso, invece che essere applicato ai campi dei vettori viene applicato ai nodi dell'albero. In base ad un operatore stabilito inizialmente, alcune parti dei geni delle soluzioni candidate all'evoluzione vengono mescolate per ricavare nuove soluzioni.

Gli operatori più comunemente utilizzati sono il crossover ad un punto, il crossover a due punti ed il crossover uniforme di seguito esposti.

- *Crossover ad un punto*: consiste nel considerare due soluzioni adatte all'evoluzione e nel tagliare i loro vettori di codifica in un punto casuale o predefinito per ottenere due teste e due code. La prima nuova soluzione ottenuta sarà data dalla combinazione della testa della prima soluzione con la coda della seconda, mentre la seconda nuova soluzione sarà data dalla coda della prima soluzione con la testa della seconda.

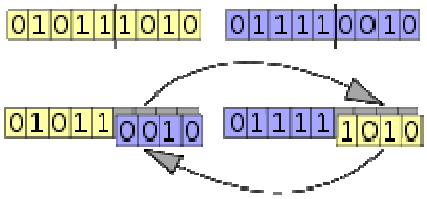

Figura 3: Crossover ad un punto

- *Crossover a due punti:* consiste nel considerare due soluzioni adatte all'evoluzione e nel tagliare i loro vettori di codifica in due punti predefiniti o casuali al fine di ottenere una testa, una parte centrale ed una coda dalla prima e dalla seconda soluzione. La prima nuova soluzione sarà data dalla testa e della coda della prima soluzione e dalla parte centrale della seconda soluzione. La seconda nuova soluzione sarà data dalla parte centrale della prima soluzione e dalla testa e dalla coda della seconda soluzione.

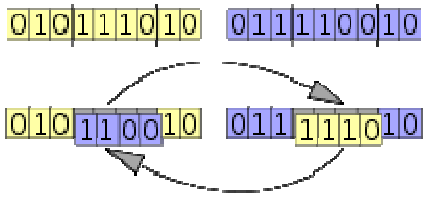

Figura 4: Crossover a due punti

- *Crossover uniforme:* consiste nello scambiare casualmente dei bit tra le soluzioni candidate all'evoluzione. Si segnala l'esistenza anche di crossover uniformi parziali ossia dei crossover uniformi dove lo scambio di bit è limitato ad una percentuale fissa o dinamica dei cromosomi candidati all'evoluzione.

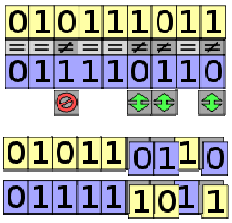

Figura 5: Crossover uniforme

- *Crossover aritmetico:* consiste nell'utilizzare un'operazione aritmetica per creare la nuova soluzione. (es. XOR o un AND).

Non è detto che il crossover debba avvenire ad ogni iterazione dell'algoritmo genetico. Generalmente la frequenza di crossover è regolata da un apposito parametro p_c.

3.3.5 La mutazione

In campo biologico, per *mutazione genetica* si intende ogni modifica stabile ed ereditabile nella sequenza nucleotidica di un genoma o più generalmente di materiale genetico (sia DNA che RNA) dovuta ad agenti esterni o al caso, ma non alla ricombinazione genetica. Una mutazione modifica quindi il genotipo di un individuo e può eventualmente modificarne il fenotipo a seconda delle sue caratteristiche e delle interazioni con l'ambiente.

Le mutazioni sono gli elementi di base grazie ai quali possono svolgersi i processi evolutivi. Le mutazioni determinano infatti la cosiddetta variabilità genetica, ovvero la condizione per cui gli organismi differiscono tra loro per uno o più caratteri. Su questa variabilità, tramite la ricombinazione genetica, opera la selezione naturale, la quale promuove le mutazioni favorevoli a scapito di quelle sfavorevoli o addirittura letali.

Nel caso degli algoritmi genetici, la mutazione consiste nella modifica pseudocasuale di alcune parti dei geni in base a coefficienti definiti inizialmente. Queste modifiche alle volte sono utilizzate per migliorare il valore della funzione di fitness per la soluzione in questione e altre volte sono utilizzate per ampliare lo spazio di ricerca ed attuare la tecnica dell'elitismo per non far ricadere l'evoluzione in ottimi locali. La frequenza con cui deve avvenire una mutazione è generalmente regolata da un apposito parametro p_m.

Nel caso di codifica vettoriale binaria la mutazione riguarda un singolo bit che viene cambiato nel valore opposto, da 0 in 1 o viceversa, con una probabilità prefissata p_m di solito molto piccola. La mutazione è applicata a due elementi discendenti generati per mezzo del crossover.

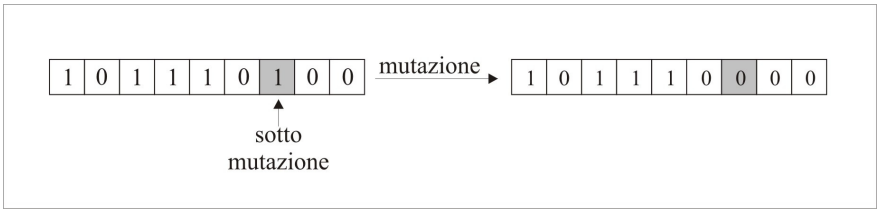

Figura 6: Esempio di mutazione di un gene in un codice binario

Di seguito il processo per l'applicazione dell'operatore di mutazione.

E' possibile agire su un unico gene oppure su tutti i geni all'interno di un cromosoma, nel primo caso, il gene su cui agire è scelto in modo casuale.

1. Azione su tutti i geni del cromosoma.

Per ogni gene di ogni cromosoma, si genera un numero casuale q con distribuzione uniforme in [0,1]:

a. se $q<p_m$ il gene è selezionato per la mutazione;

b. se si usa un operatore di *mutazione forte*, il gene viene mutato da *0* in *1* oppure da *1* in *0*; se si usa un operatore di *mutazione debole*, il nuovo valore è scelto casualmente come uno dei due valori *0* o *1*.

Nel caso della mutazione debole, di fatto, il gene potrebbe alla fine risultare immutato.

2. Azione su un unico gene del cromosoma.

Si genera un numero casuale q con distribuzione uniforme in [0,1]:

a. se $q<p_m$ per la mutazione si seleziona un gene in modo casuale all'interno del cromosoma;

b. si applica il precedente punto "1.b".

Nel caso di codifica vettoriale con numeri reali, l'operatore di mutazione funziona in modo analogo al caso della codifica binaria, infatti, i precedenti punti "1.a" e "2.a" restano invariati, il punto "b" si realizza in uno dei seguente modi:

- Ogni singolo gene selezionato all'interno di un cromosoma può essere rimpiazzato da un altro numero reale scelto casualmente all'interno dell'insieme dei valori di quel gene.
- Ogni singolo gene selezionato all'interno di un cromosoma viene rimpiazzato con un nuovo numero reale, all'interno dell'insieme dei valori di quel gene, ottenuto variando il valore dello stesso gene aggiungendo o sottraendo un valore costante prefissato δ. L'aggiunta o la sottrazione della costante viene decisa in modalità casuale equiprobabile. Se la modifica comporta l'ottenimento di un valore che supera il *max* o il *min* dell'insieme dei valori del gene, il nuovo valore considerato sarà pari al *max* o al *min*.

Per codifiche vettoriali a caratteri o interi, la mutazione funziona in modo analogo.

3.3.6 Definizione del codice genetico per l'AG

Per l'implementazione dell'AG per il presente studio è stato definito un opportuno *codice genetico* di seguito descritto.

In considerazione dell'equazione 2.1 della misura di dissimilarità tra due pattern, attraverso l'utilizzo dell'AG si vogliono ottimizzare i pesi p_i delle variabili considerate ed il valore di K del numero di cluster per ottenere la partizione ottima.

A tale scopo è stata utilizzata la codifica vettoriale reale in cui ogni cromosoma è un individuo della popolazione ed è rappresentato da un vettore di 7 campi, dove i primi 6 contengono numeri reali ed il settimo un numero intero, in particolare:

- i primi 6 corrispondono ai pesi delle variabili considerate e sono compresi nell'intervallo di valori reali [0, 1];
- il settimo corrisponde al valore di K utilizzato per il clustering numero intero compreso tra 2 ed un prefissato valore *K-max*.

Ogni posizione del vettore corrisponde ad un peso di una variabile della Tabella 12 come mostrato nella figura seguente:

prima sezione
[0] peso p_1 di *x1 Età* (Età del lavoratore all'anno del certificato)
[1] peso p_2 di *x2 Durata mansione*
[2] peso p_3 di *x3 Età inizio*
[3] peso p_4 di *x4 Genere*
[4] peso p_5 di *x5 Mansione specifica*
[5] peso p_6 di *x6 Ateco*
---
seconda sezione
[6] *K*

Figura 7: Vettore codificante un individuo della popolazione

Per l'implementazione dell'AG è stato necessario individuare un'apposita funzione di fitness affinché si potesse realizzare l'evoluzione della popolazione in esame, in particolare, sono state individuate più funzioni di fitness.

Sono state utilizzate meta euristiche di ottimizzazione nell'analisi di clustering, infatti, per la realizzazione del classificatore di malattie professionali, è stato necessario ricercare la migliore combinazione dei pesi delle variabili e del *K* ottimo utilizzando:

1 – Un modellamento *non supervisionato* realizzato attraverso un algoritmo genetico con funzione di fitness implementata per mezzo degli indici di validazione, ossia la funzione di fitness è il valore dell'indice di validazione calcolato in corrispondenza della clusterizzazione ottenuta con un dato individuo della popolazione.

2 - Un modellamento *supervisionato* realizzato attraverso un algoritmo genetico con funzione di fitness definita dalla prestazione su un insieme di validazione; il dataset iniziale viene suddiviso in tre parti: training, validation e test set.

Nel caso del modellamento non supervisionato, le funzioni di fitness sono state individuate in corrispondenza di ciascuno dei tre indici di validazione precedentemente introdotti (Indice di Maulik-Bandyopadhyay (indice I), di Davies-Bouldin, di Kalinski-Harabasz). Come già detto nel paragrafo 3.2, gli indici di validazione danno una misura della compattezza e dispersione dei cluster di una partizione e quindi danno un'indicazione per l'individuazione della partizione ottima in prossimità di un valore minimo o massimo; in particolare del valore minimo per l'indice di Davies-Bouldin e del valore massimo per l'indice I e Kalinski-Harabasz. La partizione individuata come ottima darà l'indicazione del miglior *K* che sarà quello utilizzato per ottenerla, inoltre, i pesi delle variabili utilizzati per il calcolo della dissimilarità, saranno considerati ottimi in quanto costituenti un individuo della popolazione sottoposta all'AG, che sarà assunto come il migliore.

Gli indici di validazione non hanno un limite superiore e/o inferiore, pertanto non è possibile stabilire un criterio in base al quale è affermabile di aver ottenuto il valore del *K* ottimo e dei pesi ottimi in quanto il valore della fitness (valore dell'indice di validazione) è uguale o molto prossimo al loro limite superiore e/o inferiore.

Allo scopo di migliorare i risultati ottenuti attraverso l'utilizzo dell'AG per la determinazione del *K* ottimo e dei pesi ottimi, si è definita un'apposita funzione di fitness che fosse indipendente da qualunque limitazione non propria dei pattern del dataset.

Una funzione di fitness basata sugli indici di validazione non consente di sfruttare l'informazione di classe, pertanto si è adottato un metodo di valutazione della fitness che fosse indipendente dagli indici.

Si è passati quindi al modellamento *supervisionato* del secondo caso. Ai fini della definizione delle nuove funzioni di fitness, il dataset di analisi considerato, è stato suddiviso in tre

sottoinsiemi: un insieme costituente il dataset di training, un insieme costituente il dataset di validazione ed infine il dataset di test. La nuova funzione di fitness degli individui della popolazione è stata definita come l'accuratezza di classificazione sul dataset di validazione.

I tre distinti dataset sono stati generati mediante un metodo di stratificazione random, ottenendo il seguente risultato:

- *Training Set*: il 50% del numero totale dei pattern disponibili, denotato con S_{TR};
- *Validation Set*: il 25% del numero totale dei pattern disponibili, denotato con S_{VAL};
- *Test Set*: il restante 25% del numero totale dei pattern disponibili, denotato con S_{TEST}.

Per i due tipi di modellamento sono stati utilizzati due dataset diversi.

Per il modellamento non supervisionato con fitness pari agli indici di validazione, il dataset utilizzato è stato quello non filtrato costituito dai 3427 record contenente tutte le patologie rilevate.

Per quanto riguarda il modellamento supervisionato il dataset utilizzato è stato quello filtrato costituito dall'80% dei 3427 record iniziali, di cui sono stati eliminati i record delle malattie professionali con una frequenza inferiore al 5%, ottenendo così un dataset composto di 2722 record rappresentanti le 6 malattie professionali più frequenti. La stratificazione random è stata eseguita su quest'ultimo dataset.

La seguente tabella mostra la distribuzione delle 6 malattie più frequenti, in ordine decrescente, nei tre sottoinsiemi S_{TR}, S_{VAL} ed S_{TEST} per il DB regionale; inoltre, ogni patologia è indicata con la propria etichetta codificata con un numero intero progressivo che sarà utilizzata per l'indicazione delle patologie nelle successive rappresentazioni tabellari:

Tabella 17: Distribuzione delle malattie nei tre sottoinsiemi su DB regionale

Patologia	**Training set (S_{TR})**	**Validation set (S_{VAL})**	**Test set (S_{TEST})**	**Totale**
1 - Sordità	747 54,89%	373 54,77%	373 54,85%	1493
2 - Malattie del rachide	167 12,27%	84 12,33%	83 12,21%	334
3 - Malattie muscolo scheletriche (escluso rachide)	144 10,58%	72 10,57%	72 10,59%	288
4 - Tumori pleura e peritoneo	116 8,52%	58 8,52%	58 8,53%	232
5 – Sindrome tunnel carpale	99 7,27%	50 7,34%	50 7,35%	199
6 – Malattie della pelle	88 6,47%	44 6,47%	44 6,47%	176
Totali	**1361** **100%**	**681** **100%**	**680** **100%**	**2722**

La seguente tabella mostra la distribuzione delle 6 malattie considerate nei tre sottoinsiemi ottenuti tramite la stratificazione random per il DB nazionale; vengono utilizzate le stesse etichette del DB regionale:

Tabella 18: Distribuzione delle malattie nei tre sottoinsiemi su DB nazionale

Patologia	Training set (S_{TR})	Validation set (S_{VAL})	Test set (S_{TEST})	Totale
1 - Sordità	1487	743	743	2973
2 - Malattie del rachide	908	454	454	1816
3 - Malattie muscolo scheletriche (escluso rachide)	1331	665	665	2661
4 - Tumori pleura e peritoneo	301	150	150	601
5 – Sindrome tunnel carpale	657	328	328	1313
6 – Malattie della pelle	156	78	78	312
Totali	**4840 100%**	**2418 100%**	**2418 100%**	**9676**

La nuova funzione di fitness, denominata f_1 degli individui della popolazione è stata definita come l'accuratezza di classificazione sul dataset di validazione S_{VAL}, secondo la funzione:

$$f_1 = accuracy = \frac{1}{|S|}\sum_{x \in S} h(\omega_x, \omega_{Kx}) \tag{3.4}$$

Dove:

S è il pattern set etichettato (in tal caso S_{VAL}) sul quale è calcolata l'accuratezza;

Ω = *{sordità, malattie del rachide, malattie muscolo scheletriche, tumori pleura e peritoneo, sindrome del tunnel carpale, malattie della pelle}* è il label set considerato;

$\omega_x \in \Omega$ è la patologia reale del lavoratore $x \in S$

$\omega_{Kx} \in \Omega$ è l'etichetta assegnata dal modello di classificazione al pattern (corrispondente all'etichetta del cluster *k-esimo* avente il minimo valore di dissimilarità con il pattern in esame)*;*

$h(\omega_x, \omega_{Kx}) = 1$ *se* $\omega_x = \omega_{Kx}$;

$h(\omega_x, \omega_{Kx}) = 0$ *se* $\omega_x \neq \omega_{Kx}$;

Osservando la distribuzione delle patologie all'interno del dataset si può evidenziare che queste non sono ben bilanciate; di conseguenza i valori della fitness possono essere distorti, dando

eccessiva importanza alla patologia più frequente. Pertanto, è stata introdotta una seconda funzione di fitness, variante della prima, con lo scopo di bilanciare con un peso ciascun eventuale errore di classificazione a prescindere dal loro numero e dalla frequenza della patologia. La nuova fitness è data dalla accuratezza pesata, ossia il valore medio delle percentuali di risposte corrette per ciascuna patologia.

La nuova funzione di fitness, denominata f_2, è definita nel modo seguente:

$$f_{2(weighted)} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \frac{1}{|S_\omega|} \sum_{x \in S} h(\omega_x, \omega_{Kx}) \tag{3.5}$$

Dove:

S_ω è il sottoinsieme di S di tutti gli elementi associati con la patologia $\omega \in \Omega$ $(S_\omega \subset S)$

Gli altri operatori hanno lo stesso significato della precedente espressione.

Per il modellamento supervisionato, sono state implementate due versioni di applicazione dell'algoritmo genetico: una detta di base denominata *Basic Algorithm (BA)* ed una variante denominata *Class-aware Basic Algorithm (CBA)*, per ciascuna di esse sono state utilizzate le due funzioni di fitness introdotte f_1 ed f_2.

Per il calcolo di ciascuna delle due fitness è necessario stabilire un criterio di etichettatura dei cluster. Per ciascuno dei due algoritmi, *BA* e *CBA*, è stato adottato un diverso criterio di etichettatura dei cluster. Le etichette, per entrambi gli algoritmi, sono state individuate nelle *n=6* malattie più frequenti contenute in *Ω*. In generale, si può pensare di estendere tale insieme, includendo due ulteriori elementi (anche se in tale fase questa condizione non è stata considerata), un elemento *B* rappresentante tutte le altre malattie, ed un elemento *C* rappresentante la situazione di *"non malattia"*.

3.4 BA (Basic Algorithm): algoritmo base

La procedura di base per la sintesi di un modello di classificazione consiste nel clustering del dataset di training S_{TR} attraverso il noto algoritmo *k-means*. A questo scopo, è stata definita una misura di dissimilarità δ ad hoc tra i *pattern*, come una combinazione lineare convessa di misure interne δ_i tra caratteristiche omologhe utilizzando la (2.1).

Il sistema globale di classificazione è stato progettato per determinare automaticamente i pesi p_i della misura di dissimilarità ed il numero ottimale di cluster K, al fine di massimizzare l'accuratezza della classificazione attraverso la funzione f_1 (3.4) ed f_2 (3.5)

Allo scopo di realizzare tale ottimizzazione, è stata realizzata un'adeguata implementazione di un algoritmo genetico. Il generico individuo della popolazione sottoposta all'evoluzione per mezzo degli operatori genetici, è formato da due strutture dati (sezioni) per un totale di 7 parametri da ottimizzare, come descritto all'inizio del paragrafo 3.3.6 e che qui ricordiamo:

1. Un vettore di 6 numeri reali con valori assunti nell'intervallo [0, 1], corrispondenti ai pesi associati con le variabili nella funzione di distanza δ;
2. Un numero intero con valore assunto tra 2 ed il massimo valore fissato tra i parametri di sistema, rappresentante il numero di cluster da utilizzare per il clustering del training set;

Da una generazione alla successiva, ogni individuo in AG viene valutato per mezzo di una funzione di fitness definita come accuratezza dalla (3.4) e dalla (3.5) calcolate su $S_{VAL.}$ La *selezione*

degli individui è simulata utilizzando l'operatore di tipo *selezione a roulette*; il *crossover* e la *mutazione* incidono sull'intero individuo formato dai sei pesi e dal numero di cluster. Gli individui della popolazione iniziale dell'AG sono generati utilizzando la modalità di campionamento casuale. Per ogni individuo, viene eseguita la procedura di clustering utilizzando l'algoritmo *k-means* sull'insieme di training S_{TR} calcolando la distanza tra i campioni di S_{TR} utilizzando i pesi contenuti nella prima sezione del codice genetico degli individui, ed utilizzando il numero intero memorizzato nella seconda sezione per realizzare il clustering in un numero di cluster pari a tale numero intero (vedi Figura 7).

Una volta ottenuta una partizione di S_{TR}, per poter procedere al calcolo di ciascuna delle due funzioni di fitness è stato necessario definire un criterio di etichettatura dei cluster e per l'algoritmo *BA*, si è deciso di etichettare ogni cluster unicamente con una delle malattie considerate, e precisamente con la patologia che risulta essere la più frequente tra i campioni contenuti all'interno del cluster considerato. In tal caso potremo avere che una stessa malattia, altamente diffusa nella popolazione iniziale, potrà essere la più frequente in più di un cluster, e quindi si potranno avere più cluster con la stessa etichetta. Tale situazione può essere vista come il caso che per diverse caratteristiche dei lavoratori è possibile contrarre la stessa malattia.

Il criterio di etichettatura per l'algoritmo BA può essere rappresentato con la seguente figura, in cui l'etichetta di ciascun cluster è rappresentata da una bandierina del colore (patologia) con frequenza più alta in quel cluster.

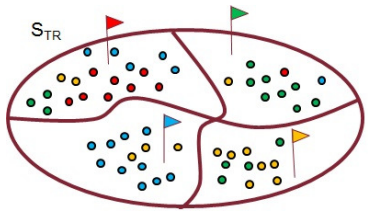

Figura 8: Etichettatura con BA.

Successivamente, il valore della fitness, è calcolato come l'accuratezza di classificazione sul validation set S_{VAL} secondo la f_1 (3.4) e la f_2 (3.5). Gli operatori genetici della riproduzione, del crossover e della mutazione dell'AG, vengono applicati agli individui della popolazione affinché possa evolvere. Questo processo continua fino al raggiungimento del numero di generazioni prefissato che rappresenta, nel nostro caso, il verificarsi del criterio di stop per l'AG.

Come si vedrà successivamente, i test sono stati eseguiti con entrambe le funzioni di fitness ed i loro risultati messi a confronto.

Nella tabella seguente viene mostrata la sintesi dell'algoritmo base BA.

Tabella 19: Sintesi dell'algoritmo base BA

Parametri di Input:

- Numero massimo di cluster: *Kmax*

- Numero di individui della popolazione in AG: *Pop*

- Numero di generazioni di AG: *nGeneration.*

1. Lettura dei dati da S_{TR} ed S_{VAL}.

2. Inizializzazione *(Generation = 0*).

   For *j* = 1 to *Pop*

   - Assegnazione random dei pesi p_i delle 6 features e dei valori $K \leq Kmax$.
   - Clustering degli elementi di S_{TR} in *K* cluster utilizzando la funzione di distanza (*2.1*) con i parametri codificati nell'individuo *j*
   - Valutazione della funzione di fitness (*3.4*) oppure (*3.5*) su S_{VAL}

3. For *q* = 1 to *nGeneration*

   - Applicazione dell'elitismo.
   - Repeat

     – *Selezione* di individui della vecchia popolazione per mezzo dell'operatore *ruota della roulette*.

     – *Crossover* tra coppie di individui selezionati.

     – *Mutazione* con una bassa probabilità su ogni elemento.

     – Clustering di S_{TR} in *K* cluster utilizzando la funzione di distanza (*2.1*) con i parametri codificati nell'individuo.

     – Valutazione della funzione di fitness (*3.4*) oppure (*3.5*) su S_{VAL}

3.5 CBA (Class-aware Basic Algorithm): una variante dell'algoritmo base

L'algoritmo base BA dà luogo alla formazione di cluster contenenti più di una malattia. L'etichetta associata con il cluster, coincide con la patologia più frequente in esso. Questa procedura non può assicurare la presenza di almeno un cluster per ogni classe di malattia. Al fine di assicurarci che tutte le patologie vengano rappresentate nel modello di classificazione finale, è stata progettata una seconda versione del sistema di classificazione precedentemente proposto. A questo scopo, il training set S_{TR} è stato partizionato in sei sottoinsiemi, uno per patologia. Il nuovo algoritmo esegue sei analisi di cluster in parallelo, una per ciascuno dei sei sottoinsiemi di S_{TR}. Il risultato è che ogni cluster conterrà pattern associati con un'unica classe di etichette, di conseguenza sarà direttamente etichettato da quella classe. L'unione dei sei gruppi di cluster etichettati sarà direttamente impiegato per la definizione del modello di classificazione.

Il modo in cui un individuo generico nell' AG è codificato è stato adattato al nuovo algoritmo. In particolare, la seconda parte dell'individuo non contiene più un solo intero ma sei interi distinti, ognuno rappresentante il numero di cluster da generare con l'analisi di clustering eseguita in parallelo su ogni sottoinsieme del training set S_{TR} (uno per ogni classe-etichetta). La fase di inizializzazione per ottenere la prima generazione dell' AG, è simile al precedente algoritmo base BA.

Anche per la presente versione, per la valutazione del valore di fitness di ciascun individuo, vengono considerate entrambe le funzioni di fitness f_1 ed f_2 calcolate sull'insieme di valutazione S_{VAL}.

Il criterio di etichettatura per l'algoritmo CBA può essere rappresentato con la seguente figura, in cui l'etichetta di ciascun cluster è rappresentata da una bandierina di un dato colore rappresentante una patologia. In questo caso, ogni cluster sarà costituito da elementi di una stessa patologia e ciò è evidenziato dal fatto che ciascun cluster contiene elementi tutti dello stesso colore.

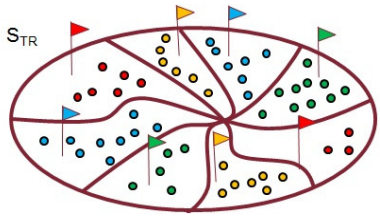

Figura 9: Etichettatura con CBA.

Riassumendo:

- S_{TR} è stato suddiviso in sei sottoinsiemi, uno per patologia.
- Vengono eseguite sei analisi di clustering in parallelo, una per ognuno dei 6 sottoinsiemi.
- Ogni cluster contiene pattern associati con un unica patologia.

3.6 NCBA (Non-exclusive Class-aware Basic Algorithm): una variante non esclusiva dell'algoritmo base.

Entrambi gli algoritmi precedentemente presentati BA e CBA sono stati progettati per affrontare un problema di classificazione di tipo esclusivo, in cui il modello di classificazione finale assegna un'unica *classe-etichetta* al pattern in input, le classi sono considerate mutuamente esclusive. La presenza di fattori di rischio per una particolare malattia logicamente non esclude che lo stesso lavoratore possa sviluppare anche un'altra patologia. Per questa ragione, è stato sviluppato un nuovo approccio di classificazione, dove le *classi-etichetta* sono considerate come *non-esclusive.* A questo scopo, uno specifico modello di classificazione è stato addestrato per ogni patologia (classe) contenuta nel data set in questione. Una volta sintetizzato il modello di classificazione M_i agirà come un riconoscitore per la *i-esima* patologia. La sintesi di M_i si basa su una procedura di ri-etichettatura del training set originale. In particolare, considerata la *i-esima* classe-etichetta ω_i, ogni pattern in S_{TR} originariamente appartenente ad una classe diversa da ω_i verrà associato ad un codice comune (ad esempio "0"), mentre ω_i sarà ricodificato con uno diverso (ad esempio "1").

Sia $S_{TR}^{(i)}$ la ricodifica del training set; M_i sarà sintetizzato dall'algoritmo CBA partendo da $S_{TR}^{(i)}$. Così, ogni cluster conterrà tutti i pattern associati con un'unica classe-etichetta e di conseguenza sarà etichettato dalla stessa. Tutti i modelli di classificazione (sei nel nostro caso) saranno incorporati in un insieme di classificatori in modo da costituire un unico strumento diagnostico generale. I classificatori sono concepiti come un insieme in cui lavorano in parallelo sullo stesso pattern di input, in grado di riconoscere più di una possibile patologia contemporaneamente.

Nella figura successiva è rappresentato l'algoritmo di training.

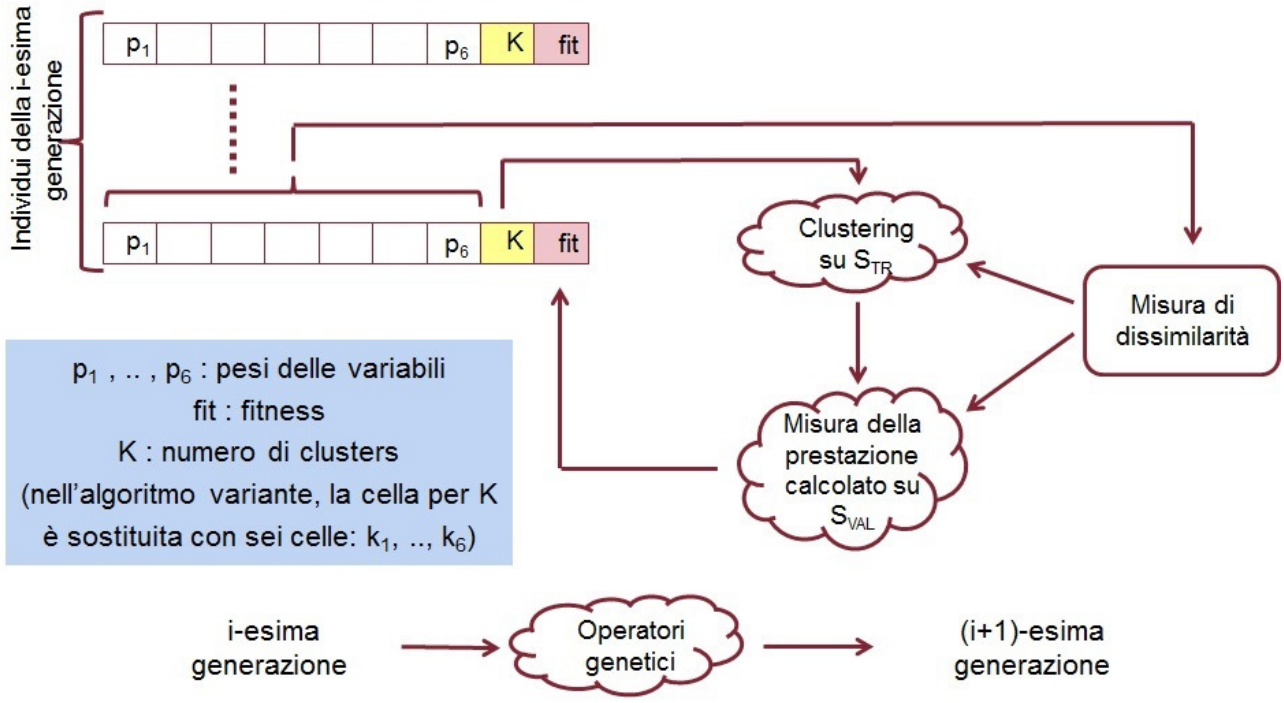

Figura 10: L'algoritmo di training.

4. I test e i risultati

4.1 La realizzazione dei test

I test sono stati condotti su due diversi DB con dati omologhi, uno regionale relativo ai dati della Lombardia ed uno nazionale.

Per il DB regionale, sono state eseguite 8 serie di esperimenti con l'AG. Le prime tre serie sono state realizzate utilizzando come fitness ciascuno degli indici di validazione, le successive 5 serie sono state realizzate utilizzando i nuovi algoritmi, in particolare per queste 5 serie di esperimenti, le prime due con l'algoritmo BA utilizzando entrambe le funzioni di fitness f_1 ed f_2; la terza e la quarta serie eseguite con l'algoritmo CBA utilizzando f_1 ed f_2; ed infine la quinta serie è stata eseguita con l'algoritmo NCBA utilizzando solo la funzione di fitness f_2.

Tutti gli esperimenti delle ultime 5 serie, sono stati realizzati utilizzando una popolazione di 100 individui in 50 generazioni. Il numero massimo di cluster è stato fissato a 20 per gli algoritmi BA e CBA, a 50 per l'algoritmo NCBA. Per l'addestramento iniziale è stato utilizzato l'insieme S_{VAL}, quindi, utilizzando il miglior individuo risultante dall'ottimizzazione genetica, è stata eseguita la classificazione utilizzando il dataset S_{TEST}. I dati riportati in tutte le tabelle che seguono, si riferiscono ai risultati ottenuti utilizzando il miglior individuo trovato dall'ottimizzazione genetica.

Inoltre, con l'algoritmo CBA con funzione di fitness f_2, sono state eseguite 11 elaborazioni in ciascuna delle quali è stato utilizzato un differente seme iniziale per il generatore di numeri casuali utilizzato per la creazione della popolazione iniziale. Sui risultati di queste 11 elaborazioni si è calcolato per ogni malattia sulla totalità delle patologie, il valore medio delle prestazioni, la deviazione standard e i valori minimi e massimi. Il ciclo delle 11 elaborazioni è stato eseguito sull'algoritmo CBA, in quanto ha mostrato prestazioni computazionali migliori rispetto all'algoritmo BA riducendo il tempo di esecuzione di un terzo, poiché le procedure di clustering sono state eseguite su gruppi di dati meno numerosi.

Per quanto riguarda il DB nazionale, i tempi di elaborazione sono aumentati notevolmente, pertanto, sono stati condotti test solo relativamente all'algoritmo NCBA con funzione di fitness pesata f_2, eseguendo per ciascuna patologia 3 serie di test considerando una popolazione di 50 individui per 10 generazioni con un numero massimo di cluster fissato a 50.

I risultati sono stati rappresentati attraverso vari strumenti come grafici, matrici di confusione, tabelle di confusione, sensibilità e specificità.

4.2 AG con indici di validazione

Vengono mostrati i risultati ottenuti con l'esecuzione dell'AG utilizzando come fitness ciascuno degli indici di validazione discussi precedentemente, ossia: indice di Maulik-Bandyopadhyay (indice I), indice di Calinski-Harabasz, indice di Davies-Bouldin.

Per l'esecuzione dei test, sono stati utilizzati gli stessi parametri iniziali per l'AG con ciascuno dei tre indici di validazione. Di seguito, per ciascun indice, vengono mostrati i relativi risultati.

I parametri iniziali:

- Valore Kmax 10
- Numero Record Dataset 3445
- Individui della popolazione 30
- N. Generazioni 10

Dove il valore *Kmax* è il numero massimo di cluster consentito per la realizzazione del partizionamento del dataset utilizzato per il clustering.

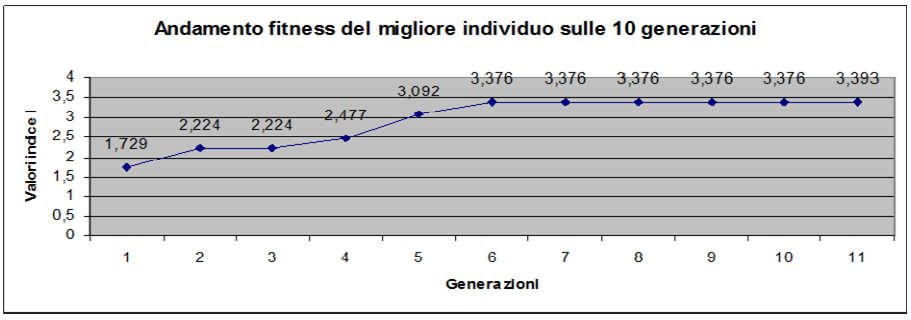

Figura 11: Fitness con indice di Maulik-Bandyopadhyay (indice I)

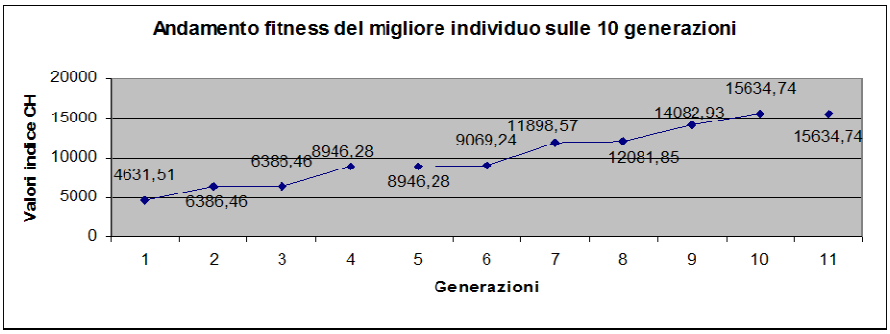

Figura 12: Fitness con indice di Calinski-Harabasz

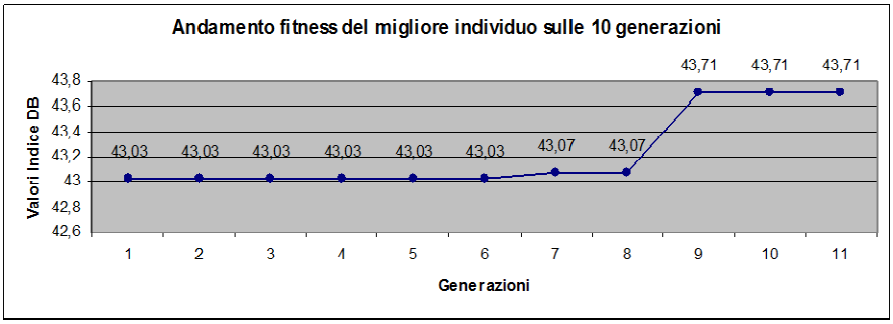

Figura 13: Fitness con indice di Davies-Bouldin

Con tutti e tre i tipi di indice la fitness migliora nelle generazioni e tende a stabilizzarsi dopo poche generazioni.

Il miglior valore di *k* lo si è ottenuto per Maulik-Bandyopadhyay (indice I) in corrispondenza del valore 2, per Calinski-Harabasz nel valore 3 e per l'indice di Davies-Bouldin in corrispondenza del valore 8. Quindi per i primi due indici in corrispondenza dei minimi valori possibili di cluster, mentre per l'ultimo tipo di indice in corrispondenza del massimo numero di cluster possibile per il prefissato parametro iniziale *Kmax.*

Di seguito viene mostrata una tabella riepilogativa dei pesi delle variabili per i migliori individui ottenuti con i tre indici come fitness. Evidenziati in grigio chiaro i valori più rilevanti.

Tabella 20: Pesi delle variabili dei migliori individui sui tre indici.

		I	**DB**	**CH**
[0]	**ETA' ANNO CERTIFICATO**	0,118998	**0,957166949752718**	0,044446
[1]	**MESI DURATA ATTIVITA'**	0,004783	0,109861750621349	0
[2]	**ETA' INIZIO MANSIONE**	**0,75774**	**0,895791934113950**	**0,438744**
[3]	**GENERE**	**0,992881**	**0,715838963175192**	**0,992881**
[4]	**CODICE MANSIONE SPECIFICA**	**0,967507**	**0,950222045416012**	0,149294
[5]	**CODICE ATECO**	**0,996461**	**0,807928615007550**	0,007774

Si noti che per la variabile *"MESI DURATA ATTIVITA'"*, si sono ottenuti dei valori nulli o prossimi a *0* per tutti e tre gli indici, risultando quindi la variabile meno influente delle 6 considerate. Un peso considerevole, per tutti e tre gli algoritmi, lo si è ottenuto per le due variabili *"ETA' INIZIO MANSIONE"* e *"GENERE"*. Dei tre indici utilizzati come fitness, l'indice di Davis-Bouldin (DB), è quello per il quale le variabili sono tutte significative tranne *"MESI DURATA ATTIVITA'"*.

Nella tabella successiva viene mostrata la distribuzione di ciascuna malattia sia sull'intero dataset che per alcuni cluster appartenenti all'ultima partizione ottenuta utilizzando come fitness l'indice DB. Viene mostrato sia il valore assoluto che quello relativo in percentuale. Nella prima colonna viene mostrata la distribuzione per l'intero dataset, nelle altre tre colonne quella per ciascuno dei cluster 3, 6 e 7. Sono stati considerati solo i tre cluster contenenti i risultati più significativi.

Tabella 21: Distribuzione malattie su intero dataset e per cluster - con indice DB.

Patologia	**Popolaz. (%)**	**Cluster 3 (%)**	**Cluster 6 (%)**	**Cluster 7 (%)**
Sordità	1503 (43,6)	23 (7,2)	17 (15,2)	**413 (64,1)**
Malattie rachide	334 (9,7)	**76 (23,8)**	7 (6,2)	25 (3,9)
Muscoloscheletriche	291(8,4)	**85 (26,6)**	12 (10,7)	22 (3,4)
Tumori maligni pleura	233 (6,8)	20 (6,3)	5 (4,5)	22 (3,4)
Tunnel carpale	200 (5,8)	**59 (18,5)**	11 (9,8)	6 (0,9)
Malattie pelle	176 (5,1)	15 (4,7)	2 (1,8)	18 (2,8)
Disturbi orecchio (no sordità)	137 (4,0)	0	0	51 (7,9)
Malattie psichiche	98 (2,8)	9 (2,8)	**43 (38,4)**	12 (1,9)
Altre	473 (13,8)	32 (10,1)	15 (13,4)	75 (11,7)
Totale	**3445 (100)**	**319 (100)**	**112 (100)**	**644 (100)**

Nella tabella successiva, vengono mostrati i risultati rappresentati dai cluster 3, 6 e 7 appartenenti all'ultima partizione ottenuta utilizzando come fitness l'indice DB. Viene mostrata la distribuzione delle *mansioni professionali* sull'intero dataset e per ciascun cluster sia in valore assoluto che relativo (valore percentuale). Per i cluster 3 e 7 la mansione più frequente è quella degli *"Artigiani operai specializzati e agricoltori"* mentre per il cluster 6 la mansione più significativa è quella delle *"Professioni esecutive relative all'amministrazione e gestione"*

Tabella 22: Distribuzione mansioni professionali su intero dataset e per cluster - con indice DB.

	DATASET		**cluster 3**		**cluster 6**		**cluster 7**	
Mansione Specifica	**Pop.**	**%**	**Pop.**	**%**	**Pop.**	**%**	**Pop.**	**%**
ARTIGIANI, OPERAI SPECIALIZZATI E AGRICOLTORI	**1910**	**55,4**	**146**	**45,8**	3	2,7	**482**	**74,8**
CONDUTTORI DI IMPIANTI, OPERATORI DI MACCHINARI FISSI E MOBILI (ANCHE IN AGRICOLTURA) E OPERAI DI MONTAGGIO INDUSTRIALE	590	17,1	2	0,6	2	1,8	121	18,8
PERSONALE NON QUALIFICATO	401	11,6	86	27,0	6	5,4	8	1,2
PROFESSIONI RELATIVE ALLE VENDITE ED AI SERVIZI PER LE FAMIGLIE	210	6,1	48	15,0	8	7,1	7	1,1
PROFESSIONI INTERMEDIE (TECNICI)	169	4,9	18	5,6	10	8,9	15	2,3
PROFESSIONI ESECUTIVE RELATIVE ALL'AMMINISTRAZIONE E GESTIONE	**90**	**2,6**	4	1,3	**79**	**70,5**	1	0,2
PROFESSIONI INTELLETTUALI, SCIENTIFICHE E DI ELEVATA SPECIALIZZAZIONE	56	1,6	14	4,4	2	1,8	6	0,9
LEGISLATORI, DIRIGENTI E IMPRENDITORI	17	0,5	1	0,3	1	0,9	4	0,6
FORZE ARMATE	2	0,1	0	0	1	0,9	0	0
TOTALE	**3445**	**100**	**319**	**100**	**112**	**100**	**644**	**100**

4.3 Algoritmo BA

L'algoritmo BA è stato applicato alla base dati regionale utilizzando la funzione di fitness f_1 ed f_2. I risultati illustrati sono quelli ottenuti in corrispondenza del miglior individuo e sono stati rappresentati in forma di *matrice di confusione*, insieme ad ogni risultato sono indicati tra parentesi i valori normalizzati delle percentuali del totale delle occorrenze di riga. Sulla diagonale della matrice sono stati evidenziati i valori predetti in modo corretto, si osservi che il valore più alto tra essi è stato ottenuto in corrispondenza della patologia *"1 - sordità"*. Le patologie vengono indicate utilizzando la codifica delle etichette introdotte nella Tabella 17.

Tabella 23: Matrice di confusione per algoritmo BA con f_1 su DB regionale

		Patologia predetta					
		1	**2**	**3**	**4**	**5**	**6**
Patologia reale	**1**	**351** **(94,1)**	5 (1,3)	0 (0)	7 (1,9)	9 (2,4)	1 (0,3)
	2	45 (54,2)	**36** **(43,4)**	0 (0)	0 (0)	1 (1,2)	1 (1,2)
	3	32 (44,4)	20 (27,8)	**0** **(0)**	0 (0)	16 (22,2)	4 (5,6)
	4	18 (31,0)	8 (13,8)	0 (0)	**32** **(55,2)**	0 (0)	0 (0)
	5	9 (18,0)	16 (32,0)	0 (0)	1 (2)	**23** **(46)**	1 (2)
	6	20 (45,5)	7 (15,9)	0 (0)	0 (0)	4 (9,1)	**13** **(29,5)**

Dalla matrice di confusione, per ognuna delle 6 patologie, è possibile estrarre ulteriori viste in forma di *tabella di confusione.* Le tabelle di confusione permettono una migliore comprensione dei risultati rispetto alla semplice indicazione di percentuale di risposte corrette (accuratezza). Dato il contenuto del dataset, formato solo da lavoratori affetti almeno da una patologia, per ogni malattia, vengono considerati come sani quei lavoratori non affetti da quella patologia. Di seguito viene rappresentato lo schema tipo di una tabella di confusione, le colonne "Test Positivo" e "Test Negativo", contengono il numero di lavoratori che l'algoritmo predice rispettivamente come *malati* (ossia affetti dalla malattia in questione), o come *sani* (ossia affetti da un'altra malattia). Le righe "Vero reale" e "Falso reale" contengono il numero di coloro che realmente sono, rispettivamente, *malati* e *sani.*

Tabella 24: Schema della tabella di confusione per valutare la capacità predittiva di un test.

	Test Positivo	**Test Negativo**
Vero reale	*Veri positivi*	Falsi negativi
Falso reale	Falsi positivi	*Veri negativi*

Di seguito viene mostrata la *tabella di confusione* per la patologia *"1 – sordità"* dove i valori delle righe, sono stati indicati come "Veri malati" e "Veri sani", ed i valori delle colonne come "Malato predetto" e "Sano predetto". Tra parentesi sono indicati i valori normalizzati delle percentuali del totale delle occorrenze di riga.

Patologia predetta				
2	**3**	**4**	**5**	**6**
5	0	7	9	1

Tabella 25: Tabella di confusione per la patologia *"1 – sordità"*

	Malato predetto	**Sano predetto**	**Totale**
Vero malato	*351* *(94,1%)* *Veri positivi*	22 (5,9%) Falsi negativi	373
Vero sano	124 (40,4%) Falsi positivi	*183* *(59,6%)* *Veri negativi*	307

Patologia reale	
2	45
3	32
4	18
5	9
6	20

Si noti che il valore dei "Falsi positivi" è dato dalla somma dei restanti valori di "Patologia reale" presi nella colonna della "Patologia predetta" in questione (1 - sordità); il valore dei "Falsi negativi" è dato dalla somma dei restanti valori di "Patologia predetta" presi nella riga della "Patologia reale" in questione (1 - sordità); infine il valore dei "Veri negativi" è dato dalla somma degli altri restanti valori. Allo stesso modo si ricavano le tabelle di confusione per le restanti patologie.

Nella tabella seguente sono mostrati i risultati dell'algoritmo BA eseguito con f_2 rappresentati come *matrice di confusione*, anche qui, tra parentesi sono indicati i valori normalizzati delle percentuali del totale delle occorrenze di riga. Si noti che in questo caso per la patologia *"3 - malattie muscoloscheletriche"*, si sono ottenute delle predizioni che non sono state ottenute nel caso precedente con fitness f_1.

Tabella 26: Matrice di confusione per algoritmo BA con f_2

		Patologia predetta					
		1	**2**	**3**	**4**	**5**	**6**
Patologia reale	**1**	**338 (90,6)**	7 (1,9)	7 (1,9)	15 (4)	6 (1,6)	0 (0)
	2	36 (43,4)	**35 (42,2)**	4 (4,8)	0 (0)	7 (8,4)	1 (1,2)
	3	32 (44,4)	12 (16,7)	**14 (19,4)**	0 (0)	13 (18,1)	1 (1,4)
	4	18 (31)	1 (1,7)	7 (12,1)	**32 (55,2)**	0 (0)	0 (0)
	5	8 (16)	6 (12)	12 (24)	2 (4)	**21 (42)**	1 (2)
	6	21 (47,7)	3 (6,8)	0 (0)	0 (0)	8 (18,2)	**12 (27,3)**

Di seguito sono mostrati gli *schemi di sintesi* delle 6 tabelle di confusione dell'algoritmo BA eseguito sia con f_1 che con f_2 e il valore medio sulle 6 patologie. Nello schema, ogni colonna rappresenta una tabella di confusione per quella data patologia. Per entrambi gli esperimenti il massimo numero di cluster è stato fissato a 20. La tabella seguente mostra i risultati ottenuti con l'esperimento utilizzando f_1.

Tabella 27: Schema di sintesi delle 6 tabelle di confusione e valori medi per BA con f_1

	Patologia						Valori medi
	1	**2**	**3**	**4**	**5**	**6**	
Veri positivi	351	36	0	32	23	13	75,8
Falsi positivi	124	56	0	8	30	7	37,5
Falsi negativi	22	47	72	26	27	31	37,5
Veri negativi	183	541	608	614	600	629	529,2

Nella tabella seguente sono mostrati i risultati ottenuti con l'esperimento utilizzando f_2.

Tabella 28: Schema di sintesi delle 6 tabelle di confusione e valori medi per BA con f_2

	Patologia						Valori medi
	1	**2**	**3**	**4**	**5**	**6**	
Veri positivi	338	35	14	32	21	12	75,3
Falsi positivi	115	29	30	17	34	3	38
Falsi negativi	35	48	58	26	29	32	38
Veri negativi	192	568	578	605	596	633	528,7

4.4 Algoritmo CBA

Negli esperimenti basati sull'algoritmo CBA eseguiti sia con fitness f_1 che con fitness f_2, il numero massimo di cluster è stato fissato a 20 per ogni patologia. In corrispondenza dell'individuo migliore, il numero totale di cluster è stato 36 con la seguente distribuzione per classe: 20 con etichetta *"1 - sordità"*, 6 con etichetta *"2 - Malattie del rachide"*, 2 con etichetta *"3 - Malattie muscolo scheletriche"*, 4 con etichetta *"4 - Tumori pleura e peritoneo"*, 2 con etichetta *"5 - Sindrome tunnel carpale"* e 2 con etichetta *"6 - Malattie della pelle"*.

I risultati vengono mostrati in forma di schema di sintesi di *tabelle di confusione*. In particolare, di seguito vengono mostrati i risultati ottenuti con la funzione di fitness f_1.

Tabella 29: Schema di sintesi delle 6 tabelle di confusione per CBA con f_1

	Patologia						Valori medi
	1	**2**	**3**	**4**	**5**	**6**	
Veri positivi	312	42	12	35	1	19	70,2
Falsi positivi	127	59	19	15	20	19	43,2
Falsi negativi	61	41	60	23	49	25	43,2
Veri negativi	180	538	589	607	610	617	523,5

Anche per l'esperimento eseguito con l'algoritmo CBA con fitness f_2 come per quello eseguito con fitness f_1, il numero massimo di cluster è stato fissato a 20 per ogni patologia. In corrispondenza dell'individuo migliore, il numero totale di cluster è stato 71 con la seguente distribuzione per classe: 18 con etichetta *"1 - sordità"*, 10 con etichetta *"2 - Malattie del rachide"*, 7 con etichetta *"3 - Malattie muscolo scheletriche"*, 11 con etichetta *"4 - Tumori pleura e peritoneo"*, 15 con etichetta *"5 - Sindrome tunnel carpale"* e 10 con etichetta *"6 - Malattie della pelle"*. I risultati sono mostrati di seguito attraverso lo schema di sintesi delle sei *tabelle di confusione*.

Tabella 30: Schema di sintesi delle 6 tabelle di confusione per CBA con f_2

	Patologia						Valori medi
	1	**2**	**3**	**4**	**5**	**6**	
Veri positivi	210	44	17	35	30	25	60,2
Falsi positivi	48	61	45	30	89	46	53,2
Falsi negativi	163	39	55	23	20	19	53,2
Veri negativi	259	536	563	592	541	590	513,5

Un altro strumento importante per l'analisi delle prestazioni, comunemente utilizzato nella valutazione dei test diagnostici da parte dei medici del lavoro, consiste nel calcolo della *sensitività* e della *specificità* e del *valore predittivo negativo* e *positivo*. Prendiamo in considerazione l'analisi dei risultati attraverso questo nuovo strumento che testa la popolazione di lavoratori per una data malattia. Il test può risultare *positivo* (predicendo che l'individuo in esame è affetto dalla malattia in questione), oppure può risultare *negativo* (predicendo che l'individuo è sano). Il risultato del test può o non può corrispondere all'attuale stato di salute del soggetto in esame. Si consideri la seguente impostazione:

- Veri positivi: individui malati correttamente diagnosticati come malati
- Falsi positivi: individui sani erroneamente diagnosticati come malati
- Veri negativi: individui sani correttamente diagnosticati come sani
- Falsi negativi: individui malati erroneamente diagnosticati come sani.

I quattro risultati possono essere espressi per mezzo di una tabella di confusione 2x2 come nella tabella seguente:

Tabella 31: Tabella di confusione

	Condizione Positiva	**Condizione Negativa**
Test Positivo	*Veri positivi*	Falsi positivi
Test Negativo	Falsi negativi	*Veri negativi*

Nella tabella seguente sono definiti gli indicatori usati nei test diagnostici:

Tabella 32: Indicatori test diagnostici

Sensibilità	Veri positivi / Σ Condizioni positive
Specificità	Veri negativi / Σ Condizioni negative
Valore predittivo negativo	Veri negativi / Σ Test negativi
Valore predittivo positivo	Veri positivi / Σ Test positivi

Nella tabella seguente sono riassunti i risultati per *sensibilità e specificità* per patologia rappresentati con il valore medio, è illustrata anche la deviazione standard ed il valore minimo e massimo per ciascuna patologia risultanti dopo l'esecuzione di 11 cicli di elaborazioni utilizzando la funzione di fitness f_2.

Tabella 33: Sensibilità e specificità per CBA per patologia dopo 11 elaborazioni con f_2

	Patologia	**1**	**2**	**3**	**4**	**5**	**6**
	Pattern per patologia	**373**	**83**	**72**	**58**	**50**	**44**
Sensibilità	Media	0,482	0,567	0,158	0,718	0,633	0,610
	St. dev.	0,059	0,047	0,082	0,058	0,061	0,063
	Min	0,375	0,470	0,083	0,603	0,560	0,523
	Max	0,563	0,614	0,306	0,810	0,760	0,727
Specificità	Media	0,875	0,871	0,948	0,926	0,873	0,892
	St. dev.	0,017	0,022	0,039	0,025	0,021	0,019
	Min	0,844	0,841	0,863	0,897	0,846	0,866
	Max	0,899	0,899	0,979	0,968	0,903	0,928

Nella tabella seguente sono riassunti i risultati per i *valori predittivi positivi e negativi* per patologia rappresentati con il valore medio, è illustrata anche la deviazione standard ed il valore minimo e massimo per ciascuna patologia risultanti dopo l'esecuzione di 11 cicli di elaborazioni utilizzando la funzione di fitness f_2.

Tabella 34: Valori predittivi per CBA per patologia risultanti dopo 11 elaborazioni con f_2

	Patologia	**1**	**2**	**3**	**4**	**5**	**6**
	Pattern per patologia	**373**	**83**	**72**	**58**	**50**	**44**
Valore predittivo negativo = Veri negativi / Σ Test negativi	Media	0,58	0,94	0,90	0,97	0,97	0,97
	St. dev.	0,02	0,01	0,01	0,01	0,01	0,00
	Min	0,54	0,92	0,90	0,96	0,96	0,96
	Max	0,61	0,94	0,91	0,98	0,98	0,98
Valore predittivo positivo = Veri positivi / Σ Test positivi	Media	0,82	0,38	0,26	0,48	0,28	0,28
	St. dev.	0,01	0,03	0,05	0,09	0,03	0,04
	Min	0,81	0,34	0,21	0,40	0,22	0,23
	Max	0,85	0,43	0,35	0,67	0,34	0,35

4.5 Algoritmo NCBA

La realizzazione di tale algoritmo è stata effettuata utilizzando solo la funzione di fitness f_2. Questo algoritmo viene realizzato attraverso un distinto classificatore per ogni malattia. I test eseguiti su ciascuna malattia sono stati realizzati considerando la parte complementare, contenente il resto dei lavoratori affetti da altre patologie, considerati come lavoratori sani. In tale algoritmo per ogni classificatore, il numero massimo di cluster è stato fissato a 50 sia per i soggetti malati che per quelli sani.

Di seguito vengono mostrati i risultati dei test eseguiti con il classificatore per ciascuna malattia attraverso lo schema di sintesi delle sei *tabelle di confusione*.

Tabella 35: Schema di sintesi delle 6 tabelle di confusione per NCBA con f_2

	Patologia						Valori medi
	1	**2**	**3**	**4**	**5**	**6**	
Veri positivi	314	62	56	54	44	38	94,7
Falsi positivi	84	88	204	73	116	110	112,5
Falsi negativi	59	21	16	4	6	6	18,7
Veri negativi	223	509	404	549	514	526	454,2

Nelle tabelle seguenti dalla 36 alla 41 vengono mostrati gli indicatori diagnostici dei risultati ottenuti per ciascuna malattia con l'algoritmo NCBA con funzione di fitness f_2 e confrontati con quelli ottenuti con l'algoritmo BA e l'algoritmo CBA con funzione di fitness f_1 *ed* f_2.

Tabella 36: Indicatori diagnostici dei test per la patologia "1 - sordità"

	BA, f1	**BA, f2**	**CBA, f1**	**CBA, f2**	**NCBA, f2**
Sensibilità	0,941	0,906	0,836	0,563	0,842
Specificità	0,596	0,625	0,586	0,844	0,726
Valore predittivo negativo	0,893	0,846	0,747	0,614	0,791
Valore predittivo positivo	0,739	0,746	0,711	0,814	0,789
Media	**0,792**	**0,781**	**0,720**	**0,709**	**0,787**

Tabella 37: Indicatori diagnostici dei test per la patologia "2 - rachide"

	BA, f1	**BA, f2**	**CBA, f1**	**CBA, f2**	**NCBA, f2**
Sensibilità	0,434	0,422	0,506	0,530	0,747
Specificità	0,906	0,951	0,901	0,898	0,853
Valore predittivo negativo	0,920	0,922	0,929	0,932	0,960
Valore predittivo positivo	0,391	0,547	0,416	0,419	0,413
Media	**0,663**	**0,711**	**0,688**	**0,695**	**0,743**

Tabella 38: Indicatori diagnostici dei test per la patologia "3 - malattie muscolo scheletriche"

	BA, f1	BA, f2	CBA, f1	CBA, f2	NCBA, f2
Sensibilità	0,000	0,194	0,167	0,236	0,778
Specificità	1,000	0,951	0,969	0,926	0,664
Valore predittivo negativo	0,894	0,909	0,908	0,911	0,962
Valore predittivo positivo	0,000	0,318	0,387	0,274	0,215
Media	**0,474**	**0,593**	**0,608**	**0,587**	**0,655**

Tabella 39: Indicatori diagnostici dei test per la patologia "4 - tumori della pleura e peritoneo"

	BA, f1	BA, f2	CBA, f1	CBA, f2	NCBA, f2
Sensibilità	0,552	0,552	0,603	0,603	0,931
Specificità	0,987	0,973	0,976	0,952	0,883
Valore predittivo negativo	0,959	0,959	0,963	0,963	0,993
Valore predittivo positivo	0,800	0,653	0,700	0,538	0,425
Media	**0,825**	**0,784**	**0,811**	**0,764**	**0,808**

Tabella 40: Indicatori diagnostici dei test per la patologia "5 - tunnel carpale"

	BA, f1	BA, f2	CBA, f1	CBA, f2	NCBA, f2
Sensibilità	0,460	0,420	0,020	0,600	0,880
Specificità	0,952	0,946	0,968	0,859	0,816
Valore predittivo negativo	0,957	0,954	0,926	0,964	0,988
Valore predittivo positivo	0,434	0,382	0,048	0,252	0,275
Media	**0,701**	**0,675**	**0,490**	**0,669**	**0,740**

Tabella 41: Indicatori diagnostici dei test per la patologia "6 - malattie della pelle"

	BA, f1	BA, f2	CBA, f1	CBA, f2	NCBA, f2
Sensibilità	0,295	0,273	0,432	0,568	0,864
Specificità	0,989	0,995	0,970	0,928	0,827
Valore predittivo negativo	0,953	0,952	0,961	0,969	0,989
Valore predittivo positivo	0,650	0,800	0,500	0,352	0,257
Media	**0,722**	**0,755**	**0,716**	**0,704**	**0,734**

Nella tabella seguente sono riassunti i valori medi al fine di una migliore comprensione nel confronto degli algoritmi considerati.

Tabella 42: Indicatori diagnostici dei test relativi ai valori medi

	BA, f1	**BA, f2**	**CBA, f1**	**CBA, f2**	**NCBA, f2**
Sensibilità	0,447	0,461	0,427	0,517	0,840
Specificità	0,905	0,907	0,895	0,901	0,795
Valore predittivo negativo	0,929	0,923	0,906	0,892	0,947
Valore predittivo positivo	0,503	0,574	0,460	0,442	0,396
Media	**0,696**	**0,716**	**0,672**	**0,688**	**0,744**

4.6 AG su base dati nazionale

Gli algoritmi di classificazione fin qui presentati e applicati al dataset costituito solo dai dati della regione Lombardia, hanno prodotto un insieme di risultati utilizzabile come base di confronto.

L'algoritmo NCBA è stato applicato ad un dataset di dati più esteso comprendente dati di tipo analogo a quelli della regione Lombardia, raccolti e registrati a livello nazionale da parte di tutti i servizi di prevenzione delle Aziende Sanitarie Locali (ASL) di tutte le regioni italiane secondo il modello strutturato MALPROF. L'estensione del dataset, ha riguardato solo la variazione del numero di record (pattern) trattati e non la loro struttura sia per quanto riguarda il tipo di dati che il numero delle variabili già considerate includendo i lavoratori e le relative patologie di tutte le regioni del territorio italiano registrati nella base dati nazionale dell'INAIL.

I dati costituenti il nuovo dataset nazionale, trattati nel rispetto della normativa del codice in materia di protezione dei dati personali dl 30 giugno 2003, n. 196 (legge sulla privacy), hanno consentito l'ottenimento di interessanti risultati nell'ambito della prevenzione sanitaria nei luoghi di lavoro, e che confrontati con i risultati ottenuti nelle precedenti elaborazioni sui dati della regione Lombardia hanno consentito la conferma di alcuni risultati e chiarito altri per i quali si erano presentati elementi di ambiguità e quindi di impossibilità di ricavare delle informazioni utili e interessanti ai fini della prevenzione riguardo le malattie lavoro-correlate.

Di seguito vengono mostrati i risultati ottenuti con l'algoritmo NCBA con utilizzo della funzione di fitness pesata f_2 utilizzando il DB nazionale e confrontati con quelli ottenuti sul DB regionale. Sono stati considerati i valori dei migliori individui tra tutte le elaborazioni effettuate per ciascun DB.

Di seguito lo schema di sintesi delle sei *tabelle di confusione* dei valori ottenuti con l'algoritmo NCBA su DB nazionale utilizzando la funzione di fitness f_2. I valori si riferiscono al miglior individuo ottenuto su tutte le generazioni di tutti i cicli di elaborazione eseguiti.

Tabella 43: Schema di sintesi delle 6 tabelle di confusione per NCBA con f_2 su DB nazionale

	Patologia					
	1	**2**	**3**	**4**	**5**	**6**
Veri positivi	604	292	466	136	218	47
Falsi positivi	693	700	729	211	551	241
Falsi negativi	139	162	199	14	110	31
Veri negativi	982	1264	1024	2057	1539	2099

Di seguito lo schema di sintesi delle sei *tabelle di confusione* dei valori medi ottenuti utilizzando la funzione di fitness f_2. La media è stata calcolata sui valori in corrispondenza del miglior individuo di ciascun ciclo di elaborazione.

Tabella 44: Schema di sintesi dei valori medi per NCBA con f_2 su DB nazionale

	Patologia					
	1	**2**	**3**	**4**	**5**	**6**
Veri positivi	618,333	288,667	485,333	132,667	218	51,667
Falsi positivi	747,333	697,667	810	206,333	560,333	403,333
Falsi negativi	124,667	165,333	179,667	17,333	110	26,333
Veri negativi	927,667	1266,333	943	2061,667	1529,667	1936,667

Nella tabella seguente viene mostrato il confronto dei risultati ottenuti con l'algoritmo NCBA utilizzando la funzione di fitness f_2 eseguito su base dati nazionale (Nat) e lo stesso algoritmo e stesso tipo di fitness su base dati regionale (Reg). I valori si riferiscono al miglior individuo ottenuto su tutte le generazioni di tutti i cicli di elaborazione eseguiti.

Tabella 45: NCBA con f_2: confronto dei risultati su base dati nazionale e regionale

	Patologia											
	1		**2**		**3**		**4**		**5**		**6**	
	NCBA											
	Nat	**Reg**	**Nat**	**Reg**	**Nat**	**Reg**	**Nat**	**Reg**	**Nat**	**Reg**	**Nat**	**Reg**
Sensib.	0,813	0,842	0,643	0,747	0,701	0,778	0,907	0,931	0,665	0,880	0,603	0,864
Specif.	0,586	0,726	0,644	0,853	0,584	0,664	0,907	0,883	0,736	0,816	0,897	0,827
Valore pred. neg.	0,876	0,791	0,886	0,960	0,837	0,962	0,993	0,993	0,933	0,988	0,985	0,989
Valore pred. pos.	0,466	0,789	0,294	0,413	0,390	0,215	0,392	0,425	0,283	0,275	0,163	0,257
Media dei 4 param.	**0,685**	**0,787**	**0,617**	**0,743**	**0,628**	**0,655**	**0,800**	**0,808**	**0,654**	**0,740**	**0,662**	**0,734**

Nella tabella seguente sono riassunti i risultati per i *valori predittivi positivi e negativi* per patologia rappresentati con il valore medio, la deviazione standard ed il valore minimo e massimo per ciascuna patologia risultanti dopo l'esecuzione di 3 cicli di elaborazioni per NCBA utilizzando la funzione di fitness f_2 su DB nazionale.

Tabella 46: Valori predittivi negativi e positivi per patologia per NCBA con f_2 su DB nazionale

	Patologia	**1**	**2**	**3**	**4**	**5**	**6**	**Totale**
	Patterns	**743**	**454**	**665**	**150**	**328**	**78**	**2418**
Valore predittivo negativo = Veri negativi / Σ Test negativi	Media	0,882	0,885	0,840	0,992	0,933	0,987	
	St. dev.	0,032	0,003	0,028	0,001	0,000	0,001	
	Min	0,859	0,881	0,818	0,991	0,932	0,985	
	Max	0,920	0,887	0,873	0,993	0,933	0,988	
Valore predittivo positivo = Veri positivi / Σ Test positivi	Media	0,453	0,293	0,375	0,391	0,280	0,114	
	St. dev.	0,019	0,002	0,013	0,036	0,003	0,036	
	Min	0,433	0,290	0,367	0,359	0,278	0,096	
	Max	0,466	0,294	0,390	0,430	0,283	0,163	

4.7 Confronto SVM, BA, CBA e NCBA

I risultati ottenuti con gli algoritmi BA, CBA ed NCBA su DB regionale sono stati messi a confronto con quelli ottenuti con l'ottimizzazione genetica eseguita attraverso SVM (Support Vector Machine). Di seguito sono mostrate le tabelle che mettono a confronto tali risultati. Viene mostrata per prima la tabella riassuntiva dei risultati a confronto ottenuti da NCBA ed SVM utilizzando la fitness f_2, successivamente, con una serie di tabelle riassuntive, vengono illustrati i risultati ottenuti da BA e CBA con f_1 ed f_2 e confrontati con quelli ottenuti con SVM.

Di seguito la tabella riassuntiva dei risultati a confronto ottenuti da NCBA ed SVM utilizzando la fitness f_2 su base regionale.

Tabella 47: Confronto NCBA con SVM con f_2 su DB regionale.

	Patologia											
	1		**2**		**3**		**4**		**5**		**6**	
	NCBA f_2	**SVM f_2**	**NCBA f_2**	**SVM f_2**	**NCBA f_2**	**SVM f_2**	**NCBA f_2**	**SVM f_2**	**NCBA f_2**	**SVM f_2**	**NCBA f_2**	**SVM f_2**
Sensib.	0,842	0,83	0,747	0,25	0,778	0,08	0,931	0,65	0,88	0,04	0,864	0,23
Specif.	0,726	0,79	0,853	0,98	0,664	0,99	0,883	0,99	0,816	1	0,827	0,99
Valore pred. neg.	0,791	0,8	0,96	0,9	0,962	0,9	0,993	0,97	0,988	0,93	0,989	0,95
Valore pred. pos.	0,789	0,83	0,413	0,64	0,215	0,46	0,425	0,89	0,275	0,57	0,257	0,64
Media dei 4 param.	**0,787**	**0,813**	**0,743**	**0,693**	**0,655**	**0,608**	**0,808**	**0,875**	**0,74**	**0,635**	**0,734**	**0,703**

Nelle tabelle seguenti dalla 48 alla 53 vengono mostrati gli indicatori diagnostici dei risultati ottenuti per ciascuna malattia con gli algoritmi BA e CBA con funzione di fitness f_1 *ed* f_2 e confrontati con quelli ottenuti con SVM con funzione di fitness f_1 *ed* f_2.

Tabella 48: Indicatori diagnostici dei test per la patologia "1 - sordità"

	BA, f1	**BA, f2**	**CBA, f1**	**CBA, f2**	**SVM, f1**	**SVM, f2**
Sensibilità	0,941	0,906	0,836	0,563	0,96	0,95
Specificità	0,596	0,625	0,586	0,844	0,65	0,62
Valore predittivo negativo	0,893	0,846	0,747	0,614	0,77	0,75
Valore predittivo positivo	0,739	0,746	0,711	0,814	0,93	0,92
Media	**0,792**	**0,781**	**0,720**	**0,709**	**0,828**	**0,810**

Tabella 49: Indicatori diagnostici dei test per la patologia "2 - rachide"

	BA, f1	BA, f2	CBA, f1	CBA, f2	SVM, f1	SVM, f2
Sensibilità	0,434	0,422	0,506	0,530	0,45	0,43
Specificità	0,906	0,951	0,901	0,898	0,95	0,95
Valore predittivo negativo	0,920	0,922	0,929	0,932	0,57	0,56
Valore predittivo positivo	0,391	0,547	0,416	0,419	0,92	0,92
Media	**0,663**	**0,711**	**0,688**	**0,695**	**0,723**	**0,715**

Tabella 50: Indicatori diagnostici dei test per la patologia "3 - malattie muscolo scheletriche"

	BA, f1	BA, f2	CBA, f1	CBA, f2	SVM, f1	SVM, f2
Sensibilità	0,000	0,194	0,167	0,236	0,20	0,19
Specificità	1,000	0,951	0,969	0,926	0,96	0,95
Valore predittivo negativo	0,894	0,909	0,908	0,911	0,36	0,33
Valore predittivo positivo	0,000	0,318	0,387	0,274	0,91	0,91
Media	**0,474**	**0,593**	**0,608**	**0,587**	**0,608**	**0,595**

Tabella 51: Indicatori diagnostici dei test per la patologia "4 - tumori della pleura e peritoneo"

	BA, f1	BA, f2	CBA, f1	CBA, f2	SVM, f1	SVM, f2
Sensibilità	0,552	0,552	0,603	0,603	0,78	0,75
Specificità	0,987	0,973	0,976	0,952	0,99	0,99
Valore predittivo negativo	0,959	0,959	0,963	0,963	0,89	0,88
Valore predittivo positivo	0,800	0,653	0,700	0,538	0,98	0,98
Media	**0,825**	**0,784**	**0,811**	**0,764**	**0,910**	**0,900**

Tabella 52: Indicatori diagnostici dei test per la patologia "5 - tunnel carpale"

	BA, f1	BA, f2	CBA, f1	CBA, f2	SVM, f1	SVM, f2
Sensibilità	0,460	0,420	0,020	0,600	0,26	0,23
Specificità	0,952	0,946	0,968	0,859	0,96	0,97
Valore predittivo negativo	0,957	0,954	0,926	0,964	0,34	0,38
Valore predittivo positivo	0,434	0,382	0,048	0,252	0,94	0,94
Media	**0,701**	**0,676**	**0,491**	**0,669**	**0,625**	**0,630**

Tabella 53: Indicatori diagnostici dei test per la patologia "6 - malattie della pelle"

	BA, f1	BA, f2	CBA, f1	CBA, f2	SVM, f1	SVM, f2
Sensibilità	0,295	0,273	0,432	0,568	0,26	0,30
Specificità	0,989	0,995	0,970	0,928	0,99	0,98
Valore predittivo negativo	0,953	0,952	0,961	0,969	0,58	0,56
Valore predittivo positivo	0,650	0,800	0,500	0,352	0,95	0,95
Media	**0,722**	**0,755**	**0,716**	**0,704**	**0,695**	**0,698**

4.8 Confronto codici genetici

Si ricordi che i codici genetici sono rappresentati per ciascun algoritmo da vettori costituiti dai pesi delle 6 variabili e dal numero di cluster. Di seguito, vengono rappresentati e confrontati i *codici genetici* dei migliori individui prodotti dall'AG per gli algoritmi BA e CBA su data base regionale. I pesi delle variabili sono normalizzati ed il numero massimo di cluster considerato è stato pari a 20.

Tabella 54: Confronto codice genetico per BA e CBA su DB regionale

	Algoritmo BA usando f_1	**Algoritmo BA usando f_2**	**Algoritmo CBA usando f_1**	**Algoritmo CBA usando f_2**
Peso della variabile x1	1	1	0,870	0,660
Peso della variabile x2	0,041	0,134	0,894	0,709
Peso della variabile x3	0,078	0,346	0,569	1
Peso della variabile x4	0,592	0,519	1	0,196
Peso della variabile x5	0,189	0,220	0,280	0,726
Peso della variabile x6	0,265	0,076	0,096	0,141
N. Cluster	10	20	–	–
N. cluster patologia 1	–	–	20	18
N. cluster patologia 2	–	–	6	10
N. cluster patologia 3	–	–	2	7
N. cluster patologia 4	–	–	4	11
N. cluster patologia 5	–	–	2	15
N. cluster patologia 6	–	–	2	10

Di seguito il confronto tra il codice genetico ottenuto sul DB nazionale (Nat) con l'algoritmo NCBA corrispondente al miglior individuo di tutte le elaborazioni confrontato con quello ottenuto sul DB regionale (Reg) della Lombardia utilizzando sempre l'algoritmo NCBA con fitness f_2. I valori mostrati dei pesi delle variabili sono in forma normalizzata. Il numero massimo di cluster considerato è stato pari a 50. Nella prima riga sono indicate le patologie con la codifica

Tabella 55: Confronto tra codici genetici per NCBA con f_2 su base dati nazionale e regionale

	Patologia											
	1		**2**		**3**		**4**		**5**		**6**	
	NCBA											
	Nat	**Reg**	**Nat**	**Reg**	**Nat**	**Reg**	**Nat**	**Reg**	**Nat**	**Reg**	**Nat**	**Reg**
Peso var. x1	0,585	1	0,395	0,039	0,067	0,522	1	1	0,454	0,906	0,277	0,795
Peso var. x2	0,567	0,352	0,547	1	0,750	1	0,892	0,455	0,054	0,961	1	1
Peso var. x3	0,821	0,514	0,013	0,039	0,171	0,236	0,538	0,269	0,723	1	0,522	0,403
Peso var. x4	1	0,450	0,920	0,237	0,961	0,856	0,274	0,147	0,851	0,880	0,182	0,859
Peso var. x5	0,276	0,237	0,724	0,546	1	0,970	0,098	0,911	1	0,695	0,776	0,694
Peso var. x6	0,123	0,118	1	0,377	0,114	0,544	0,080	0,107	0,626	0,970	0,295	0,124
N. cluster malati	47	48	12	10	5	34	37	47	31	14	15	31
N. cluster sani	30	44	18	27	3	29	43	41	17	13	36	41

5. Il software

5.1 Ambiente di sviluppo

Per la realizzazione di tutte le fasi dello sviluppo del software, analisi, progettazione e scrittura del codice, sono stati utilizzati strumenti open source. Per l'analisi e la progettazione è stato utilizzato StarUML, per lo sviluppo MS Visual C++ 2008 Express Edition su sistema operativo Windows 7.

Sono state utilizzate le librerie SPARE sviluppate presso il POMOS, le librerie BOOST, librerie windows per la grafica, le librerice C++ STD e Windows. L'utilizzo delle librerie SPARE comporta l'uso delle librerie Boost di cui è stata utilizzata la versione precompilata per Windows.

5.2 Gli strumenti di lavoro e le tecniche di realizzazione

Sul paradigma dell'Object-Oriented (OO), è stato basato lo sviluppo di analisi, progettazione e programmazione dello strumento software realizzato per l'elaborazione dei dati inerenti le malattie professionali. Il paradigma dell'OO avendo una visione del problema più vicina alla realtà ha consentito una maggiore concentrazione sulle problematiche di interesse di questo studio, e meno su quelle meramente tecniche di cui comunque è stato necessario affrontare ed occuparsene. I risultati ottenuti possono essere considerati al quanto interessanti.

Nell'ambito del paradigma OO, per l'Object-Oriented Analysis and Design (OOAD) è stato utilizzato il linguaggio di modellazione e specifica UML (Unified Modeling Language); mentre per l'Object-Oriented Programming, per la codifica del software, è stato utilizzato il linguaggio C++ scelto anche nell'ottica di riuso delle librerie SPARE (Something for PAttern REcognition) sviluppate presso il POlo di MObilità Sostenibile (POMOS) dell'Università "Sapienza" di Roma. Le librerie SPARE sono un insieme di classi C++ (principalmente template) utili per la creazione di software rivolto alla soluzione di problemi di Soft Computing e Pattern Recognition come: Classificazione, Clustering, Approssimazione Funzionale e Predizione implementanti alcune classiche routine di machine learning come algoritmi di clustering e algoritmi genetici in stile generico ed altamente flessibile. Le classi contenute nelle SPARE sono state riadattate e personalizzate per gli scopi del presente lavoro e combinate con le nuove classi implementate, abilitando nuovi scenari operativi.

Oltre alle SPARE, sono state utilizzate le librerie Boost (versione 1.47), utilizzate anche per le SPARE, le librerie standard STL e le Visual C++ libraries. Lo sviluppo è avvenuto su sistema operativo Windows 7 (ampiamente utilizzato in svariati ambienti di lavoro e di studio) attraverso tecnologia .NET utilizzando un ambiente di sviluppo open source quale Visual C++ 2008 Express Edition. La connessione al DB è stata realizzata mediante Open DataBase Connectivity (ODBC), un driver che attraverso un'API standard (Application Programming Interface) implementa la connessione dal client al DBMS; l'ODBC è indipendente dai linguaggi di programmazione, dai sistemi di database e dal sistema operativo.

L'ambiente di sviluppo Visual C++ Express Edition supporta il lavoro per lo sviluppo di software dalla creazione alla gestione dei progetti attraverso la scrittura di codice ed esecuzione del debug e contiene strumenti per la compilazione. Visual C++ supporta direttamente la compilazione per computer x86 e include i compilatori basati su x64. Ottimizza le prestazioni per tutte le piattaforme.

Le librerie utilizzabili in questo ambiente di sviluppo, comprendono:

- libreria runtime C (CRT), che include soluzioni avanzate di sicurezza alternative a quelle funzioni note per porre problemi di sicurezza;

- libreria standard di C++, contenente la libreria iostream e la Standard Template Library (STL);

- Active Template Library (ATL), librerie per la creazione di componenti e applicazioni COM;

- librerie Microsoft Foundation Class (MFC), per la creazione di applicazioni desktop con interfacce tradizionali o in stile Office;

- librerie PPL (Parallel Patterns Library), per gli algoritmi paralleli e asincroni che vengono eseguiti dalla CPU;

- librerie C++ AMP (C++ Accelerated Massive Parallelism), per gli algoritmi paralleli eseguiti dalla GPU in modalità massiccia;

- libreria di modelli di Windows Runtime C++ (WRL), per lo sviluppo di tipo COM;

- librerie di classi di .NET Framework (utilizzate tramite C++/CLI), STL/CLR e la libreria di supporto C++.

Inoltre, attraverso l'utilizzo di Visual C++ è possibile accedere alle API Windows sia per le applicazioni di Windows Store che per le applicazioni desktop.

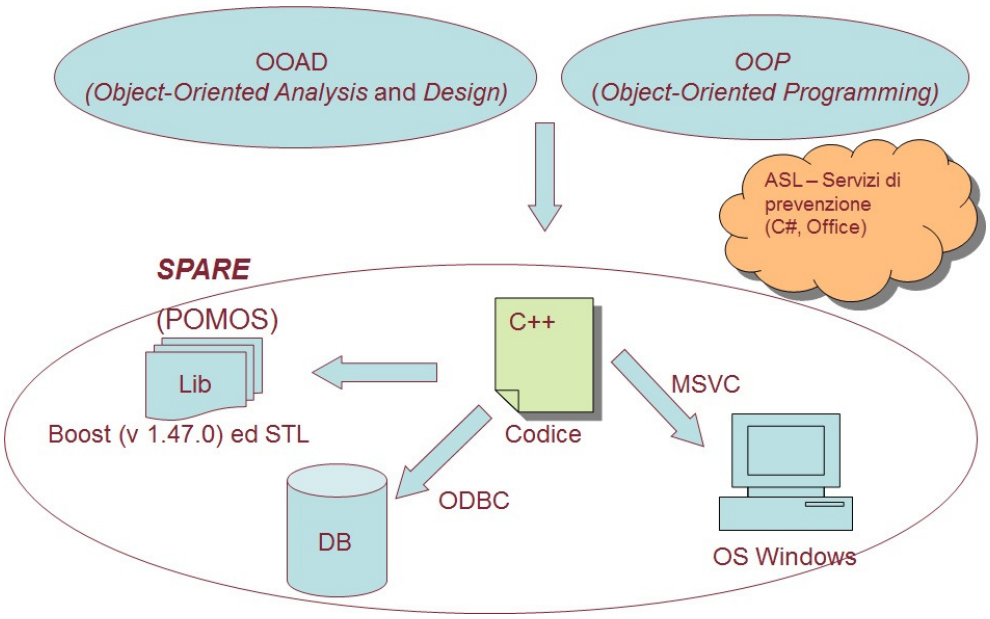

Figura 14: Analisi e sviluppo del software

Per quanto riguarda la modellazione in UML è stato utilizzato il prodotto open source denominato StarUML. Essa, è un'applicazione completa per la realizzazione di modelli e progetti di sviluppo software, permette di creare diagrammi sintetici e analitici per descrivere i blocchi di codice necessari allo sviluppo di un programma.

Una caratteristica interessante di questa applicazione, è la possibilità di generare codice in automatico partendo dai grafici, come anche di eseguire reverse engineering partendo dal codice.

Questa opzione è compatibile con i linguaggi di programmazione C++, C# e Java. Inoltre è possibile generare documentazione in maniera quasi automatica utilizzando degli appositi template.

Tali strumenti di lavoro sono stati utilizzati per la realizzazione dell'idea proposta in questo lavoro che è basata sulla combinazione di Cluster Analysis e di un Algoritmo Genetico. Per tale realizzazione sono state utilizzate le tecniche inerenti l'area dell'apprendimento automatico ed in particolare dei metodi di *apprendimento non supervisionato e supervisionato*, del *cluster analysis*, e degli *algoritmi genetici.* Sono state utilizzate sia le tecniche di apprendimento non supervisionato, che lavorano confrontando i dati e ricercando similarità o differenze, sia le tecniche di apprendimento supervisionato. Come algoritmo di clustering è stato utilizzato il *K-means*, che consente di suddividere gruppi di oggetti in partizioni di *K cluster* sulla base dei loro attributi; e per il quale sono state definite opportune misure di dissimilarità parametriche. Nell'implementazione del *K-means*, sono state percorse varie fasi di sviluppo a crescente complessità: semplici cicli di clustering indipendenti dai pesi delle variabili, diverse definizioni di distanze, analisi di criteri di stop, algoritmi di inizializzazione dei centroidi, implementazioni degli indici di validazione per la valutazione della compattezza e separazione dei clusters. Per quanto riguarda l'algoritmo genetico sono state utilizzate le più note tecniche di realizzazione presenti in letteratura.

5.3 Architettura software

Sulla base dello sviluppo software con utilizzo del paradigma ad oggetti, il software è stato implementato con il modello three-layer: livello interfaccia utente, livello di business e livello dati.

Ciascun livello contiene le proprie classi necessarie alla realizzazione del proprio scopo, infatti il livello di interfaccia utente contiene tutte le classi per la creazione di maschere per l'interazione con l'utente e tramite le quali è possibile attivare le principali funzioni software; il livello di business contiene tutte le classi di calcolo ed il livello dei dati contiene tutte le classi necessarie all'interfacciamento con la base dati per la loro lettura e scrittura.

Di seguito vengono descritte brevemente le principali classi utilizzate. Solo per le classi *RealCode* e *Genetic* viene data una descrizione leggermente più ampia rispetto alle altre classi.

Le Classi di INTERFACCIA

`FormOpenApp.h`

Maschera iniziale visualizzata all'avvio dell'applicazione.

`FormOutputGroup.h`

Maschera di visualizzazione dei dati di un cluster.

Le Classi di BUSINESS

`FormOpenApp.h`

Contiene la logica ed il controllo degli eventi da interfaccia controllando il flusso delle chiamate delle varie funzioni.

`LWD.hpp`

Classe per il calcolo della dissimilarità - Linearly Weighted Distance

`RIV.h`

RIV (Relative Index of Validation) classe adibita al calcolo dell'indice di validazione, per il calcolo della partizione ottimale. Vengono implementati gli indici di validazione: indice di Davies-Bouldin; indice di Calinski-Harabasz; indice I (Maulik-Bandyopadhyay)

`Utility.h`

Contiene la logica di utility necessarie all'esecuzione del codice.

`Analyzer.h`

Classe valutatrice per l'algoritmo genetico con indici di valutazione viene passata alla classe RealCode.hpp

`AnalyzerClassifier.h`

Classe valutatrice per l'algoritmo genetico per la realizzazione dei classificatori per gli algoritmi BA, CBA ed NCBA il classificatore , viene passata alla classe RealCode.hpp

`Kmeans.hpp e MinSod.hpp`

Classi SPARE utilizzate per il clustering

`RealCode.hpp`

E' la classe template che implementa il concetto di ambiente principalmente per l'algoritmo genetico. In tale ambiente il codice è un vettore di reali che implementa la descrizione fatta nei paragrafi precedenti. Brevemente seguono i principali attributi e funzioni.

Attributi

`RealParam` *`mUniformRate`*`:`

Probabilità di crossover in distribuzione uniforme. Default *33% (0,33)*.

`RealParam` *`mTwoPointRate`*`:`

Probabilità di crossover two point. Default *50% (0,5)*.

Funzioni

`Public:`

`CodeType` *`Breed()`*

Crea il secondo elemento di un individuo in modo casuale. Il secondo elemento è un vettore di reali (o di interi) della dimensione fissata al momento della creazione dell'oggetto *Genetic* (è il primo parametro del suo costruttore).

(La *Breed()* genererà il vettore dei pesi)

Private:

`void UniformCrossover(CodeType& rCodeA, CodeType& rCodeB) const;`

Esegue il crossover di due vettori scambiando gli elementi di una stessa posizione scelta in modo casuale; potrebbero essere scambiati tutti gli elementi, nessuno o solo alcuni. La modalità casuale è eseguita con una funzione *random* che rilascia *true* o *false* (metodo del lancio della monetina). Per ogni elemento del vettore, viene eseguita tale funzione, se il risultato è *true* l'elemento viene scambiato altrimenti no.

`Genetic.hpp`

E' la classe template che implementa l'algoritmo genetico.

Parametri di template

`typename Environment:`

Deve essere il tipo contenente le funzioni tipiche per l'evoluzione genetica: *riproduzione, crossover* e *mutazione*. Viene utilizzata la classe *RealCode.*

Attributi

Protected:

`Environment mEnvAgent:`

E' il contenitore dell'ambiente creato esternamente ed assegnato esternamente. E' dello stesso tipo passato come parametro del template di classe. (*RealCode*).

`NaturalParam mPopSize:`

Dimensione della popolazione è un *intero senza segno*.

`Population mPopBuffA, mPopBuffB:`

Buffer contenenti la popolazione degli individui. Viene utilizzata *mPopBuffA* per la popolazione iniziale.

`RealParam mElitism:`

Contiene il valore della percentuale di elementi da preservare della vecchia popolazione in fase di generazione della nuova popolazione. Il valore di default è pari a *10% (0,1).*

`RealParam mCrossRate:`

Contiene il valore della probabilità di effetturare il *crossover*. Il valore di default è pari a *80% (0,8).*

`RealParam mMuteRate`:

Contiene il valore della probabilità di effetturare un'operazione di *mutazione*. Il valore di default è pari a *30% (0,3)*.

Tipi di dati

`typedef std::set<Individual> Population:`

E' la raccolta degli elementi che costituiscono la popolazione.

`typedef std::pair<RealType, CodeType> Individual`:

Rappresenta un individuo della popolazione costituito da una coppia di elementi dove il *primo* rappresenta il suo valore di fitness ed è un valore di tipo real; il *secondo* è un vettore (*std::vector<GeneType> CodeType* definito nella classe *ambiente*) rappresentante la costituzione dell'individuo, ossia contiene in ogni elemento un valore (discreto o reale) inizialmente assegnato in modo casuale all'interno di un intervallo prefissato di valori (all'atto della creazione della classe Genetic)

Funzioni

`void Initialize()`

Funzione di inizializzazione di creazione della popolazione di individui dove ogni individuo è una coppia <fitness, elemento> (<*RealType, CodeType*>). Gli elementi vengono generati in modo casuale e viene calcolata la relativa funzione di fitness. Le coppie vengono inserite in un apposito buffer (*mPopBuffA*).

(Nel nostro caso ogni elemento è il vettore dei pesi e la funzione di fitness sarà la funzione che calcola l'indice che per essere calcolato deve eseguire *kmax* clustering)

`void StepUp()`

Tale funzione viene utilizzata per realizzare un passo di evoluzione di una generazione della popolazione attuale e chiama la funzione *Genetic::NextGeneration(vecchiaPopolazione, nuovaPopolazione)*

`void NextGeneration(Population& OldPop, Population& NewPop)`

Sulla base della vecchia popolazione genera la nuova popolazione eseguendo i seguenti passi:

- Viene preservata una percentuale della vecchia popolazione in base al valore prefissato di elitismo *mElitism* fissata al 10%.
- Viene eseguita la funzione *RouletteWheelSelection(OldPop)* che seleziona due nuovi individui ossia esegue la *"Riproduzione"*. I due nuovi individui sono i genitori utilizzati nella fase successiva.
- In base al valore di *mCrossRate* (probabilità di eseguire il crossover) viene eseguita la funzione di crossover *RealCode::Crossover(genitore1, genitore2)*.

- In base al valore del *muteRate* (probabilità di eseguire una mutazione su un individuo) viene eseguita la funzione *RealCode::Mutate(genitore1)* (analogamente con *genitore2*)
- Calcolo delle fitness corrispondenti ai due nuovi individui.
- Vengono inseriti i nuovi individui nel buffer *NewPop* (nuova popolazione)
- Viene svuotato il buffer *OldPop*.

Le Classi di DATA

`DB.h`

Realizza la connessione fisica al database

`QueryClass.h`

Realizza le query di lettura/scrittura sul DB

`Lavoratore.h`

E' una classe trasversale ai tre layer in quanto classe *entità*. Contiene i dati di un record di lavoratore del database e quindi le variabili di elaborazione.

Di seguito viene mostrato il principale diagramma delle classi che realizzano l'algoritmo genetico e la cluster analysis.

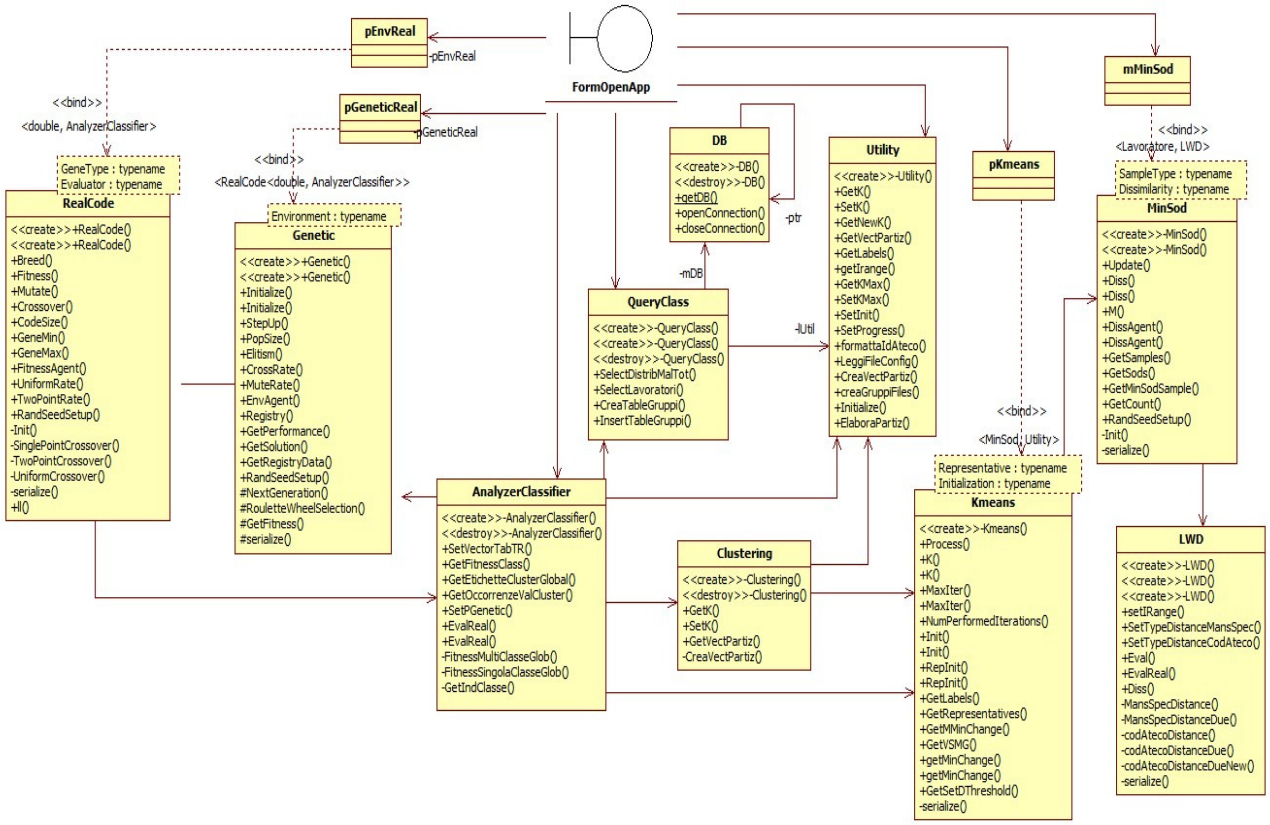

Figura 15: Diagramma UML delle classi per la realizzazione dell'AG e cluster analysis

5.4 L'interfaccia grafica del prototipo

Di seguito l'interfaccia utente del prototipo sviluppato. Sono visibili anche tutte le sezioni utilizzate per i test di analisi iniziale dei dati, per il loro preprocessamento e quindi per la ricerca delle features utilizzate con l'esecuzione dell'AG nella fase successiva.

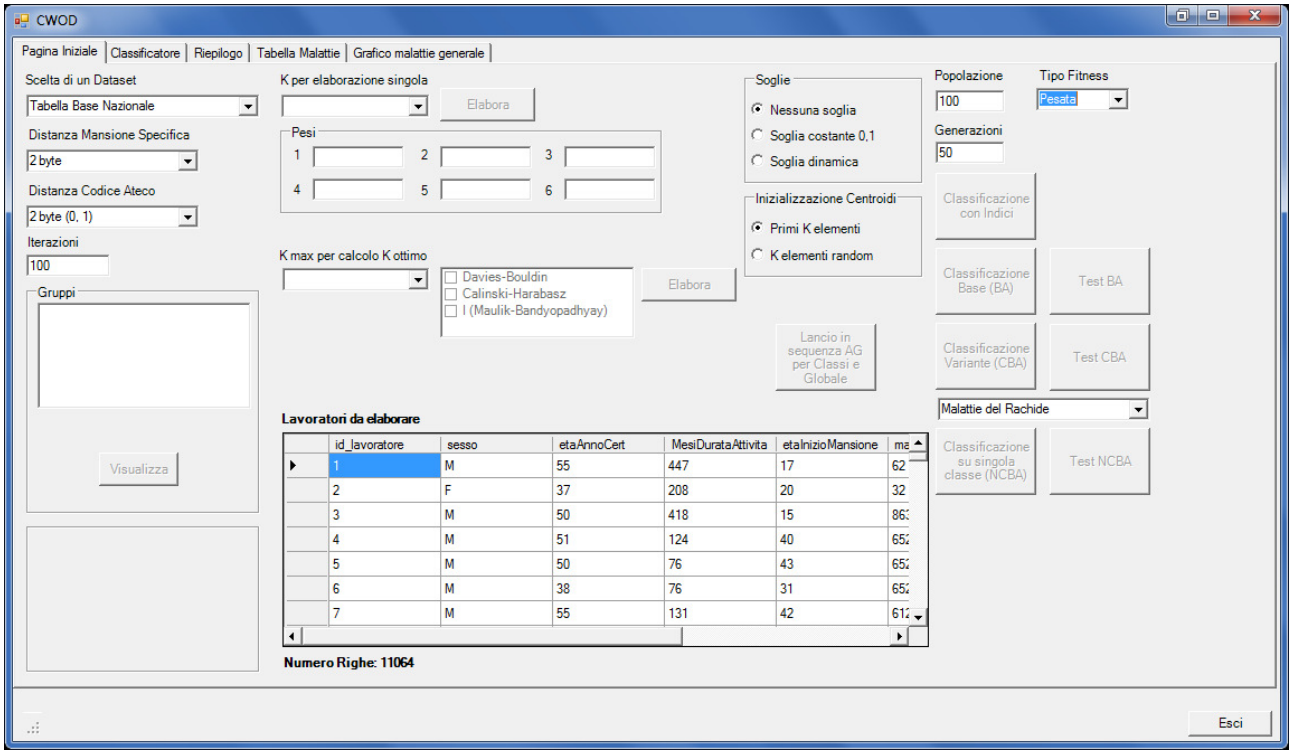

Figura 16: Interfaccia grafica del prototipo

L'interfaccia grafica del prototipo consente l'esecuzione dei test sui diversi dataset consentendo di scegliere un particolare algoritmo di classificazione.

L'interfaccia grafica del software consente l'esecuzione dei classificatori in parallelo o la singola esecuzione su una determinata malattia scelta a priori.

6. Conclusioni

Sono stati esaminati tre differenti algoritmi di classificazione e per uno di questi sono stati eseguiti dei test su due basi di dati, omologhe per struttura dei dati ma differenti nella copertura del territorio: una base dati con copertura della sola regione Lombardia e l'altra con copertura di tutto il territorio nazionale. Inoltre è stato possibile confrontare i risultati ottenuti applicando l'SVM al database regionale con gli algoritmi BA e CBA, sia con fitness f_1 che con fitness f_2.

I test eseguiti sul DB regionale mostrano che l'algoritmo base (BA) fornisce le migliori risposte in corrispondenza della patologia più frequente ("sordità"); con l'utilizzo della funzione di fitness f_2 si sono ottenute predizioni sull'intero set di patologie, mentre con la funzione di fitness f_1 non è stato possibile ottenere predizioni per le "malattie muscoloscheletriche".

Gli esperimenti condotti sempre sul DB regionale con l'algoritmo CBA, utilizzando sia f_1 che f_2, mostrano valori di prestazione simili rispetto all'algoritmo Base BA, ed un miglioramento dei tempi di esecuzione che si sono ridotti a un terzo rispetto a quelli della versione BA, ciò grazie al clustering eseguito su insiemi di dati di dimensioni ridotte avendo partizionato il DB in gruppi selezionati per patologia.

Osservando i valori predittivi negativi (v. Tabella 34) particolarmente elevati ottenuti con l'algoritmo CBA, è possibile considerare che una previsione negativa per un determinato lavoratore è sufficiente ad accertare con alta attendibilità il suo stato non patologico, e quindi si può pensare di usare il sistema all'interno di una procedura di screening automatica, progettata per ridurre i costi di esecuzione delle visite mediche su tutti i lavoratori interessati.

I pesi delle caratteristiche (v. Tabella 54) ottenuti nei test sia sul DB regionale che nazionale, mostrano un'ampia variabilità tra i differenti algoritmi. In tutti gli esperimenti, solo la 6^ variabile (attività economica) sembra essere meno importante delle altre, e in un ulteriore lavoro potrebbe essere sostituita con qualche altra caratteristica non considerata in questo studio.

I test eseguiti sul DB nazionale attraverso l'algoritmo NCBA hanno prodotto risultati peggiori rispetto a quelli ottenuti con lo stesso algoritmo sul DB regionale. La causa andrebbe ricercata attraverso ulteriori studi analizzando le variazioni introdotte attraverso una base dati la cui differenza con la precedente è l'ampliamento del territorio. Proprio questa nuova feature del territorio potrebbe sostituire l' "attività economica", oppure si potrebbe semplicemente aggiungerla a quelle già esistenti.

Dai risultati dei test eseguiti con l'SVM si evince non esserci nessun miglioramento sostanziale.

Si nota che per 2 patologie si ottengono risultati migliori con l'algoritmo SVM e per 4 patologie invece si hanno risultati migliori con il nuovo software. Le differenze sono comunque abbastanza contenute. Inoltre occorre tener conto che per quanto riguarda i nuovi algoritmi di classificazione sviluppati nell'ambito del presente lavoro sono riportati i risultati del test più favorevole, mentre per l'algoritmo SVM, è riportata la media su 5 test, per cui nel caso del test migliore i risultati potrebbero essere più incoraggianti.

L'utilizzo dei 4 indicatori (sensibilità, specificità, valore predittivo esito negativo e valore predittivo esito positivo) al posto dell'accuratezza (% di risposte corrette) è conseguenza del fatto che la sola indicazione dell'accuratezza sarebbe fuorviante poiché le classi non sono bilanciate: si potrebbero avere valori elevati di accuratezza in presenza di alta sensitività e pessima specificità o viceversa (in dipendenza della numerosità della classe in esame), mentre l'obiettivo dovrebbe essere quello di ottenere predittori con valori buoni per tutti i 4 indicatori.

I risultati simili dei due approcci considerati (SVM e classificatori basati sul clustering) fanno pensare che probabilmente, indipendentemente dalla tecnica usata, non si può andare molto oltre con i risultati che si ottengono da questi test, tuttavia è possibile trarre spunti per ulteriori sviluppi.

Il software implementato (prototipo unico) e confrontato con l'SVM ha evidenziato di essere in grado di produrre risultati all'altezza di quelli prodotti con strumenti allo stato dell'arte, come appunto le SVM. Il lavoro presentato in questa tesi è stato finalizzato allo sviluppo ed implementazione di un nuovo strumento utilizzando come SO di base (produzione ed esercizio) quello probabilmente più diffuso nell'ambito della PA (Windows) che possa essere eventualmente utilizzabile come strumento coadiuvante nella prevenzione delle malattie professionali dai preposti a tale attività.

Lo strumento realizzato ha riscontrato un interesse positivo in ambito sanitario per la prevenzione ed il monitoraggio delle malattie lavoro-correlate, ponendosi quale ulteriore mezzo di supporto per la previsione del rischio di contrazione di una data malattia in funzione di alcune caratteristiche del lavoratore e dell'ambiente di lavoro.

Ulteriori sviluppi potrebbero riguardare i seguenti punti:

- Arricchimento del data base con nuove caratteristiche.
- Applicazione di tecniche di "Ensemble" ai modelli di classificazione sintetizzati per riconoscere le singole malattie.

Bibliografia

[1] A. K. Jane, R. C. Dubes - *"Algorithms for Clustering Data"* - Prentice-Hall. Englewood Cliffs - (1988)

[2] Kumara Sastry, David Goldberg, Graham Kendall - *"Genetic Algorithms"* - In Search Methodologies. Springer US - (2005)

[3] M. Riccò - *Bernardino Ramazzini. Vita ed Opere* - Atti del 68° Congresso Nazionale della Società Italiana di Medicina del Lavoro e Igiene Industriale SIMLII - Parma, 5 - 8 ottobre 2005 (2005).

[4] Alexandr A. Savinov - *Mining Possibilistic Set-Valued Rules by Generating Prime Disjunctions*. In PKDD'99, 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases. Vol. 1704 Springer (1999), p. 536-541.

[5] Istat - *Classificazione delle professioni. "Metodi e norme", serie C, n. 12.* - Supplemento all'annuario statistico italiano - Roma, Istituto Poligrafico e Zecca dello Stato (1991).

[6] F. Gallo, P. Scalisi, B. Loré - *La Classificazione delle professioni 2011. "Metodi e norme"* - Roma, Istat (2013).

[7] Istat - *ATECO 91. Classificazione delle attività economiche. Metodi e norme, serie C, n. 11.* - Roma, Istat (1991).

[8] Istat - *ATECO 2002. Classificazione delle attività economiche. "Metodi e norme", n. 18* - Roma, Istat (2003)

[9] G. Campo, M.G. Magliocchi, P. Montanari, A. di Noia, A. Papale - *"Quarto rapporto ISPESL-Regioni sulle malattie professionali - MALPROF 2005-2006"* - ISPESL, Dipartimento Processi Organizzativi, Roma - supplemento alla rivista "Prevenzione Oggi" - (giugno 2009).

[10] G. Campo, M.G. Magliocchi, P. Montanari, A. di Noia, A. Papale - *"Quinto rapporto INAIL-Regioni sulle malattie professionali - MALPROF 2007-2008"* - INAIL, Dipartimento Processi Organizzativi - ex ISPESL, Roma - (2011).

[11] P. Montanari, A. di Noia, A. Silvetti, S. Mineo - "*MaProWeb (Versione 1.1). Applicativo web based per l'inserimento delle segnalazioni di malattie professionali raccolte dai servizi di prevenzione delle ASL. - Manuale utente"* - ISPESL, Dipartimento Processi Organizzativi, Roma - (luglio 2008)

[12] Chinmoy Mukherjee, Komal Gupta, Rajarathnam Nallusamy - *"A Decision Support System for Employee Healthcare"* - In Third International Conference on Services in Emerging Markets (2012)

[13] Razan Paul, Abu Sayed Md. Latiful Hoque - *"Clustering Medical Data to Predict the Likelihood of Diseases"* - In Fifth International Conference - IEEE - Digital Information Management (ICDIM) - (2010)

[14] Zhaohui Huang Daoheng Yu Jianye Zhao - *"Application of Neural Networks with Linear and Nonlinear Weights in Occupational Disease Incidence forecast"*. In Circuits and systems. - IEEE APCCAS 2000 - (2000)

[15] L. Fabbris – "*Statistica multivariata. Analisi esplorativa dei dati"* - McGraw-Hill, Milano - (1997)

[16] M. Gherghi, C. Lauro – *"Appunti di analisi dei dati multidimensionali. Metodologie ed esempi"* – RCE Edizioni, Napoli - (2004)

[17] G. Campo, M.G. Magliocchi, G. Capozza, A. Papale, P. Montanari, A. di Noia - *"Il sistema di sorveglianza MalProf per le patologie correlate al lavoro"* - In Giornale Italiano di Medicina del Lavoro ed Ergonomia (GIMLE), Vol. XXX, Supplemento 2 al N. 3 – (Luglio-Settembre 2008) (pp 105-106).

[18] Tilmann Steinberg, James C. Ford, Yuhang Wang, Fillia S. Makedon – *"Similarity Searches in Heterogeneous Feature Spaces"* - Department of Computer Science Dartmouth College Hanover, NH 03755 USA – (2003)

[19] MacQueen J.B. – *"Some Methods for Classification Analysis of Multivariate Observations"* - Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability - Berkeley, University of California Press – (1967), (pp 281-297)

[20] Roiger R.J., Geatz M.W. - *"Introduzione al Data Mining"* - McGraw-Hill – (2004)

[21] Russell S.J., Norvig P. – *"Intelligenza Artificiale. Un approccio moderno"* - Voll. 1 e 2, Pearson Education Italia – (2005)

[22] Tan P., Steinbach M., Kumar V. – *"Introduction to Data Mining"* - Pearson Addison Wesley – (2005)

[23] G. Del Vescovo, A. Rizzi – *"Algoritmi simbolici per la classificazione automatica di dati strutturati"* – Tesi di dottorato ciclo XX - Università degli Studi di Roma "La Sapienza", Facoltà di Ingegneria, Dipartimento Infocom – (2008)

[24] Grady Booch – *"Object oriented analysis and design with applications"* Second Edition – Addison Wesley – (1994)

[25] A. Rizzi - *"Introduzione al modellamento data driven"* - Dispense del corso "Circuiti ed Algoritmi per il Riconoscimento" alias "Riconoscimento e Classificazione" - Università degli Studi di Roma "La Sapienza", Facoltà di Ingegneria, Dipartimento Infocom – (2008)

[26] John Holland - *"Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence"* - Bradford Books - (1992)

[27] Kathryn A. Dowsland - *"Genetic Algorithms-A Tool for OR?"* - The Journal of the Operational Research Society, Vol. 47, No. 4. (Apr., 1996), (pp. 550-561)

[28] J.R. Koza. - *"Genetic Programming: On the Programming of Computers by Means of Natural Selection"* - MIT Press - (1992)

[29] Melanie Mitchell - *"An Introduction to Genetic Algorithms"* - MIT Press - (1996)

[30] B. Lazzerini - *"Introduzione agli Algoritmi Genetici"* - Dipartimento di Ingegneria della Informazione - Università di Pisa

[31] D.B. Fogel – "*Evolutionary computation: toward a new philosophy of machine intelligence"* - IEEE Press, NewYork – (1995)

[32] Nicholas Freitag McPhee, Riccardo Poli, William B. Langdon – "*Field Guide to Genetic Programming"* - University of Minnesota Morris Digital Well, Computer Science Faculty – (2008)

[33] Alden H. Wright - *"Genetic Algorithms for Real Parameter Optimization"* - CiteSeerX, College of Information Sciences and Technology, Pennsylvania State University - (1991)

[34] Alfredo Rizzi - *"Inferenza Statistica"* - UTET, Torino - (1992)

Sitografia

[1] Centro nazionale per la prevenzione ed il controllo delle malattie

http://www.ccm-network.it/home.html

[2] Società Italiana di Medicina del Lavoro ed Igiene Industriale

http://www.simlii.org

[3] Wikipedia the free encyclopedia

https://it.wikipedia.org/wiki/

[4] Istat - Istituto Nazionale di Statistica – Sistema delle classificazioni

http://www.istat.it/it/strumenti/definizioni-e-classificazioni/

[5] Scientific literature digital library and search engine for literature in computer and information science

http://citeseerx.ist.psu.edu/index

[6] Portale europeo per la ricerca e l'accesso alle tesi di dottorato e di laurea in formato elettroniche pubblicate in Europa

http://www.dart-europe.eu/basic-search.php

[7] Portale Area Ricerca INAIL (ex ISPESL)

http://www.inail.it/internet/default/INAILcosafa/Ricerca/index.html

[8] International Labour Organization

http://www.ilo.org/

[9] Institute of Electrical and Electronics Engineers

www.ieee.org

[10] National Center for Biotechnology Information. Fornisce l'accesso alle informazioni di tipo biomedico e genomico riguardo ai progressi scientifici sulla salute.

http://www.ncbi.nlm.nih.gov/

[11] World Health Organization

http://www.who.int/en/

[12] Office of Scientific and Technical Information of U.S. Department of Energy

http://www.osti.gov

[13] CORDIS - Servizio Comunitario di Informazione in materia di Ricerca e Sviluppo

http://cordis.europa.eu/home_it.html

[14] Occupational Health and Safety - American Magazine.

https://ohsonline.com/Home.aspx

[15] Occupational Safety & Health Administration - Agenzia Europea per la sicurezza e la salute sul lavoro

https://osha.europa.eu/it