Università Politecnica delle Marche

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

Tesi di Laurea

Applicazione di tecniche di Intelligenza Artificiale per l'estrazione di conoscenza a partire dai dati di monitoraggio di pazienti affetti da Covid-19

Application of Artificial Intelligence techniques for knowledge extraction from monitoring data of patients with Covid-19

Relatore

Prof. Domenico Ursino

Candidata

Sara Pantalone

Anno Accademico 2020-2021

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Nel 2020, l'Italia, così come tutto il mondo, è stata investita dal nuovo coronavirus Sars-CoV-2. Quest'ultimo ha portato alla dichiarazione di un'emergenza sanitaria per Covid-19, a cui è stata data risposta immediata con una serie di misure urgenti fin dalla dichiarazione dello stato di emergenza del 31 Gennaio 2020.

Gli ospedali di tutta Italia, sin dall'inizio della diffusione della pandemia, si sono trovati in grave difficoltà, non avendo mezzi e fondi per poter affrontare, in modo ottimale, la situazione.

L'avvento della pandemia da un un lato, l'innovazione tecnologica e l'aumento dell'età media dall'altro, hanno spinto sempre di più verso il miglioramento dell'assistenza sanitaria e hanno favorito l'adozione della telemedicina.

Infatti, secondo i dati dell'Osservatorio Innovazione Digitale in Sanità della School of Management del Politecnico di Milano, il 47% degli specialisti ricorre attualmente al teleconsulto, percentuale che arrivava appena al 21% prima della pandemia.

Per telemedicina si intende una modalità di erogazione di servizi di assistenza sanitaria, tramite il ricorso a tecnologie innovative, in particolare alle Information and Communication Technologies (ICT), in situazioni in cui il professionista della salute e il paziente si trovano a distanza. La telemedicina comporta la trasmissione sicura di informazioni e dati di carattere medico nella forma di testi, suoni, immagini o altre forme necessarie per la prevenzione, la diagnosi, il trattamento e il successivo controllo dei pazienti.

I servizi di telemedicina vanno assimilati a qualunque servizio sanitario diagnostico/ terapeutico. Tuttavia la prestazione in telemedicina non sostituisce la prestazione sanitaria tradizionale nel rapporto personale medico-paziente, ma la integra per migliorare efficacia, efficienza e appropriatezza.

In questo scenario di emergenza sanitaria, in cui la telemedicina è protagonista del progresso, è stato organizzato ad Offagna il progetto RicovAi-19, ovvero RICOVero ospedaliero con strumenti di "Artificial Intelligence" nei pazienti con Covid-19. Lo scopo principale è quello di alleggerire il più possibile l'ospedalizzazione, in modo tale da evitare la saturazione dei reparti Covid e non, sfruttando, appunto, la telemedicina.

Il progetto è frutto di una stretta collaborazione tra Almawave, Ospedali Riuniti di Ancona, Università Politecnica delle Marche, Asur Marche e le società Vivisol e

Aditech.

Si tratta di uno studio pilota di fattibilità (non-farmacologico interventistico) e di una sperimentazione clinica in cui l'Intelligenza Artificiale supporta concretamente, a più livelli, pazienti, medici e ospedali, nella complessa sfida del contrasto alla pandemia da Covid-19 e in prospettiva applicabile alla diagnosi e prognosi di altre e diverse patologie.

Il ruolo dell'università, inizialmente solo di supporto alla creazione della piattaforma e alla gestione dei dati, è stato principalmente di ricerca. I dati raccolti durante l'intera durata del progetto, sono stati condivisi in forma anonima e su di essi sono state applicate tecniche di Intelligenza Artificiale per poter estrarre informazioni utili riguardanti la malattia da Covid-19.

La prima fase progettuale, in cui i pazienti venivano reclutati e il database veniva riempito di informazioni, ha riguardato principalmente le attività di ETL, ovvero di pulizia e di manipolazione dei dati. Durante questa fase si sono riscontrati diversi problemi nell'avere a che fare con dati biomedici reali; la raccolta di dati da parte di medici e pazienti è risultata essere spesso inconsistente o incompleta. Di conseguenza, grazie alla collaborazione con tutte le parti impegnate nel progetto, sono state messe a punto soluzioni puntuali.

Una volta ottenute tutte le informazioni, e una volta manipolati e corretti i dati, il database era composto da numerosi dettagli importanti per conoscere l'evoluzione della sintomatologia da Covid-19.

Avere un dataset ordinato e ben assortito contenente le informazioni di ogni paziente prima, durante e dopo l'insorgenza dei sintomi, ha aperto a differenti idee da cui estrarre conoscenze. Di conseguenza, grazie alla collaborazione di tutto il team, sono state vagliate diverse ipotesi per poter portare avanti studi interessanti e di ricerca.

Dopo aver considerato diverse proposte, in accordo con i medici coinvolti, è stato deciso di creare diversi gruppi di pazienti sulla base di parametri invariati nel tempo che caratterizzano il paziente prima di ammalarsi di Covid-19, per poi soffermarsi sul decorso della malattia di ogni gruppo.

In questo modo, un nuovo ipotetico paziente, una volta esaminati i parametri coinvolti, potrà essere assegnato ad un determinato gruppo e si potrà presumibilmente prevedere come il virus influirà sugli altri parametri presenti nel dataset in modo tale da porre più attenzione al loro monitoraggio.

Nell'elaborato verranno analizzati ed implementati nello specifico tali aspetti. La struttura della presente tesi è la seguente:

- Nel Capitolo 1 verrà introdotto il progetto RicovAi-19 ponendo l'attenzione sulle esigenze che hanno portato alla sua realizzazione, agli obiettivi e, infine, alla struttura informatica.
- Nel Capitolo 2 si esplorerà il dataset di riferimento chiarendo il meccanismo di reclutamento dei pazienti fino ad approfondire il significato dei dati nel dataset.
- Nel Capitolo 3 si descriverà il processo di estrazione, trasformazione e caricamento dei dati in una prima fase iniziale.
- Nel Capitolo 4 verrà illustrata la seconda fase di ETL soffermandosi principalmente sull'individuazione dei campi vuoti del dataset, sulla logica di assegnazione dei valori ai campi NaN e sulle problematiche ad asse associate.

- Nel Capitolo 5 sarà mostrato il pre-processing dei dati, in particolar modo verrà effettuato un focus sulle variabili di interesse, sul ragionamento che ha portato alla loro scelta, nonchè problematiche e le tecniche di risoluzione ad esse associate
- Nel Capitolo 6 verrà presentato il fulcro del progetto, ovvero l'estrazione di conoscenza attraverso l'implementazione della tecnica di clustering al fine di raggruppare i pazienti illustrando i vari algoritmi studiati e, infine, mostrando i risultati ottenuti dalla tecnica scelta.
- Nel Capitolo 7 saranno illustrati i risultati delle analisi svolte precedentemente; più in particolare verrà approfondito l'evolversi della sintomatologia da un punto di vista globale e basandosi sul raggruppamento effettuato.
- Nel Capitolo 8 saranno tratte le conclusioni, mostrando dei possibili sviluppi futuri.

1

Progetto RicovAi-19

In questo primo capitolo verranno introdotte il progetto RicovAi-19, le esigenze che hanno indotto alla sua realizzazione, gli obiettivi e infine la struttura informatica.

1.1 Emergenza sanitaria e obiettivi del progetto

1.1.1 Il Covid-19

Il 30 gennaio 2020, in seguito alla segnalazione da parte della Cina (31 dicembre 2019) di un cluster di casi di polmonite ad eziologia ignota (poi identificata come un nuovo coronavirus Sars-CoV-2) nella città di Wuhan, l'Organizzazione Mondiale della Sanità (OMS) ha dichiarato emergenza di sanità pubblica di interesse internazionale l'epidemia di coronavirus in Cina.

Il giorno successivo il Governo italiano, dopo i primi provvedimenti cautelativi adottati a partire dal 22 gennaio, tenuto conto del carattere particolarmente diffusivo dell'epidemia, ha proclamato lo stato di emergenza e messo in atto le prime misure di contenimento del contagio sull'intero territorio nazionale.

I sintomi di Covid-19 variano sulla base della gravità della malattia, dall'assenza di sintomi (essere asintomatici) a presentare febbre, tosse, mal di gola, debolezza, affaticamento e dolore muscolare. I casi più gravi possono presentare polmonite, sindrome da distress respiratorio acuto e altre complicazioni, tutte potenzialmente mortali.

Perdita improvvisa dell'olfatto (anosmia) o diminuzione dell'olfatto (iposmia), perdita del gusto (ageusia) o alterazione del gusto (disgeusia) sono stati riconosciuti come sintomi di Covid-19. Altri sintomi meno specifici possono includere cefalea, brividi, mialgia, astenia, vomito e/o diarrea. Gli ultimi dati riguardanti i casi confermati di Covid-19 e il numero di decessi in Italia risalgono al 21 Ottobre 2021 e sono rispettivamente 4.722.188 e 131.655.

Nelle Marche dall'inizio della crisi pandemica sono stati individuati 25.914 casi in provincia di Pesaro-Urbino, 35.656 in provincia di Ancona, 24.367 in quella di Macerata, 11.805 nel Fermano e 12.651 nel Piceno; inoltre, sono 5.127 i casi che si riferiscono a residenti fuori regione.

Gli ospedali di tutta Italia dall'inizio della diffusione della pandemia si sono trovati in grave difficoltà non avendo mezzi e fondi per affrontare in modo ottimale la situazione.

In Figura 1.1 vengono presenti i grafici elaborati dall'Agenzia Nazionale per i Servizi Sanitari Regionali; essi mostrano la percentuale di occupazione di posti letto in terapia intensiva e in area non critica dal 1 Marzo 2020 al 25 Ottobre 2021.

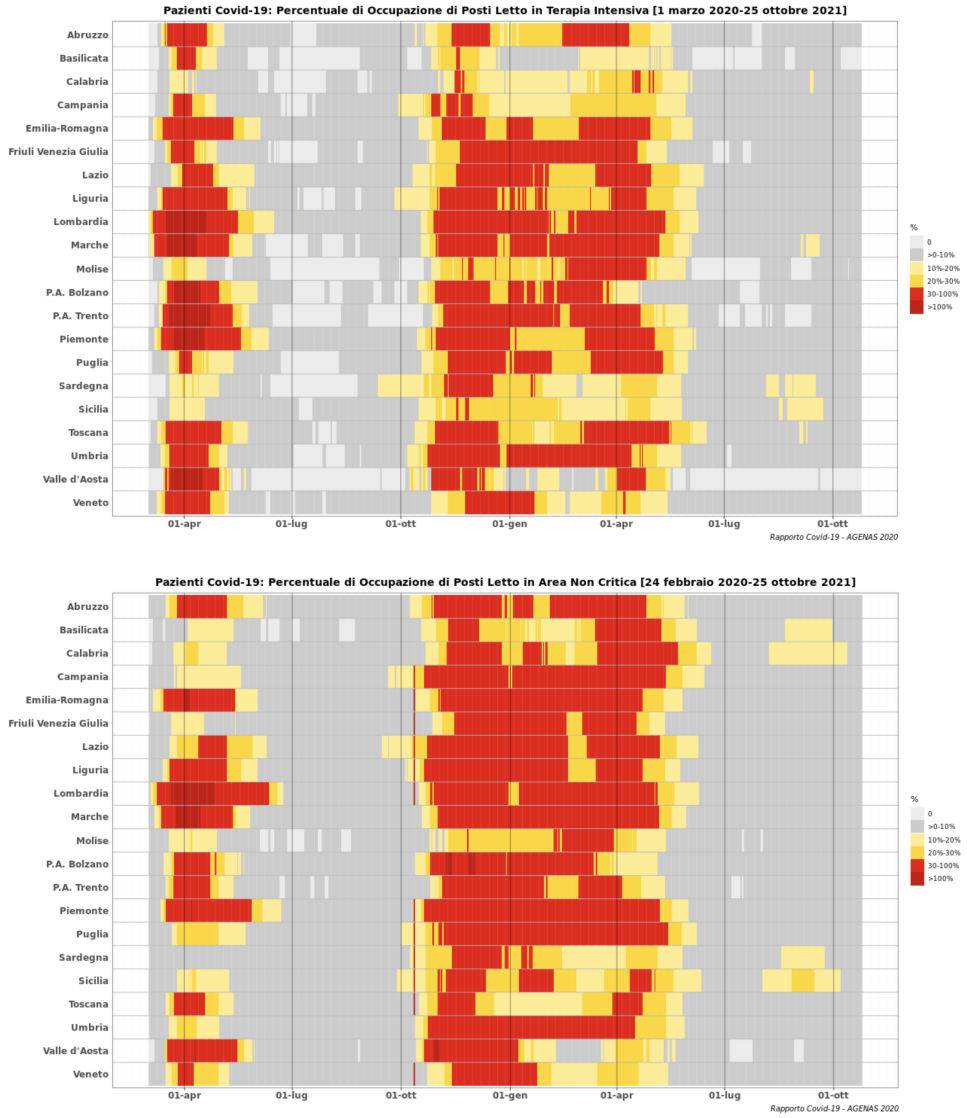


Figura 1.1. Valore percentuale di occupazione di posti letto in terapia intensiva e in area non critica dal 1 Marzo 2020 al 25 Ottobre 2021

Risulta evidente il simile andamento: durante la prima fase di diffusione del

Covid e nei mesi invernali del 2021 gli ospedali di tutto il Paese hanno raggiunto percentuali di occupazione di posti letto tendenti al 100%. Questa situazione, decisamente critica, ha portato alla luce l'esigenza di intervenire sull'organizzazione ospedaliera anche grazie all'ausilio di nuove tecnologie.

1.1.2 Il progetto

In questa situazione di emergenza sanitaria è stato organizzato ad Offagna il progetto RicovAi-19 (il cui logo è riportato in Figura 1.2) ovvero *RICOV*ero ospedaliero con strumenti di "*A*rtificial Intelligence" nei pazienti con Covid-*19* .

Figura 1.2. Logo del Progetto RicovAi-19.

Lo scopo principale, infatti, è quello di alleggerire il più possibile l'ospedalizzazione in modo tale da evitare la saturazione dei reparti Covid e non.

"Tra gli obiettivi dello studio scientifico pilota vi è quello di monitorare l'appropriatezza dell'accesso ospedaliero, perché avvenga solo quando è necessario, oltre, ovviamente, quello di capire quanto l'Intelligenza Artificiale possa essere utile a monitorare lo stato di salute dei pazienti" ha spiegato Marco Mazzanti, Direttore Scientifico di RicovAI-19, già in prima linea nei reparti degli ospedali Barts Heart Centre di Londra e Riuniti di Ancona, ed impegnato su ulteriori fronti di attuazione dell'AI Health con Almawave.

L'Intelligenza Artificiale (AI) scende in campo al servizio della Sanità nella lotta al Covid-19: i cittadini e medici del Comune di Offagna, in provincia di Ancona, sono stati infatti protagonisti di uno studio scientifico e di una sperimentazione clinica pilota per analizzare in tempo reale i parametri clinici del paziente.

Il progetto è frutto di una stretta collaborazione tra Almawave, Ospedali Riuniti di Ancona, Università Politecnica delle Marche, Asur Marche e le società Vivisol e Aditech. Almawave spiega che si tratta di uno studio pilota di fattibilità, non-farmacologico interventistico, e di una sperimentazione clinica in cui l'Intelligenza Artificiale supporta concretamente, a più livelli, pazienti, medici e ospedali nella complessa sfida del contrasto alla pandemia da Covid-19 e in prospettiva applicabile alla diagnosi e prognosi di altre e diverse patologie.

L'Intelligenza Artificiale, infatti, consente di analizzare in tempo reale molteplici parametri clinici dell'utente positivo al virus e di trasmettere i risultati tempestivamente ai medici che, a distanza, effettuano tutte le successive valutazioni, diagnosi e prescrizioni di eventuali cure ed iniziative idonee alla gestione di ciascun caso.

Il tutto avviene grazie a un dispositivo portatile abbinato ad uno smartphone con applicazione dedicata.

Il sistema si avvale di "medical device" CE marked (classe IIA), piattaforma di telemedicina (Classe I) conforme al GDPR e dispositivo di IA non-marcato CE. Ovviamente esso non ha lo scopo principale di valutare l'efficacia di un intervento o di una diagnosi: i medici infatti valutano il paziente agendo sempre secondo le linee guida di buona pratica clinica e incondizionatamente dai risultati del sistema di supporto alle decisioni.

1.2 Parametri e variabili cliniche

Il sistema di AI, dopo addestramento su base dati retrospettiva rilasciato da esso, calcola l'Indicatore di Stabilità Clinica (ICS) per consentire una stima dell'andamento delle condizioni cliniche di ogni paziente. I valori di output del ICS sono classificati secondo le seguenti fasce presenti in Tabella 1.1:

ICS	Classi	Condizione Paziente
0,00 - 2,00	Classe 1	Critico
2,01 - 3,25	Classe 2	Severamente Instabile
3,26 - 4,50	Classe 3	Moderatamente Instabile
4,51 - 7,00	Classe 4	Lievemente Instabile
7,01 - 10,0	Classe 5	Stabile

Tabella 1.1. Classi rispetto all'indice di stabilità clinica

I parametri clinici misurati dal dispositivo in dotazione, fondamentali per il monitoraggio del paziente e il calcolo dell' ICS, sono raggruppati in diverse categorie per una fruizione più chiara e precisa dei dati. Essi sono:

- **Sintomi allarme:**
  - difficoltà a respirare (SOB +/++/+++);
  - saturazione 02 <90%;
  - coscienza alterata;
  - pressione sistolica bassa: $\leq 100 mmHg$;
  - frequenza cardiaca > 100 bpm (0.6) o < 50 bpm.
- **Sintomi maggiori:**
  - febbre >37,5;
  - tosse +/++/+++;
  - sintomi minori;
  - stanchezza;
  - mal di gola;
  - mal di testa;
  - dolori muscolari;
  - congestione nasale;
  - perdita totale capacità percezione odori (Anosmia);
  - distorsione o indebolimento del senso del gusto (Disgeusia).

- **Link epidemiologico:**
  - esposizione a casi accertati;
  - esposizione a casi sospetti;
  - contatti con familiari di casi accertati/sospetti;
  - frequenza ambienti sanitari con casi accertati/sospetti.
- **Comorbilità:**
  - malattie polmonari;
  - malattie cardiache;
  - malattie renali;
  - malattie sistema immunitario;
  - malattie oncologiche;
  - malattie metaboliche.
- **Fattori rischio:**
  - ipertensione arteriosa sistemica;
  - diabete mellito tipo 1;
  - diabete mellito tipo 2;
  - abitudine tabagica;
  - ipertrigliceridemia;
  - sindrome metabolica.
- **Altre condizioni di rischio:**
  - gravidanza;
  - isolamento sociale;
  - non autosufficienza;
  - operatore sanitario;
  - RSA/lungodegente;
  - comunità chiuse.

1.3 Struttura Informatica

1.3.1 Software Ricovai-19

Il software Ricovai-19, schematizzato in Figura 1.3, si compone di un modulo client ed uno server. Il modulo client, consiste in una App mobile in grado di comunicare con il server della Control Room per acquisire ed aggiornare i dati dei pazienti e salvare lo stato dei dati inseriti.

L'App propone al paziente una serie di questionari, presentati secondo un ordine logico ben determinato, per acquisire diverse informazioni relative ad una eventuale sintomatologia ritenuta significativa nel caso di Covid 19, link epidemiologici, stato di salute generale ed eventuali patologie.

Essa è in grado di acquisire misure di parametri vitali tramite il dispositivo elettromedicale messo a disposizione da Ricovai, direttamente collegato al dispositivo mobile e all'app mediante connessione Bluetooh.

I dati inseriti, nella loro totalità e su richiesta del paziente, vengono inviati dall'App al modulo server di Ricovai, in forma completamente anonima.

Il modulo server consiste in un microservizio esposto su internet con protocollo sicuro https e scambio di messaggi POST in formato JSON (metodologia REST).

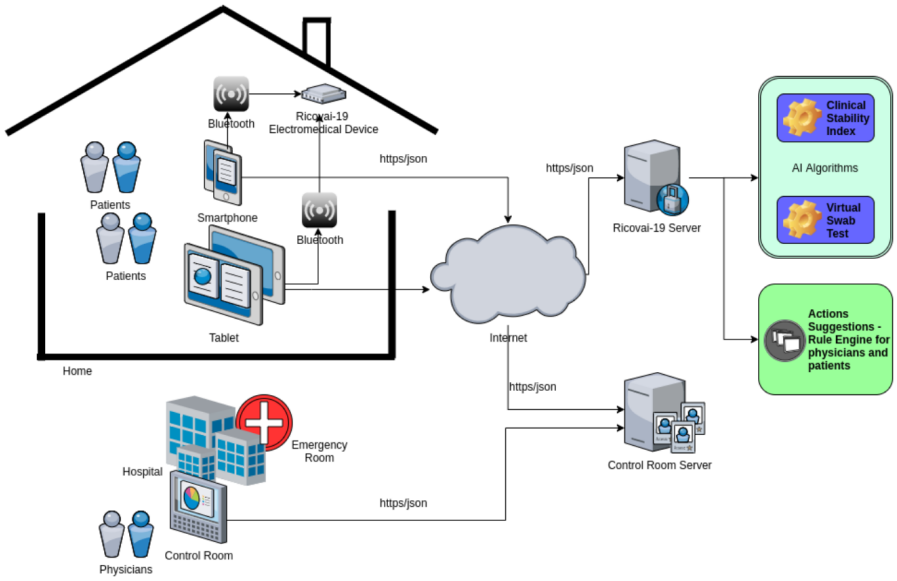


Figura 1.3. Struttura informatica

La richiesta si compone di una sezione di dati generali ed una con l'intero set di parametri clinici. Con questi dati, il sistema interroga due modelli di Intelligenza Artificiale addestrati per il calcolo dell'indice di stabilità clinica (CSI) ed il tampone virtuale (SWAB Index).

Inoltre, a seconda che il paziente sia in assistenza domiciliare pre o post ospedaliera o ricovero in ospedale, il sistema definisce dei suggerimenti su azioni da intraprendere, differenziando i messaggi per pazienti e medici.

In questo modo il motore di regole disegna un vero e proprio percorso clinico in grado di fornire un supporto nel decorso della malattia tanto al paziente quanto al personale medico (unitamente al valore dell'indice di stabilità clinica). I modelli di Intelligenza Artificiale vengono rivisti ed affinati via via che il sistema acquisisce nuovi dati.

1.3.2 L'applicazione

L'applicazione è utilizzabile sia in versione desktop che mobile. In entrambe le versioni per accedere occorrerà semplicemente inserire le proprie credenziali (Figura 1.4).

Effettuato l'accesso si visualizzerà la Control Room, dove si avrà la visibilità di tutti i pazienti attivi nel monitoraggio con gli ultimi valori rilevati, come visualizzato in Figura 1.5.

Selezionando uno dei pazienti presenti in dashboard si visualizzerà lo storico delle sue misurazioni. Questa funzionalità permette di monitorare l'andamento delle misure nel tempo.

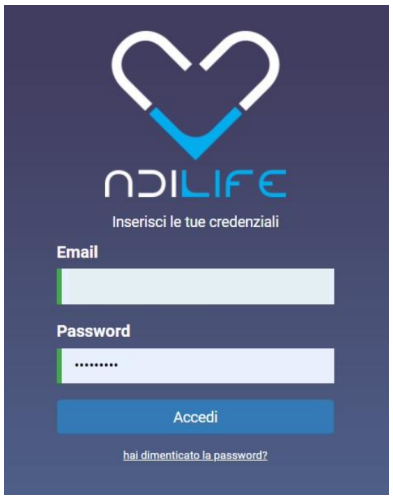


Figura 1.4. Login Utente

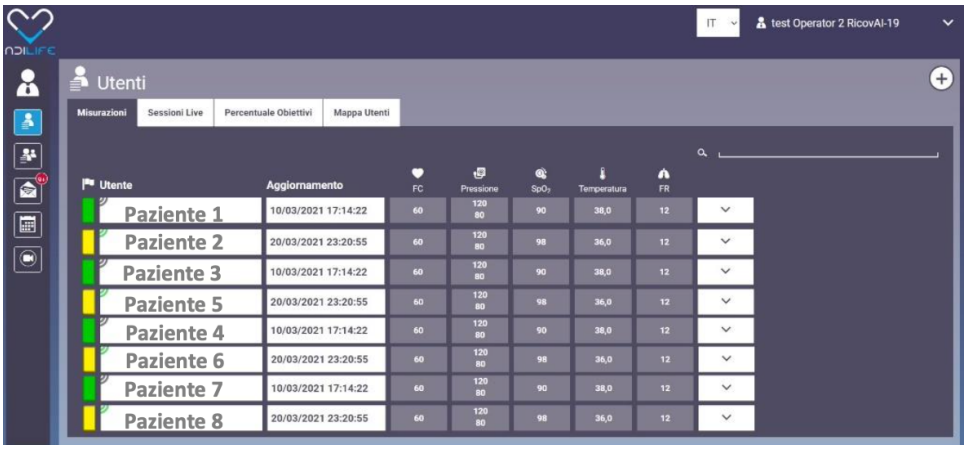


Figura 1.5. Control Room

In Figura 1.6 è mostrata la dashboard di ogni paziente; ad essa si accede cliccando sulla freccia in basso, relativa al paziente interessato nella Control Room. Da qui si aprirà una lista in cui il paziente o il medico che inserisce i dati potrà scegliere di inserire un nuovo questionario o consultare quelli precedentemente compilati. In questa finestra è presente anche uno spazio per annotazioni da parte del medico che, in base alle misurazioni e al valore atteso dei risultati, commenterà all'occorrenza e confronterà il valore dell'ISC osservato e atteso.

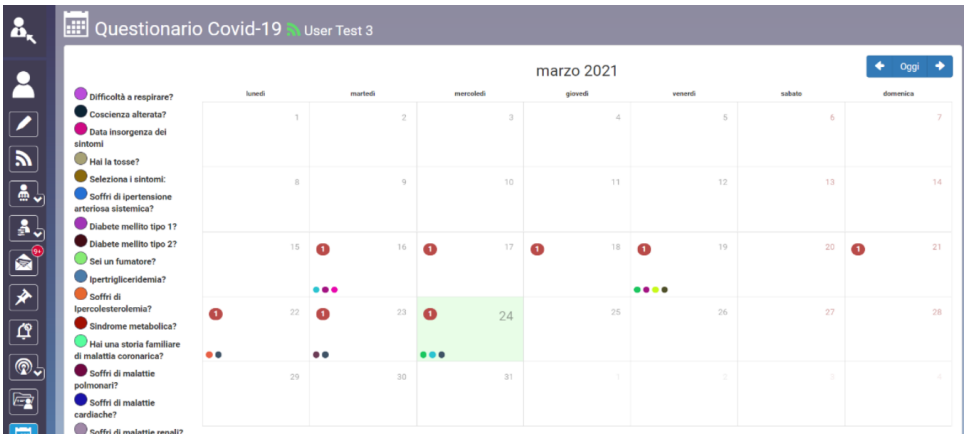


Figura 1.6. Storico Misurazioni

La piattaforma prevede, anche, la visualizzazione dei valori soglia dei parametri misurati (Figura 1.7); ciò risulta di importanza fondamentale per il coinvolgimento dei pazienti che, in questo modo, possono consultare semplicemente e immediatamente i parametri vitali. Di conseguenza, le misurazioni saranno più frequenti e anche più precise.

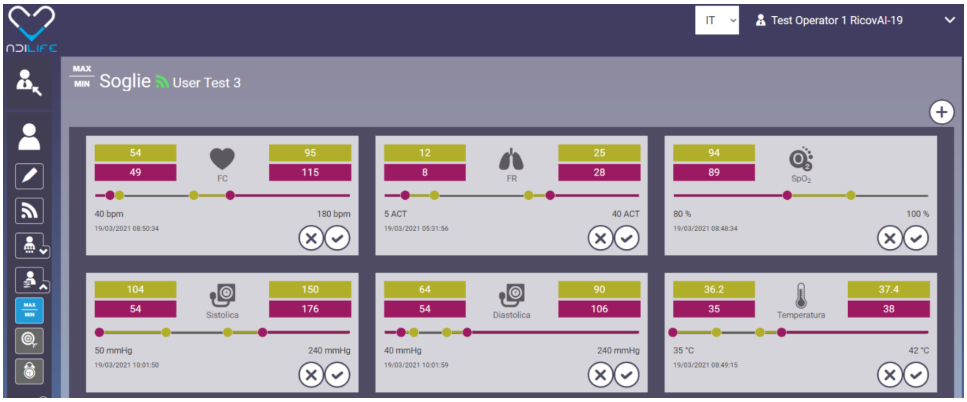


Figura 1.7. Valori soglia

2

Descrizione dataset di riferimento

In questo capitolo si esplorerà il dataset di riferimento. In particolare si i mostrerà il meccanismo di reclutamento dei pazienti fino ad approfondire il significato dei dati nel dataset stesso.

2.1 Reclutamento pazienti

I progetto Ricov-Ai ha coinvolto gli abitanti del comune di Offagna. I cittadini maggiorenni, affetti da Covid-19, su indicazioni mediche hanno ricevuto le strumentazioni per il monitoraggio autonomo recandosi in un locale messo a disposizione dal Comune. In questa occasione veniva garantita una spiegazione sull'utilizzo dell'applicazione con relativo invio dei risultati di un primo monitoraggio "test" al sistema per l'elaborazione di Alawave.

La sperimentazione è iniziata il 22 Marzo 2021 ed è terminata il 22 Ottobre 2021, di conseguenza, i dati analizzati ricoprono un intervallo di tempo della durata di otto mesi.

Durante tutto il periodo di sperimentazione sono stati contattati 158 pazienti di cui 129 hanno accettato di essere reclutati. Solo il 18% dei pazienti ha, quindi, rifiutato di partecipare al progetto.

In Tabella 2.1 e nella Figura 2.1 viene riportato l'andamento del reclutamento dei pazienti per mese. Risulta evidente che la resistenza del paziente alla sperimentazione è stata solo iniziale: già da Aprile, infatti, il numero di partecipanti è stato di 54. Ovviamente, il numero di partecipanti ha seguito il trend della curva dei contagi: durante il periodo estivo, insieme alla diminuzione dei casi di Covid, è diminuita anche la partecipazione al progetto sperimentale.

	In isolamento Covid-19	Reclutati RicovAI-19
Marzo	14	5
Aprile	70	54
Maggio	25	38
Giugno	7	2
Luglio	1	8
Agosto	22	9
Settembre	13	9
Ottobre	6	4
Totale (Marzo - Ottobre)	158	129

Tabella 2.1. Reclutamento pazienti

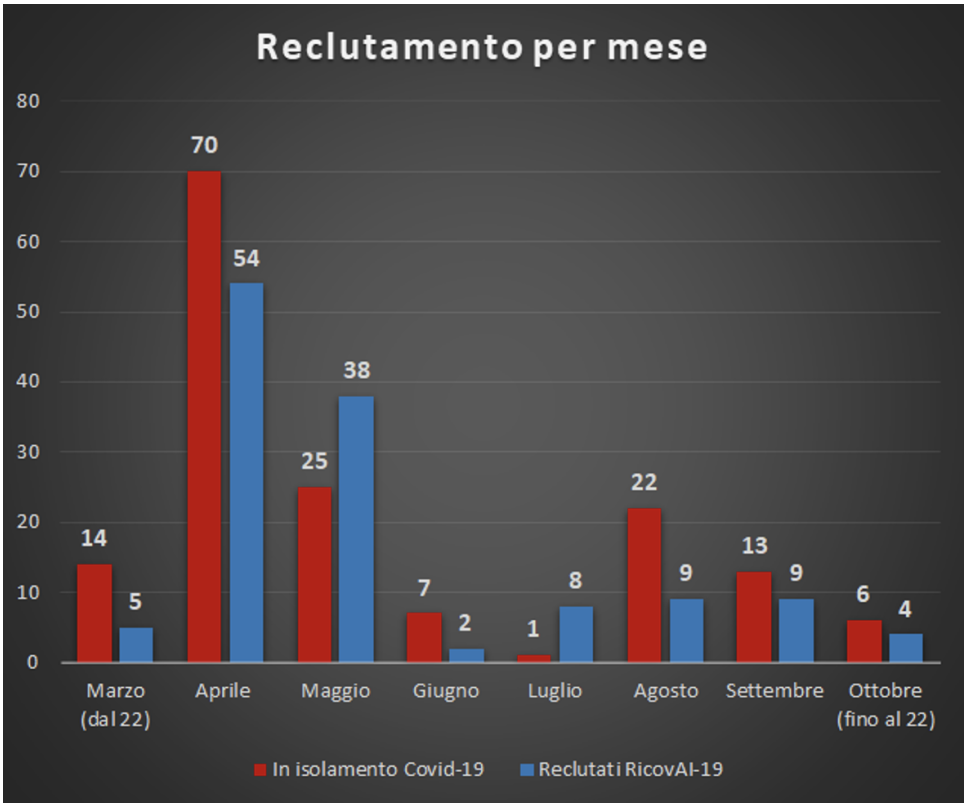


Figura 2.1. Reclutamento dei pazienti per mese

2.2 Catalogo delle variabili

I dati sono raccolti in un singolo dataset, quindi, in un'unica tabella. Le singole righe rappresentano le misurazioni, mentre le colonne denotano le variabili misurate. Ogni

variabile può essere distinta in base alla fonte, alla tipologia e al formato di dato.

Le fonti possono essere:

- *Questionari*: in tal caso, al paziente o al medico che inserisce informazioni nell'applicazione, è richiesto di compilare un semplice questionario rispondendo a domande sulla storia o sulla situazione clinica.
- *Misurazioni*: in questo caso le variabili sono direttamente misurate attraverso il dispositivo medico in dotazione; quando il paziente attua la misurazione, i dati vengono direttamente trasmessi all'applicazione.
- *Calcolate*: in questo caso i dati provengono o da risposte a questionari, o da misurazioni svolte con il dispositivo medico; tuttavia, diversamente dai casi precedenti i dati vengono processati e contestualizzati per poter essere inseriti all'interno di intervalli riconosciuti.

Le tipologie, invece, sono le seguenti:

- *dati della rilevazione*: descrivono le caratteristiche della misurazione;
- *anagrafica*: descrivono le caratteristiche del paziente;
- *misurazioni*: sono direttamente ricavate dal dispositivo in dotazione.
- *derivanti dai questionari*: sono identificati dalla lettera Q seguita da un valore numerico che distingue le varie tipologie di variabile; i valori possibili sono i seguenti:
  - Q1: sintomi allarme;
  - Q2: sintomi maggiori;
  - Q3: sintomi minori;
  - Q3: fattori di rischio;
  - Q3: comorbilità;
  - Q3: altre condizioni di rischio;
  - Q4: stato vaccinale;
  - Q5: tampone;
  - Q6: TAC/Rx torace;
  - Q7: link epidemiologico;
  - Q8: test sierologico;
  - Q9: sintomi e terapia;
  - Q10: test laboratorio;
  - Q11: stato neurologico;

  Ogni questionario ha diverse domande che, nel campo del dataset, vengono individuate dalla lettera D accompagnata dal numero corrispondente alla domanda stessa nel questionario

Per quanto riguarda il formato di dato, esso può essere uno dei seguenti:

- *data e ora*: sono indicate la data e l'ora in cui la misurazione è stata effettuata;
- *data*: è indicata la data in cui la variabile è stata registrata;
- *dicotomica*: in base alla risposta (SI o NO) o all'appartenenza ad un intervallo prestabilito, si ha 1 o 0;
- *politomica*: in base all'appartenenza ad intervalli prestabiliti, si hanno diversi valori numerici interi;
- *numerica*: generalmente riguardanti le misurazioni provenienti direttamente dal dispositivo medico.

Non tutte le misurazioni, ovvero i record del dataset, riportano tutte le variabili: alcune di esse, infatti, non sono da riportare obbligatoriamente. In tal caso all'interno del campo in questione è inserito il valore "null". Questa informazione è fondamentale per l'analisi più approfondita delle caratteristiche della situazione clinica del paziente.

2.2.1 Metadati

Il dataset è formato da 91 colonne ovvero 91 variabili.

Nelle tabelle sottostanti sono riportati i metadati ovvero tutti i dati di cui bisogna essere a conoscenza per poter correttamente formare, gestire e conservare nel tempo informazioni, del dataset di partenza. Essi sono suddivisi per tipologie.

In Tabella 2.2 vengono riportati i dati riguardanti la tipologia *Anagrafica*: essi riguardano la storia clinica del paziente.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
PATIENT_ID	Identificativo del paziente	Anagrafica	Numerica	Identificativo alfanumerico
Age_of_65	Età superiore a 65 anni	Anagrafica	Dicotomica	0=NO, 1=SI
Sex_Male	Sesso Maschile	Anagrafica	Dicotomica	0=NO, 1=SI
Admission_Discharge_ Hospital	Ricovero in ospedale	Anagrafica	Dicotomica	0=NO, 1=SI, null=Non Compilato

Tabella 2.2. Dati anagrafici presenti nel nostro dataset

In Tabella 2.3 vengono riportati i dati riguardanti la tipologia *Dati della rilevazione*: essi caratterizzano la singola misurazione indicandone informazioni temporali.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
ID	Identificativo del record	Dati della rilevazione	Numerica	Identificativo numerico
RECEIPT_TIMESTAMP	Data e ora della rilevazione	Dati della rilevazione	Data e ora	dal 22MAR2021
REQUEST_DATE	Data e ora dell'ultima misurazione	Dati della rilevazione	Data e ora	dal 22MAR2021

Tabella 2.3. Dati della rilevazione

In tabella 2.4 vengono riportati i dati riguardanti la tipologia *Comorbilità*. Il concetto di comorbilità o comorbidità in ambito sanitario indica la coesistenza di più patologie diverse nello stesso paziente.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Family_Coronary_ Artery_Disease	Storia familiare di malattia coronarica	Q3 D08	Dicotomica	0=NO, 1=SI, null=Non Compilato
Pulmonary_disease	Malattie polmonari	Q3 D09	Dicotomica	0=NO, 1=SI, null=Non Compilato
Cardiac_disease	Malattie cardiache	Q3 D10	Dicotomica	0=NO, 1=SI, null=Non Compilato
Kidney_disease_non_dialisi	Malattie renali	Q3 D11	Dicotomica	0=NO, 1=SI, null=Non Compilato
Liver_disease	Malattie del fegato	Q3 D12	Dicotomica	0=NO, 1=SI, null=Non Compilato
Immune_system_disease	Malattie del sistema immunitario	Q3 D13	Dicotomica	0=NO, 1=SI, null=Non Compilato
Known_Tumor	Malattie oncologiche	Q3 D14	Dicotomica	0=NO, 1=SI, null=Non Compilato
Metabolic_disease	Malattie metaboliche	Q3 D15	Dicotomica	0=NO, 1=SI, null=Non Compilato

Tabella 2.4. Dati relativi alle comorbilità

In Tabella 2.5 vengono riportati i dati riguardanti la tipologia *Fattori di rischio*. Queste variabili caratterizzano quei pazienti più inclini a complicazioni in caso di infezione virale da Covid-19.

In Tabella 2.6 vengono riportati i dati riguardanti la tipologia *Altre condizioni di rischio*. Queste variabili sono in aggiunta ai *Fattori di rischio* e identificano pazienti più fragili ed esposti a complicazioni.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Hypertension	Ipertensione arteriosa sistemica	Q3 D01	Dicotomica	0=NO, 1=SI, null=Non Compilato
Diabetes_type1	Diabete mellito tipo 1	Q3 D02	Dicotomica	0=NO, 1=SI, null=Non Compilato
Diabetes_type2	Diabete mellito tipo 2	Q3 D03	Dicotomica	0=NO, 1=SI, null=Non Compilato
Diabetes	Diabete	Q3 D03 Da Questionario	Dicotomica	0=NO, 1=SI, null=Non Compilato
Smoking	Abitudine tabagica	Q3 D04	Dicotomica	0=NO, 1=SI, null=Non Compilato
BMI_30	Ipertrigliceridemia	Q3 D05	Dicotomica	0=NO, 1=SI, null=Non Compilato
Dyslipidemia	Ipercolesterolemia	Q3 D06	Dicotomica	0=NO, 1=SI, null=Non Compilato
Metabolic_Syndrome	Sindrome metabolica	Q3 D07	Dicotomica	0=NO, 1=SI, null=Non Compilato

Tabella 2.5. Dati relativi ai fattori di rischio

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Pregnancy	Gravidanza	Q3 D16	Dicotomica	0=NO, 1=SI, null=Non Compilato
Social_isolation	Isolamento sociale	Q3 D17	Dicotomica	0=NO, 1=SI, null=Non Compilato
Not_self_sufficient	Non autosufficiente	Q3 D18	Dicotomica	0=NO, 1=SI, null=Non Compilato
Health_operator	Operatore sanitario	Q3 D19	Dicotomica	0=NO, 1=SI, null=Non Compilato
RSA	RSA/lungodegente	Q3 D20	Dicotomica	0=NO, 1=SI, null=Non Compilato
Closed_community	Comunità chiuse	Q3 D21	Dicotomica	0=NO, 1=SI, null=Non Compilato
Covid19_before	Covid contratto in passato	Q3 D22	Dicotomica	0=NO, 1=SI, null=Non Compilato
Covid19_before_date	Se si, data del tampone di guarigione	Q3 D23	Data	Data

Tabella 2.6. Dati relativi ad altre condizioni di rischio

In Tabella 2.7 vengono riportati i dati riguardanti la tipologia *Link epidemiologico*. Esso descrive l'ipotetico collegamento del paziente con persone risultate infette

o con aree in cui hanno circolato persone infette, da cui, quindi, può aver preso avvio una catena di contagio.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Exposition_known_ Covid_cases	Esposizione a casi accertati	Q7 D01	Dicotomica	0=NO, 1=SI, null=Non Compilato
Exposition_suspected_ Covid_cases	Esposizione a casi sospetti	Q7 D02	Dicotomica	0=NO, 1=SI, null=Non Compilato
Contacts_family_ known_cases	Contatti con familiari di casi accertati/sospetti	Q7 D03	Dicotomica	0=NO, 1=SI, null=Non Compilato
Frequentation_ healthcare_cases	Frequentazione di ambienti sanitari con casi accertati/sospetti	Q7 D04	Dicotomica	0=NO, 1=SI, null=Non Compilato

Tabella 2.7. Dati relativi al link epidemiologico

In Tabella 2.8 vengono riportati i dati riguardanti la tipologia *Misurazioni*. Le variabili in questione variano nel tempo e sono relative ai valori misurati direttamente dal dispositivo.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Systolic_BP	Pressione arteriosa sistolica (PAS)	Misurazioni	Numerica	Non negativi
Diastolic_BP	Pressione arteriosa diastolica	Misurazioni	Numerica	Non negativi, null=Non acquisito
Oxygen_Saturation	Saturazione di ossigeno	Misurazioni	Numerica	Non negativi, null=Non acquisito
Heart_rate	Frequenza cardiaca (FC)	Misurazioni	Numerica	Non negativi
Breath_rate	Frequenza respiratoria (FR)	Misurazioni	Numerica	Non negativi
Body_temp	Temperatura (TC)	Misurazioni	Numerica	Non negativi

Tabella 2.8. Dati relativi alle misurazioni

In Tabella 2.9 vengono riportati i dati riguardanti la tipologia *Sintomi* e *Sintomi e Terapia*. Queste informazioni riguardano l'insorgere dei sintomi caratterizzandone durata e terapia.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Symptom_onset_date	Data di insorgenza dei sintomi	Q2 D01	Data	Data
Persistent_symptoms _3_days	Sintomi persistenti (>3 giorni)	Q9 D01	Dicotomica	0=NO, 1=SI, null=Non Compilato
Not_Responder_ initial_therapy	Non risponde alla terapia standard iniziale	Q9 D02	Dicotomica	0=NO, 1=SI, null=Non Compilato

Tabella 2.9. Dati relativi ai sintomi e alle terapie svolte

In Tabella 2.10 vengono riportati i dati riguardanti la tipologia *Sintomi allarme*. Essi riguardano i sintomi più pericolosi per pazienti Covid-19 da dover monitorare costantemente.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Short_Of_Breath_degree_1	Difficoltà a respirare Livello +	Q1 D01	Dicotomica	0=NO, 1=SI
Short_Of_Breath_degree_2	Difficoltà a respirare Livello ++	Q1 D01	Dicotomica	0=NO, 1=SI
Short_Of_Breath_degree_3	Difficoltà a respirare Livello +++	Q1 D01	Dicotomica	0=NO, 1=SI
Impaired_consciousness	Coscienza alterata	Q1 D02	Dicotomica	0=NO, 1=SI
Systolic_Blood_Pressure_less_100	Pressione sanguigna sistolica <100	Q1 D03 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
Systolic_Blood_Pressure_over_200	Pressione sanguigna sistolica >200	Q1 D04 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
Heart_Rate_beats_per_minute	Frequenza cardiaca (battiti al minuto)	Q1 D05 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
Sat_O2_Over_97	Sat O2 >97%	Q1 D06 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
Sat_O2_91_97	Sat O2 91-97%	Q1 D07 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
Sat_O2_85_90	Sat O2 85-90%	Q1 D08 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
Sat_O2_less_85	Sat O2 <85%	Q1 D09 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI

Tabella 2.10. Dati relativi ai sintomi allarme

In Tabella 2.11 vengono riportati i dati riguardanti la tipologia *Sintomi maggiori*. Queste informazioni riguardano sintomi più frequenti nei pazienti infetti dal virus che hanno impatto più grave nel decorso della malattia.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Body_Temperature_37	Febbre tra 36,9 e 37,5	Q2 D02 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
Body_Temperature_37e5	Febbre >37,5	Q2 D02 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
cough_productive_ degree_1	tosse (produttiva) livello +	Q2 D03	Dicotomica	0=NO, 1=SI
cough_productive_ degree_2	tosse (produttiva) livello ++	Q2 D03	Dicotomica	0=NO, 1=SI
cough_productive_ degree_3	tosse (produttiva) livello +++	Q2 D03	Dicotomica	0=NO, 1=SI

Tabella 2.11. Dati relativi ai sintomi maggiori

In Tabella 2.12 vengono riportati i dati riguardanti la tipologia *Sintomi minori*. Queste informazioni riguardano sintomi più frequenti nei pazienti infetti dal virus che hanno impatto più lieve nel decorso della malattia ma che la caratterizzano.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Fatigue	Stanchezza	Q2 D04	Dicotomica	0=NO, 1=SI
Sore_Throat	Mal di gola	Q2 D05	Dicotomica	0=NO, 1=SI
Headache	Mal di testa	Q2 D06	Dicotomica	0=NO, 1=SI
Muscle_pain	Dolori muscolari	Q2 D07	Dicotomica	0=NO, 1=SI
Nasal_congestion	Congestione nasale	Q2 D08	Dicotomica	0=NO, 1=SI
Anosmia	Anosmia	Q2 D09	Dicotomica	0=NO, 1=SI
Disgeusia	Disgeusia	Q2 D10	Dicotomica	0=NO, 1=SI

Tabella 2.12. Dati relativi ai sintomi minori

In Tabella 2.13 vengono riportati i dati riguardanti la tipologia *Stato Neurologico*. Essi riguardano informazioni fondamentali sullo stato mentale del paziente.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Neurologic_state	Stato neurologico	Q11 D01	Politomica	0 = se neurologicamente attivo 1 = se neurologicamente risponde a stimolo vocale 2 = se neurologicamente risponde solo a stimolo del dolore 3 = se neurologicamente non è attivo
Breath_rate_ per_minute	Frequenza respiratoria al minuto	Q11 D02 Calcolata da Misurazioni	Dicotomica	0=NO, 1=SI
MEWS	Punteggio News	Q11 D03 Calcolata da Questionario e Misurazioni	Politomica	Valori da 0 a 3

Tabella 2.13. Dati relativi allo stato neurologico

In Tabella 2.14 vengono riportati i dati riguardanti la tipologia *Stato Vaccinale*. Queste informazioni riguardano il completamento dei cicli vaccinali che proteggono dalle malattie virali più diffuse.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Flu_vaccination	Vaccinazione antinfluenzale	Q4 D01	Dicotomica	0=NO, 1=SI, null=Non Compilato
Antipneumococcus _vaccination	Vaccinazione antipneumococco	Q4 D02	Dicotomica	0=NO, 1=SI, null=Non Compilato
Anticovid19_ vaccination	Vaccinazione anti-COVID 19	Q4 D03	Dicotomica	0=NO, 1=SI, null=Non Compilato
Anticovid19_ vaccination_date	Data ultima/unica dose di vaccinazione anti-COVID 19	Q4 D04	Data	Data

Tabella 2.14. Dati relativi allo Stato Vaccinale

In Tabella 2.15 vengono riportati i dati riguardanti la tipologia *TAC/Rx torace*. Essi riguardano informazioni relative a esami radiologici al torace per individuare patologie polmonari e cardiache.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
CTX_interstitial_ one_side_mild	CT/X-ray scan mild one side	Q6 D01	Dicotomica	0=NO, 1=SI, null=Non Eseguito
CTX_inter_ one_side_modere	CT/X-ray scan modere one side	Q6 D02	Dicotomica	0=NO, 1=SI, null=Non Eseguito
CTX_inter_ one_side_severe	CT/X-ray scan severe one side	Q6 D03	Dicotomica	0=NO, 1=SI, null=Non Eseguito
CTX_inter_ bilateral_mild	CT/X-ray scan mild bilateral	Q6 D04	Dicotomica	0=NO, 1=SI, null=Non Eseguito
CTX_inter_ bilateral_modere	CT/X-ray scan modere bilateral	Q6 D05	Dicotomica	0=NO, 1=SI, null=Non Eseguito
CTX_inter_ bilateral_severe	CT/X-ray scan severe " bilateral	Q6 D06	Dicotomica	0=NO, 1=SI, null=Non Eseguito

Tabella 2.15. Dati relativi a TAC/Rx torace

In Tabella 2.16 vengono riportati i dati riguardanti la tipologia *Tampone*. Queste variabili riportano informazioni utili relative ai tamponi effettuati specificandone tipo e data di esecuzione.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
COVID_swab_test	COVID-19 swab test positivo	Q5 D01	Dicotomica	0=NO, 1=SI, null=Non Compilato
COVID_swab_ test_type	Tipo di tampone	Q5 D02	Politomica	1 = PCR 2 = Antigenico rapido 3 = Salivare null = Non Compilato
COVID_swab_ test_date	Data esecuzione tampone	Q5 D03	Data	Data

Tabella 2.16. Dati relativi al tampone

In Tabella 2.17 vengono riportati i dati riguardanti la tipologia *Test Sierologico*. Queste variabili riportano informazioni utili relative ai test sierologici effettuati specificandone responso, tipo e data di esecuzione.

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
Sierologic_IGM	Test Sierologico IgM	Q8 D01	Dicotomica	0=Negativo, 1=Positivo, null=Non Eseguito
Sierologic_IGG	Test Sierologico IgG	Q8 D02	Dicotomica	0=Negativo, 1=Positivo, null=Non Eseguito
Sierologic_RBD	Test Sierologico RBD	Q8 D03	Numerica con 2 decimali	Valori non negativi
Sierologic_type	Tipo di Test Sierologico (Standard/Rapido/RBD)	Q8 D04	Politomica	0 = Non compilato, 1 = Standard, 2 = Rapido, 3 = RBD, null = Non Eseguito
Sierologic_date	Data del Test Sierologico	Q8 D05	Data	Data

Tabella 2.17. Dati relativi ai test sierologici

In Tabella 2.18 vengono riportati i dati riguardanti la tipologia *Test di laboratorio*. Queste variabili riportano informazioni utili relative a specifici esami istologici svolti per escludere patologie che potrebbero aggravare le condizioni del paziente

NOME DELLA VARIABILE	DESCRIZIONE	FONTE	TIPO	VALORI AMMISSIBILI
D_Dimero	D-Dimero anormale	Q10 D01	Dicotomica	0=NO, 1=SI, null=Non Compilato
Troponine	Troponine anormale	Q10 D02	Dicotomica	0=NO, 1=SI, null=Non Compilato
PCR	PCR anormale	Q10 D03	Dicotomica	0=NO, 1=SI, null=Non Compilato

Tabella 2.18. Dati relativi ai test di laboratorio

2.3 Misurazioni per paziente

I dati appena analizzati sono relativi ad ogni record ovvero ad ogni misurazione che tutti i pazienti partecipanti al progetto hanno effettuato durante lo studio.

All'interno del dataset ogni misurazione è univoca ed è identificata con un codice numerico. Per ogni misurazione è indicato il paziente relativo alla rilevazione in esame; anche quest'ultimo è rappresentato con un indice numerico univoco che rende le informazioni anonime. Nella Tabella 2.19 sono riportati alcuni dei campi del dataset relativi a due pazienti.

Come si evince, i partecipanti al progetto non hanno effettuato e inviato lo stesso numero di misurazioni.

ID	PATIENT_ID	RECEIPT_ TIMESTAMP	REQUEST_DATE	Age_of_65	Sex_Male	Admission_ Discharge_ Hospital	Impaired_ consciousness	Heart_Rate_ beats_per_minute
72378	23467	12/07/2021 19:23	12/07/2021 20:13	0	0	0	0	0
72376	23467	12/07/2021 19:20	12/07/2021 20:13	0	0	0	0	0
72347	23467	08/07/2021 17:04	08/07/2021 17:43	0	0	0	0	0
72344	23467	08/07/2021 16:59	08/07/2021 17:43	0	0	0	0	0
67562	16982	27/05/2021 16:20	27/05/2021 17:20	0	1	0	0	0
67559	16982	27/05/2021 16:15	24/04/2021 18:15	0	1	0	0	0
4785	16982	24/04/2021 17:15	24/04/2021 18:15	0	1	0	0	0
6964	16982	03/05/2021 16:29	24/04/2021 18:15	0	1	0	0	0
6962	16982	03/05/2021 16:28	24/04/2021 18:15	0	1	0	0	0
6961	16982	03/05/2021 16:25	24/04/2021 18:15	0	1	0	0	0
6960	16982	03/05/2021 16:24	24/04/2021 18:15	0	1	0	0	0
6959	16982	03/05/2021 16:22	24/04/2021 18:15	0	1	0	0	0
6938	16982	03/05/2021 14:30	24/04/2021 18:15	0	1	0	0	0
6933	16982	03/05/2021 11:00	24/04/2021 18:15	0	1	0	0	0
6930	16982	03/05/2021 10:44	24/04/2021 18:15	0	1	0	0	0

ID	Systolic_BP	Diastolic_BP	Oxygen_ Saturation	Heart_rate	Breath_rate	Body_temp	Short_Of_Breath _degree	Systolic_ Blood_Pressure	Sat_O2
72378	119	74	93	96	20	36.9	0	0	2
72376	119	74	93	96	3	36.9	0	0	2
72347	102	90	90	59	18	36.9	0	0	3
72344	102	90	90	59	12	36.9	0	0	3
67562	166	99	99	59	14	36.6	0	0	1
67559	167	103	99	59	14	36.6	0	0	1
4785	108	0	0	71	16	36.6	0	0	2
6964	114	0	0	83	20	37.3	0	0	0
6962	114	0	0	83	0	37.3	0	0	0
6961	114	0	0	83	14	37.3	0	0	0
6960	114	0	0	83	14	36.7	0	0	0
6959	114	0	0	74	14	36.7	0	0	2
6938	136	0	0	61	14	36.7	0	0	2
6933	159	0	0	53	14	36.7	0	0	2
6930	169	0	0	55	14	36.7	0	0	2

Tabella 2.19. Esempio: Pazienti 23467 e Pazienti 16892

Il primo paziente in esame, identificato con il PAZIENT_ID 23467 infatti, riporta solo 4 misurazioni; di conseguenza l'identificativo del paziente è presente in 4 record consecutivi. I rimanenti record corrispondono invece alle misure effettuate e trasmesse dal paziente identificato con il PAZIENT_ID 16982.

Nelle tabelle sono riportate solo alcune delle variabili del dataset; seppur non completo l'esempio evidenzia l'eterogeneità dei dati: sono presenti, infatti, date e orari, valori numerici interi di tipo dicotomico e politomico, e interi o float riguardanti variabili misurate o calcolate dai questionari.

3

Attività di ETL (prima parte)

In questo capitolo sarà illustrato il processo di ETL, ovvero Extract, Transform e Load. Questa espressione si riferisce al processo di estrazione, trasformazione e caricamento dei dati in un sistema di sintesi. Si esplorerà, quindi, il processo di manipolazione dei dati per poter svolgere, successivamente, le analisi necessarie all'estrazione di informazioni.

3.1 Introduzione

3.1.1 L'ambiente di calcolo

L'ambiente di calcolo utilizzato per poter estrarre conoscenza e portare avanti l'analisi dei dati è Google Colab il cui logo è riportato in Figura 3.1 .

Figura 3.1. Logo di Google Colab

Esso è uno strumento gratuito presente nella suite Google che consente di scrivere codice Python direttamente dal browser. Viene, quindi, sfruttata una piattaforma online che offre un servizio di cloud hosting per notebook Jupyter dove creare documenti che contengono righe di codice, grafici, testi, link e molto altro. Il notebook Jupiter è eseguito su macchine virtuali di server Google. Ciò consente di svincolarsi dalla parte hardware e di concentrarsi soltanto sul codice e sui contenuti che si vogliono integrare.

Il documento creato è stato condiviso con altri collaboratori interni al progetto che hanno avuto la possibilità di prendere visione di aggiornamenti sul codice in tempo reale e, in più, di lasciare commenti in modo da velocizzare il confronto tra le parti.

3.1.2 Librerie utilizzate e caricamento dei dati

Per poter gestire e studiare i dati è necessario importare diverse librerie.

Tra le tante utilizzate, una delle più importanti è NumPy, il cui logo è in Figura 3.2. Essa è una libreria open source per il linguaggio di programmazione Python, che viene utilizzata per la creazione e manipolazione di grandi matrici e array multi-dimensionali; inoltre, ha una vasta collezione di funzioni matematiche di alto livello per poter operare efficientemente su queste strutture dati.

Figura 3.2. Logo di NumPy

Altra libreria ampiamente utilizzata nello studio è Pandas, il cui logo è riportato in Figura 3.3.

Essa è una libreria open source con licenza BSD (Berkeley Software Distribution) che fornisce strutture dati e strumenti di analisi dei dati ad alte prestazioni e di facile utilizzo, per il linguaggio di programmazione Python. Pandas è costruita su NumPy ed è destinata ad integrarsi bene in un ambiente di calcolo scientifico con molte altre librerie di terze parte utilizzate successivamente per scopi più specifici.

Figura 3.3. Logo di Pandas

Acune operazioni svolte grazie all'utilizzo di Pandas sono le seguenti:

- gestione dei dati mancanti;
- inserimento e eliminazione di colonne da DataFrame e oggetti di dimensioni superiori;
- combinazione su insieme di dati, aggregazione e trasformazioni di essi;
- utilizzo di strumenti di I/O per caricare dati da file Excel.

Ultima libreria che è importante menzionare è Scikit-learn il cui logo è in Figura 3.4. Anch'essa è una libreria open source; si occupa di apprendimento automatico

per il linguaggio di programmazione Python. Scikit-learn si integra bene con le più usate librerie Python, come Matplotlib per la stampa, NumPy per la vettorizzazione degli array, Pandas, SciPy e molte altre. Essa verrà utilizzata in buona parte del progetto in quanto è il tool più semplice e, contemporaneamente, più efficiente per l'implementazione degli algoritmi di clustering e classificazione.

Figura 3.4. Logo di Scikit-learn

3.2 Analisi esplorativa

Per poter svolgere analisi descrittive e applicare algoritmi di intelligenza artificiale, il primo passaggio svolto è stato quello di conoscere e di esaminare i dati prima di manipolarli.

Grazie all'utilizzo di Pandas è stato possibile caricare i dati del progetto RicovAi-19. Essi sono stati condivisi in assoluto anonimato grazie alla collaborazione tra l'Università e l'azienda Almawave. Quest'ultima dopo aver pre-processato i dati, li ha trasferiti in un unico file Excel mettendo, quindi, a disposizione un dataset iniziale composto da 6624 righe e 91 colonne. Esso è stato condiviso il 14/06/2021, data in cui i pazienti reclutati nel progetto erano 102. Da subito quindi non è stato possibile fare delle considerazioni totali rispetto all'andamento di alcuni parametri che necessitavano di tutte le informazioni temporali e non, riguardanti la fine della malattia o l'insorgenza di alcuni sintomi.

I codici scritti, però, sono stati resi standard per poter essere adattati a qualsiasi quantità di dati e, di conseguenza, è stato semplice aggiornare il dataset di partenza inserendo nuovi pazienti e nuove rilevazioni. In data 25/10/2021, quindi in fase di chiusura del progetto, i pazienti reclutati erano 128 e il nuovo dataset a nostra disposizione era composto da 7386 righe e 91 colonne.

Una volta caricato sul drive personale, il file è stato trasformato in formato *.csv* in modo da poter manipolare il dataset agevolmente.

In Tabella 3.1 è riportato parte del dataset originale: ogni riga è indicizzata dal campo ID, ovvero un codice numerico per ogni misurazione. Nella seconda colonna è, invece, disponibile il codice del paziente che lo identifica univocamente.

ID	PATIENT_ID	RECEIPT_ TIMESTAMP	REQUEST_ DATE	Age_of _65	Sex_ Male	Admission_Discharge Hospital	Heart_Rate_ beats_per_min	SYMPTOM_ ONSET_DATE
1340	16861	22/03/2021 12:04	19/03/2021 18:07	0	1	0	0	17/03/2021 01:00
67562	16982	27/05/2021 16:20	27/05/2021 17:20	0	1	0	0	

Tabella 3.1. Prime righe e prime colonne del dataset originale

3.2.1 Numero misurazioni per ogni paziente

La prima considerazione svolta riguarda il numero di misurazioni per ogni paziente: tale informazione risulta utile per poter estrarre informazioni che lo riguardano.

Ciò è stato possibile tramite l'utilizzo della funzione *GroupBy* che permette l'aggregazione dei dati rispetto ad un parametro, che, nel caso in esame, è PATIENT_ID. Viene quindi creato un nuovo DataFrame con due colonne: ID_paziente e num_misurazioni.

Una parte di quest'ultimo è riportato nella Tabella 3.2 in cui è evidente che non tutti i pazienti hanno svolto lo stesso numero di rilevazioni; ciò è dipeso anche dal periodo di permanenza nello studio che è direttamente rapportato alla durata della malattia e alla persistenza dei sintomi.

ID_paziente	num_misurazioni
16861	1
16982	11
20329	7
22729	121
22739	112
22740	118
22814	72
22833	2
22880	3

Tabella 3.2. DataFrame riguardante tutti i pazienti e le loro rispettive misurazioni

Pur avendo accorpato i valori non si ha perdita di informazione: la colonna num_misurazioni nel DataFrame appena creato può essere, infatti, facilmente aggiunta al dataset generale ripetendo per ogni riga riguardante lo stesso paziente, lo stesso valore nella nuova colonna.

Aggiungere questo tipo di informazione risulta utile per poter svolgere considerazioni generali riguardanti il singolo paziente, e, per poter approfondire analisi di tipo temporali, è stato necessario riportare la nuova colonna generata all'interno del dataset originario.

Nella Tabella 3.3 è riportato il dataset aggiornato: è presente la nuova colonna inserita come secondo parametro. I due pazienti presi in esame sono i primi due della Tabella 3.2 precedentemente discussa.

3.2.2 Informazioni utili generali

Durante l'analisi esplorativa sono state svolte considerazioni sull'età e il sesso dei pazienti reclutati. Anche in questo caso è stata, quindi, utilizzata la funzione *GroupBy* e sono stati conteggiati i pazienti con età superiore a 65 anni, grazie al parametro Age_of_65, e i pazienti di sesso maschile, grazie al parametro Sex_Male.

I risultati sono riportati nella Tabella 3.4 : i dati sono relativi ai due dataset presi in esame nella prima fase del progetto e al termine di esso.

ID	PATIENT_ID	num_ misurazioni	RECEIPT_ TIMESTAMP	REQUEST_ DATE	Age _of_65	Sex_ Male	Admission_Discharge Hospital	Heart_Rate_ beats_per_minute
1340	16861	1	22/03/2021 12:04	19/03/2021 18:07	0	1	0	0
67562	16982	11	27/05/2021 16:20	27/05/2021 17:20	0	1	0	0
67559	16982	11	27/05/2021 16:15	24/04/2021 18:15	0	1	0	0
4785	16982	11	24/04/2021 17:15	24/04/2021 18:15	0	1	0	0
6964	16982	11	03/05/2021 16:29	24/04/2021 18:15	0	1	0	0
6962	16982	11	03/05/2021 16:28	24/04/2021 18:15	0	1	0	0
6961	16982	11	03/05/2021 16:25	24/04/2021 18:15	0	1	0	0
6960	16982	11	03/05/2021 16:24	24/04/2021 18:15	0	1	0	0
6959	16982	11	03/05/2021 16:22	24/04/2021 18:15	0	1	0	0
6938	16982	11	03/05/2021 14:30	24/04/2021 18:15	0	1	0	0
6933	16982	11	03/05/2021 11:00	24/04/2021 18:15	0	1	0	0
6930	16982	11	03/05/2021 10:44	24/04/2021 18:15	0	1	0	0

Tabella 3.3. Dataset aggiornato con la nuova colonna *num_misurazioni*

		Numero pazienti dataset Giugno	Numero pazienti dataset chiusura progetto
Età	Superiore a 65	36	48
	Inferiore a 65	66	80
Sesso	Maschile	36	60
	Femminile	66	68

Tabella 3.4. Conteggio dei pazienti in base all'età e al sesso

Inoltre sono stati anche conteggiati i pazienti ricoverati in ospedale tramite il parametro *Admission_Discarge_Hospital*. Questa variabile, però, non è obbligatoria: ha come valore ammissibile, oltre a 1 e a 0, che corrispondono al SI e al NO della risposta del questionario, anche il valore NaN, che equivale a tutti quei casi in cui il paziente o il medico non inserisce l'informazione.

Nel dataset risalente alla prima fase progettuale i valori NaN erano 94 su 102 pazienti reclutati; i partecipanti, invece, ricoverati in ospedale risultavano 7; un unico paziente riportava valori diversi durante tutte le misurazioni. Egli ha, quindi, inserito i valori sia durante l'assistenza domiciliare che durante il ricovero ospedaliero. Risultava quindi interessante analizzare questo singolo caso.

Dopo aver isolato il paziente in esame, è stato possibile conteggiare il numero di rilevazioni svolte in ospedale, ovvero 2, e quelle effettuate a casa, ovvero 78. Ci si è poi concentrati su alcuni parametri riportati nel documento ufficiale come "Sintomi Allarme" ed è stato evidente che tra il ricovero e il rientro a casa c'erano differenze sostanziali. Il paziente, infatti, riportava difficoltà a respirare al livello ++ *(Short_Of_Breath_degree_2)* in tutto il periodo ospedaliero mentre nel periodo a casa il valore era sempre a 0. Il parametro *Sat_O2_91-97%* era 1 per tutto il ricovero ospedaliero; mentre, per il periodo a casa, era a 1 per le 5 misurazioni prima di essere ricoverato, e a 0 per le restanti 73.

Ques'ultima considerazione evidenzia uno degli scopi del progetto: il dispositivo medico utilizzato riesce a monitorare il paziente, controllando i parametri vitali e unicamente quando questi sono diversi dagli intervalli standard di normalità, si ricorre all'ospedalizzazione.

Nel rispettivo dataset alla fine del progetto, i pazienti ricoverati risultano 8 mentre quelli che non hanno avuto bisogno di assistenza ospedaliera sono 120, ovvero la maggioranza. Il conteggio a zero dei valori NaN mostra un punto importante dello

sviluppo del progetto che verrà discusso nella prossima sezione.

3.3 Accorpamento dei valori

Durante una prima analisi esplorativa è emerso che molti dei parametri presi in esame, erano gestiti in modo dicotomico per intervalli.

Uno di questi è, ad esempio, la variabile *Short_Of_Breath_degree*: essa è gestita come mostrato in Tabella 3.5.

VARIABILE	DESCRIZIONE	VALORI AMMISSIBILI
Short_Of_Breath_degree_1	Difficoltà a respirare Livello +	0=NO, 1=SI
Short_Of_Breath_degree_2	Difficoltà a respirare Livello ++	0=NO, 1=SI
Short_Of_Breath_degree_3	Difficoltà a respirare Livello +++	0=NO, 1=SI

Tabella 3.5. Gestione del parametro *Short_Of_Breath_degree*

I pazienti, ad ogni misurazione, rispondevano al questionario inserendo il grado di difficoltà nella respirazione.

All'interno del dataset originale i campi relativi a questo unico parametro sono tre; ciò sta a significare che, per ogni misurazione, almeno uno dei tre campi è posto al valore 1. Quindi, il paziente può presentare uno dei tre livelli di difficoltà o nessuno di essi ma, sicuramente, non può ammetterne più di uno contemporaneamente.

Per rendere più snello il dataset ma anche per poter applicare algoritmi di intelligenza artificiale senza perdere informazioni, le colonne del parametro *Short_Of_Breath_degree* e di altri parametri con la stessa caratteristica appena spiegata, sono stati manipolati in modo tale da avere un unico parametro riassuntivo.

Nell'esempio preso in esame, è stata quindi aggiunta un'unica colonna con i seguenti quattro possibili valori:

- *0*: se il paziente è senza sintomi;
- *1*: se il paziente ha difficoltà a respirare di livello +;
- *2*: se il paziente ha difficoltà a respirare di livello + + ;
- *3*: se il paziente ha difficoltà a respirare di livello + + + .

Come è infatti possibile notare nella Tabella 3.6, in base ai valori iniziali del dataset assegnati ai tre campi, si attribuisce il valore corrispondente.

Short_Of_Breath_degree		
PARAMETRO	**VALORI DATASET INIZIALE**	**VALORE ASSEGNATO**
Short_Of_Breath_degree_1	0	
Short_Of_Breath_degree_2	0	0
Short_Of_Breath_degree_3	0	
Short_Of_Breath_degree_1	1	
Short_Of_Breath_degree_2	0	1
Short_Of_Breath_degree_3	0	
Short_Of_Breath_degree_1	0	
Short_Of_Breath_degree_2	1	2
Short_Of_Breath_degree_3	0	
Short_Of_Breath_degree_1	0	
Short_Of_Breath_degree_2	0	3
Short_Of_Breath_degree_3	1	

Tabella 3.6. Nuova assegnazione di valori del parametro *Short_Of_Breath_degree*

In questo modo, anziché avere tre variabili dicotomiche per esprimere un unico sintomo, si avrà un unico parametro di tipo politomico.

Per poter svolgere questa operazione di manipolazione, è stato scritto un codice apposito che permette di agglomerare le informazioni in un'unica colonna senza perdere l'indicizzazione delle misurazioni e dei pazienti.

In tutto il dataset i parametri modificati come spiegato sono stati i seguenti:

- **Short_Of_Breath_degree**: rappresenta la difficoltà nel respirare ed è suddiviso, a seconda del grado, nelle tre seguenti colonne:
  - *Short_Of_Breath_degree 1;*
  - *Short_Of_Breath_degree 2;*
  - *Short_Of_Breath_degree 3.*
- **Sat_O2**: rappresenta la saturazione dell'ossigeno, ovvero la percentuale di molecole di ossigeno legate all'emoglobina, ed è suddiviso, in base alle misurazione, nelle quattro seguenti colonne:
  - *Sat_O2_Over_97;*
  - *Sat_O2_91_97;*
  - *Sat_O2_85_90;*
  - *Sat_O2_less_85.*
- **Cough_productive_degree**: rappresenta l'infiammazione delle vie aeree più comunemente conosciuta come tosse e, in base al grado, è suddiviso nelle quattro seguenti colonne:
  - *cough_productive_degree_1;*
  - *cough_productive_degree_2;*
  - *cough_productive_degree_3;*
  - *cough_productive_degree_4.*
- **Body_Temperature**: rappresenta la temperatura corporea ed è suddivisa, in base alle misurazioni, nelle due seguenti colonne:

  - *Body_Temperature_37;*
  - *Body_Temperature_37e5.*
- **Systolic_Blood_Pressure**: rappresenta la pressione sanguigna a ogni battito del cuore ed è suddiviso, in base alle misurazione, nelle due seguenti colonne:
  - *Systolic_Blood_Pressure_less_100;*
  - *Systolic_Blood_Pressure_over_200.*

4

Attività di ETL seconda parte

Nel capitolo corrente verrà trattata la seconda fase di ETL che si sofferma principalmente sulla individuazione dei campi vuoti del dataset; verranno, poi, illustrata la logica di assegnazione dei valori ai campi NaN e le problematiche ad asse asssociate

4.1 Gestione valori NaN

Dall'analisi esplorativa svolta è emerso che il dataset risalente a giugno avesse molteplici parametri con campi vuoti che, quindi, erano riferiti a rilevazioni non eseguite o a domande dei questionario non compilate.

Questo risultava essere un problema: ogni riga e ogni colonna con campi NaN dovevano essere escluse dal dataset per poter implementare algoritmi di Intelligenza Artificiale. Quindi, si aveva una grossa perdita di informazione.

4.1.1 Occorrenze dei valori NaN

Per poter capire come ovviare al problema dei valori NaN il primo passaggio è stato quello di contare il numero dei campi vuoti.

In questa fase progettuale sono state contate tutte le misurazioni per cui un determinato parametro non veniva registrato. Partendo dal dataset originale sono stati contati i campi vuoti.

Per poter visualizzare immediatamente il problema è stata creato un nuovo DataFrame riportato in Tabella 4.1 in cui sono mostrati tutti i parametri studiati con il rispettivo conteggio dei valori NaN.

Un aspetto importante da notare è che il numero delle occorrenze si ripetono, sintomo del fatto che, nel momento dell'inserimento dei dati, essi sono stati compromessi nelle stesse misurazioni.

Nel grafico in Figura 4.1 sono riportati i parametri misurati che hanno occorrenze di valori NaN maggiori di 2000 e che, ricordando che il numero totale di rilevazioni e quindi di righe nel dataset risalente a Giugno era di 6625, hanno un impatto maggiore.

Parametro	Occorrenze NaN per parametro	Parametro	Occorrenze NaN per parametro	Parametro	Occorrenze NaN per parametro
Admission_Discharge_Hospital	6483	Immune_system_disease	999	CTX_inter_bilateral_mild	5886
SYMPTOM_ONSET_DATE	6074	Known_Tumor	1080	CTX_inter_bilateral_modere	5886
Hypertension	564	Metabolic_disease	1080	CTX_inter_bilateral_severe	5886
Diabetes_type1	564	Pregnancy	1119	Exposition_known_Covid_cases	1670
Diabetes_type2	564	Social_isolation	1120	Exposition_suspected_Covid_cases	1670
Diabetes	564	Not_self_sufficient	1120	Contacts_family_known_cases	1670
Smoking	564	Health_operator	1120	Frequentation_healthcare_cases	1670
BMI_30	6523	RSA	1120	Sierologic_IGM	5750
Dyslipidemia	564	Closed_community	1120	Troponine	6097
Metabolic_Syndrome	564	Covid19_before	6561	PCR	6190
Family_Coronary_Artery_Disease	564	Covid19_before_date	6623	Neurologic_state	6624
Pulmonary_disease	871	Flu_vaccination	656	MEWS	6624
Cardiac_disease	907	Antipneumococcus_vaccination	656	Diastolic_BP	6528
Kidney_disease_non_dialisi	925	Anticovid19_vaccination	656	Oxygen_Saturation	6528
Liver_disease	999	Anticovid19_vaccination_date	6138	Sierologic_IGG	5750
COVID_swab_test	2044	Sierologic_RBD	6624	CTX_interstitial_one_side_mild	5886
COVID_swab_test_type	3893	Sierologic_type	6080	CTX_inter_one_side_modere	5886
COVID_swab_test_date	4023	Sierologic_date	6624	CTX_inter_one_side_severe	5886
Persistent_symptoms_3_days	1754	Not_Responder_initial_therapy	2870	D_Dimero	6169

Tabella 4.1. Occorrenze NaN per ogni parametro del dataset

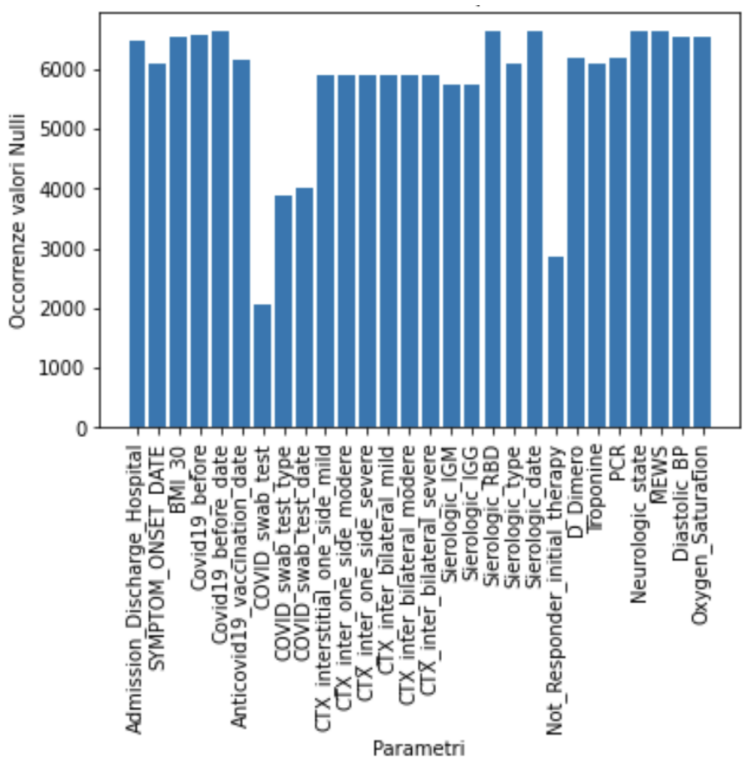


Figura 4.1. Occorrenza dei valori nulli (NaN) per i parametri più impattanti

4.1.2 Conteggio dei valori NaN per paziente

Per poter analizzare al meglio il dataset è stata condotta un'ulteriore analisi riguardante le misurazioni svolte per ogni paziente.

Il numero di pazienti è 102, ciascuno con le proprie misurazioni. Non essendo un numero elevato, è utile contare quante delle misurazioni per ogni pazienti vengono svolte e di queste quali siano effettivamente utilizzabili (non sono quindi valori NaN).

Per ogni colonna del dataset, ovvero per ogni variabile misurata, è stata valutata la percentuale di valori utilizzabili considerando il numero di valori totali, ovvero di tutte le misurazioni per quel paziente, e quelli ammissibili.

Dunque, è stata creata una tabella con 102 righe, ovvero il numero di pazienti, e 91 colonne, ovvero il numero di parametri valutati. Parte di essa è riportata nella Tabella 4.2.

Per poter visualizzare nel modo più immediato possibile la quantità di dati utilizzabili, i campi della tabella sono stati colorati come di seguito specificato:

- *verde*: percentuale di valori utilizzabili tra il 90% e il 100% ;
- *azzurro*: percentuale di valori utilizzabili tra il 70% e il 90% ;
- *giallo*: percentuale di valori utilizzabili tra il 40% e il 70%
- *arancione*: percentuale di valori utilizzabili tra il 20% e il 40%
- *rosso*: percentuale di valori utilizzabili tra lo 0 % e il 20%.

ID_paziente	num_misurazioni	Hypertension	Diabetes	BMI_30	Dyslipidemia	Cardiac_disease	Kidney_disease_non_dialisi	Liver_disease
16861	1	100	100	0	100	100	100	100
16982	11	9,09	9,09	0	9,09	9,09	9,09	9,09
20329	7	100	100	0	100	100	100	100
22729	121	97,52	97,52	0	97,52	97,52	97,52	97,52
22739	112	100	100	0	100	100	100	34,82
22740	118	100	100	0	100	100	100	100
22814	72	100	100	0	100	50	50	50
22833	2	100	100	0	100	100	100	100
22880	3	0	0	0	0	0	0	0
22881	58	0	0	0	0	0	0	0
22882	110	81,82	81,82	0	81,82	81,82	81,82	81,82
22883	150	82,67	82,67	0	82,67	82,67	82,67	82,67
22884	79	0	0	0	0	0	0	0

Tabella 4.2. Percentuale dei valori utilizzabili di tutti i parametri rispetto ad ogni paziente

In questo modo sarà più semplice valutare su quali parametri concentrarsi ed, eventualmente, isolare i pazienti con meno misurazioni per poter analizzare alcune variabili con buona percentuale di valori utilizzabili.

Anche in questo caso, risulta evidente che la percentuale di valori mancanti per ogni paziente si ripete spesso per più parametri; questo sottolinea il fatto che le rilevazioni con mancata acquisizione del dato sono molto probabilmente le stesse.

Un'analisi interessante svolta si è soffermata sull'età dei pazienti. L'idea di base era che la mancata registrazione poteva essere imputata alla difficoltà di utilizzo dell'applicazione.

Quindi, dopo aver isolato le misurazioni dei pazienti con età superiore a 65 anni, è stato svolto il conteggio. Il risultato dell'analisi ha confutato la tesi iniziale: solo il 33% dei valori NaN erano relativi a pazienti con età superiore a 65 anni.

Di conseguenza, la mancata o scorretta esecuzione della misurazione non è da collegarsi direttamente all'età più avanzata del paziente.

4.2 Assegnazione valori NaN

Una volta esplorato il problema della presenza di molti campi NaN all'interno del datase iniziale, è stata considerata l'idea di assegnare valori ai campi vuoti, sulla base delle informazioni in possesso.

Alcuni dei parametri studiati hanno valori costanti durante tutto il progetto. Essi sono quelli relativi alle comorbilità, ai fattori di rischio e alle altre condizioni di rischio. Essi sono di sequito specificati:

- **comorbilità:**
  - malattie polmonari;
  - malattie cardiache;
  - malattie renali;
  - malattie sistema immunitario;
  - malattie oncologiche;
  - malattie metaboliche.
- **fattori rischio:**
  - ipertensione arteriosa sistemica;
  - diabete mellito tipo 1;
  - diabete mellito tipo 2;
  - abitudine tabagica;
  - ipertrigliceridemia;
  - sindrome metabolica.
- **altre condizioni di rischio:**
  - gravidanza;
  - isolamento sociale;
  - non autosufficienza;
  - operatore sanitario;
  - RSA/lungodegente;
  - comunità chiuse.

I valori di questi parametri sono registrati direttamente a partire dalla risposta al questionario; tuttavia, molti dei pazienti o dei medici che inserivano le informazioni ad ogni rilevazione lasciavano il questionario non del tutto compilato immaginando che di default, il sistema assegnasse lo stesso valore, invariato dalla misurazione precedente, ai parametri costanti nel tempo.

Un esempio tra i tanti parametri è Family_Coronary_Artery_Disease ovvero "Storia familiare di malattia coronarica". Ovviamente la storia del paziente resta

invariata ad ogni misurazione; quindi, se per il primo inserimento la variabile era posta al valore SI, anche per tutte le altre rilevazioni il valore doveva ripetersi pur non avendo un inserimento manuale nel dataset.

Per ovviare, quindi, a parte del problema dei campi vuoti, sono state isolate le colonne relative a fattori di rischio, comorbilità e altre condizioni di rischio, e si è effettuato un riempimento automatico dei valori.

Grazie all'aggregazione dei dati in base all'identificativo PATIENTE_ID, sono state considerate le medie dei valori di ogni singola colonna in questione; essendo i parametri di tipo dicotomico e invariati nel tempo, la media era, per ogni campo, o 1 o 0.

Se il paziente o il medico aveva inserito almeno un valore nella colonna rispettiva ai parametri sopracitati, esso è stato prelevato e ripetuto per tutte le rilevazioni del paziente stesso. Nella Tabella 4.3 è riportato l'aggiornamento della colonna relativa al parametro riguardante la storia familiare di malattia coronarica. Nella parte superiore è presente parte del dataset iniziale; nella parte sottostante quello aggiornato con il riempimento automatico dei valori.

PATIENT_ID	**RECEIPT_ TIMESTAMP**	**Family_Coronary_ Artery_Disease**
22739	25/03/21 18:44	
22739	27/03/21 19:07	
22739	26/03/21 07:02	
22739	28/03/21 16:44	
22739	27/03/21 06:55	
22739	25/03/21 15:04	1
22739	25/03/21 19:39	
22739	...	

PATIENT_ID	**RECEIPT_ TIMESTAMP**	**Family_Coronary_ Artery_Disease**
22739	25/03/21 18:44	1
22739	27/03/21 19:07	1
22739	26/03/21 07:02	1
22739	28/03/21 16:44	1
22739	27/03/21 06:55	1
22739	25/03/21 15:04	1
22739	25/03/21 19:39	1
22739	...	1

Tabella 4.3. Aggiornamento dei valori del parametro *Family_Coronary_Artery_Disease*

Nel caso in esame, il paziente 22739 ha risposto al questionario riguardante la storia familiare in data 25/03/21 alle 15:04, ovvero durante la prima rilevazione. Nelle misurazioni successive questo campo è, però, vuoto; di conseguenza, egli non ha più inserito l'informazione.

Attraverso l'utilizzo di semplice un codice, l'inserimento dei valori nei campi vuoti è stato ottimizzato in modo automatico considerando la prima misurazione e riportandola nei campi successivi temporalmente.

Viene, quindi, aggiornata anche la tabella riguardante la percentuale dei valori utilizzabili di tutti i parametri rispetto ad ogni paziente.

Osservando la Tabella 4.4, è possibile notare che non ci sono colori intermedi tra il rosso e il verde; se nel dataset originale era presente almeno un valore tra 0 e 1 nelle colonne interessate, allora tutti i valori dei campi relativi allo stesso paziente e allo steso parametro venivano riempiti. Se nessuna delle misurazioni era inserita, non si ha riempimento e il campo resta al valore NaN; quindi, come nella tabella precendente, la percentuale resta allo 0%

ID_paziente	num_ misurazioni	Hypertension	Diabetes	BMI_30	Dyslipidemia	Cardiac_ disease	Kidney_disease_ non_dialisi	Liver_ disease
16861	1	100.0	100.0	0.0	100.0	100.0	100.0	100.0
16982	11	100.0	100.0	0.0	100.0	100.0	100.0	100.0
20329	7	100.0	100.0	0.0	100.0	100.0	100.0	100.0
22729	121	100.0	100.0	0.0	100.0	100.0	100.0	100.0
22739	112	100.0	100.0	0.0	100.0	100.0	100.0	0.0
22740	118	100.0	100.0	0.0	100.0	100.0	100.0	100.0
22814	72	100.0	100.0	0.0	100.0	0.0	0.0	0.0
22833	2	100.0	100.0	0.0	100.0	100.0	100.0	100.0
22880	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22881	58	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22882	110	100.0	100.0	0.0	100.0	100.0	100.0	100.0
22883	150	100.0	100.0	0.0	100.0	100.0	100.0	100.0
22884	79	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tabella 4.4. Aggiornamento della percentuale dei valori utilizzabili di tutti i parametri rispetto ad ogni paziente

4.2.1 Problematica assegnazione dei valori

Durante la creazione del codice relativo al riempimento dei valori NaN è emersa una problematica legata all'inserimento dei dati nel dataset.

Lavorare con dati biomedici reali porta con sé la possibilità di imbattersi, infatti, nell'errore umano che, se non individuato e trattato nel modo più opportuno, rende le considerazioni e i risultati non più veritieri.

Nella prima fase di scrittura del codice, è stata utilizzata la funzione *GroupBy* offerta da Pandas, in cui il parametro secondo cui raggruppare tutte le variabili era PATIENT_ID. Essa è stata combinata alla funzione media: per ogni gruppo

di valori, aggregati per ogni paziente, era calcolata la media e riportata nel campo relativo.

Viene, quindi, creato un nuovo DataFrame, di cui una parte viene riportata in Tabella 4.5.

PATIENT_ID	Hypertension	Diabetes	Smoking	BMI_30	Dyslipidemia	Metabolic_ Syndrome	Family_Coronary_ Artery_Disease
23080	0	0	0	0	0	0	0
16982	1	0	0	0	1	0	0
22729	0	0	1	0	0	1	0
23193	0	1	0	1	0,09	0	0

Tabella 4.5. DataFrame in cui sono presenti le medie dei valori di tutti i parametri in esame per ogni paziente.

Ogni riga corrisponde ad un paziente, mentre i campi delle colonne equivalgono alla media dei valori di quel parametro.

Come è evidente per il paziente 23193, la media dei valori del parametro *Metabolic_Syndrome* è un numero non intero.

Si è, quindi, creato un focus su questo paziente e, in particolare, relativo a questo parametro. In Tabella 4.6 è presente un DataFrame creato appositamente per potersi concentrare al meglio sulla problematica.

Sono stati, infatti, riportati dal dataset originale le colonne riguardanti l'istante in cui è stata effettuata la misurazione e i relativi valori della variabile in esame. Si noti che il paziente ha effettuato 25 misurazioni, due delle quali riportano il valore 1, mentre le altre hanno il valore 0. Sono presenti anche tre campi con valore NaN.

Il codice di riempimento automatico, quindi, in questo caso non funziona: il campo vuoto sarebbe infatti riempito con il valore media, ovvero 0.09 che non è un valore ammissibile per il parametro studiato. In più, avere un numero diverso da 0 o 1 rende il significato del parametro non più verosimile.

Ovviamente, prima di intervenire creando una soluzione al problema, sono stati interpellati i medici coinvolti nel progetto a cui sono stati mostrati casi simili a quello portato in esempio.

Dopo aver discusso della problematica, è emerso che l'erroneo inserimento dei dati o la mancata risposta a questionari, era da imputare al paziente stesso.

Per poter ovviare al problema, i medici hanno proposto di intervenire manualmente sul dataset inserendo il valore reale e da adattare a tutti gli altri, nel campo rispettivo all'ultima misurazione, rispondendo, così, ad un ultimo questionario; infatti in parte questo veniva già svolto nel progetto per la maggior parte dei pazienti: in fase di chiusura, durante il confronto tra medico e paziente, il primo ricontrollava il questionario ma, anziché correggere ogni valore per ogni misurazione, caricava un'ultima rilevazione in cui l'inserimento dei valori costanti nel tempo era certamente veritiera.

Per questo, è stato creato un nuovo codice in cui, dopo aver ordinato per data i valori di ogni paziente, non veniva isolata la media di essi, ma il campo rispettivo all'ultima misurazione svolta; esso veniva, poi, riportato in tutti gli altri campi relativi alle rilevazioni precedenti.

RECEIPT_TIMESTAMP	Metabolic_Syndrome
31/05/21 09:53	
22/05/21 15:28	0
21/05/21 22:06	0
21/05/21 11:41	0
20/05/21 14:44	0
19/05/21 21:51	0
19/05/21 09:33	0
18/05/21 22:37	0
18/05/21 14:33	0
17/05/21 19:43	0
17/05/21 19:42	0
17/05/21 13:27	0
17/05/21 09:31	1
16/05/21 21:39	0
16/05/21 06:46	0
14/05/21 18:39	
14/05/21 18:37	0
14/05/21 17:15	
14/05/21 15:31	0
14/05/21 14:56	0
14/05/21 14:53	0
09/06/21 21:44	1
09/06/21 05:54	0
08/06/21 20:43	0
08/06/21 08:35	0

Tabella 4.6. Focus per il paziente 23193 rispetto alla problematica di diversi valori sul parametro Metabolic_Syndrome

Infatti, come è evidenziato in rosa, nel caso in Tabella 4.7, in cui le rilevazioni sono in ordine crescente per data, il paziente ha come ultima misurazione in data 09/06/21 21:44, il valore 1.

In Tabella 4.8 è riportato l'aggiornamento del focus del paziente: tutti i valori dei campi relativi al parametro *Metabolic_Syndrome* sono posti ad 1.

Di conseguenza, dopo aver stabilito una nuova logica di inserimento dei valori nel campi vuoti, e dopo aver ricevuto un dataset aggiornato di ulteriori rilevazioni, i campi relativi alle colonne di comorbilità, fattori di rischio e altri fattori di rischio, non mostrano anomalie e sono utilizzabili per le analisi successive. In Tabella 4.9 è riportata la Tabella 4.5 aggiornata del valore inserito con l'ultima considerazione evidenziato in azzurro.

RECEIPT_TIMESTAMP	Metabolic_Syndrome
14/05/21 14:53	0
14/05/21 14:56	0
14/05/21 15:31	0
14/05/21 17:15	
14/05/21 18:37	0
14/05/21 18:39	
16/05/21 06:46	0
16/05/21 21:39	0
17/05/21 09:31	1
17/05/21 13:27	0
17/05/21 19:42	0
17/05/21 19:43	0
18/05/21 14:33	0
18/05/21 22:37	0
19/05/21 09:33	0
19/05/21 21:51	0
20/05/21 14:44	0
21/05/21 11:41	0
21/05/21 22:06	0
22/05/21 15:28	0
31/05/21 09:53	
08/06/21 08:35	0
08/06/21 20:43	0
09/06/21 05:54	0
09/06/21 21:44	1

Tabella 4.7. Focus per il paziente 23193 rispetto alla problematica di diversi valori sul parametro Metabolic_Syndrome, ordinati per data

PATIENT_ID	Hypertension	Diabetes	Smoking	BMI_30	Dyslipidemia	Metabolic_ Syndrome	Family_Coronary_ Artery_Disease
23080	0	0	0	0	0	0	0
16982	1	0	0	0	1	0	0
22729	0	0	1	0	0	1	0
23193	0	1	0	1	1	0	0

Tabella 4.9. Aggiornamento del DataFrame in cui sono presenti le medie dei valori di tutti i parametri in esame per ogni paziente, dopo aver cambiato la logica di assegnazione

RECEIPT_TIMESTAMP	Metabolic_Syndrome
14/05/21 14:53	
14/05/21 14:56	1
14/05/21 15:31	1
14/05/21 17:15	1
14/05/21 18:37	1
14/05/21 18:39	1
16/05/21 06:46	1
16/05/21 21:39	1
17/05/21 09:31	1
17/05/21 13:27	1
17/05/21 19:42	1
17/05/21 19:43	1
18/05/21 14:33	1
18/05/21 22:37	1
19/05/21 09:33	1
19/05/21 21:51	1
20/05/21 14:44	1
21/05/21 11:41	1
21/05/21 22:06	1
22/05/21 15:28	1
31/05/21 09:53	1
08/06/21 08:35	1
08/06/21 20:43	1
09/06/21 05:54	1
09/06/21 21:44	1

Tabella 4.8. Aggiornamento del focus del paziente 23193 rispetto alla problematica di diversi valori sul parametro Metabolic_Syndrome

5

Pre-processing dei dati

In questo capitolo sarà illustrato il pre-processing dei dati ovvero tutte quelle attività volte a preparare il dataset in vista delle analisi successive. Saranno quindi mostrati le variabili di interesse, il ragionamento che ha portato alla loro scelta, le problematiche e le tecniche di risoluzione ad esse associate.

5.1 Introduzione

Dopo aver acquisito i dati finali e aver applicato le tecniche di ETL, il dataset su cui poter lavorare era di 7368 righe e 82 colonne.

Avere un dataset ordinato e ben assortito riguardante le informazioni di ogni paziente prima, durante e dopo l'insorgenza dei sintomi da Covid-19, ha aperto a differenti idee da cui estrarre conoscenze. Di conseguenza, grazie alla collaborazione di tutto il team, sono state vagliate diverse ipotesi per poter portare avanti studi interessanti e di ricerca.

Dopo aver considerato diverse proposte, in accordo con i medici coinvolti, è stato deciso di creare diversi gruppi di pazienti sulla base di parametri invariati nel tempo che caratterizzano il paziente prima di ammalarsi di Covid-19, per poi soffermarsi sul decorso della malattia di ogni gruppo.

In questo modo, un nuovo ipotetico paziente, una volta esaminati i parametri coinvolti, potrà essere assegnato ad un determinato gruppo e si potrà presumibilmente prevedere come il virus influirà sugli altri parametri presenti nel dataset in modo tale da porre più attenzione al loro monitoraggio.

5.2 Scelta delle variabili

In comune accordo con i medici, gli infermieri e i collaboratori delle aziende coinvolte, sono state scelte le variabili su cui poter svolgere il raggruppamento. Esse sono le seguenti:

- **età del paziente:**
  - superiore a 65 anni;

  - inferiore a 65 anni.
- **comorbilità:**
  - malattie polmonari;
  - malattie cardiache;
  - malattie renali;
  - malattie del sistema immunitario;
  - malattie oncologiche;
  - malattie metaboliche.
- **fattori rischio:**
  - ipertensione arteriosa sistemica;
  - diabete di mellito tipo 1;
  - diabete di mellito tipo 2;
  - abitudine tabagica;
  - ipertrigliceridemia;
  - sindrome metabolica.
- **altre condizioni di rischio:**
  - gravidanza;
  - isolamento sociale;
  - non autosufficienza;
  - operatore sanitario;
  - RSA/lungodegente;
  - comunità chiuse.

Prima di applicare gli algoritmi di clustering sono state studiate le variabili interessate.

5.3 Preparazione del dataset

5.3.1 Matrice di correlazione

In prima istanza, si è creata una matrice di correlazione. Essa è una tabella che riporta al suo interno gli indici di correlazione tra due o più variabili.

Graficare la matrice di correlazione è immediato grazie all'utilizzo della funzione *heatmap()* della libreria Seborn. È stato scelto l'indice di correlazione di Pearson che esprime l'eventuale relazione di linearità tra due variabili statistiche; esso ha un valore compreso tra -1 e 1, dove +1 corrisponde alla perfetta correlazione lineare positiva, 0 corrisponde a un'assenza di correlazione lineare e -1 corrisponde alla perfetta correlazione lineare negativa; esso è definito come la covarianza delle due variabili interessate divisa per il prodotto delle deviazioni standard di esse.

Nella Figura 5.1 è presente la matrice di correlazione in cui sono state riportate tutte le variabili utilizzabili come feature dell'algoritmo di clustering.

Si nota immediatamente che le variabili *Diabetes_type1*, *Pregnancy*, *RSA* e *Closed_community* non vengono analizzate dalla funzione *heatmap()*; i valori di queste variabili restano settati sempre o a 1, o a 0. Per tale ragione, non viene considerata la loro correlazione con tutte le altre variabili che, invece, hanno un andamento differente per le diverse misurazioni e quindi per i diversi record del dataset.

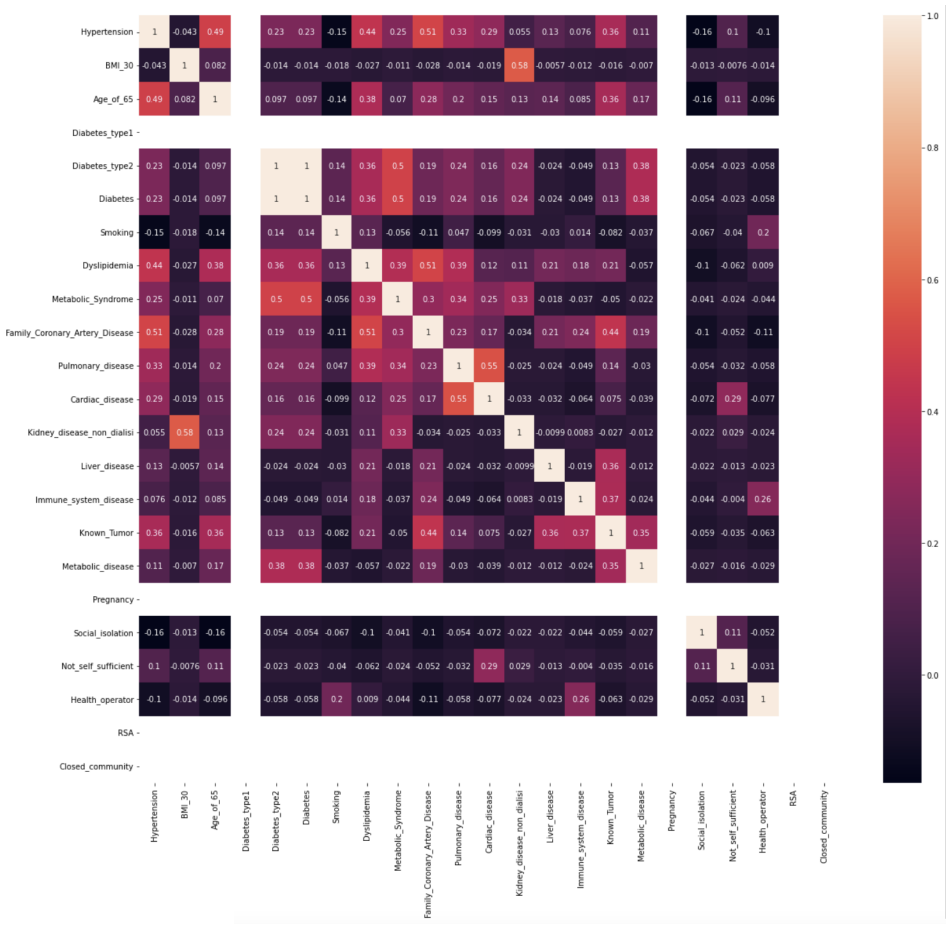


Figura 5.1. Matrice di correlazione

Altra considerazione importante su cui porre l'attenzione riguarda la correlazione lineare perfetta tra i parametri *Diabetes_type2* e *Diabetes* il cui indice di Pearson è, infatti, 1; questo risultato è in linea con il significato medico delle variabili: se un paziente è affetto da diabete di tipo 1 avrà sicuramente posto a 1 anche il parametro riguardante il diabete.

Risulta, quindi, inutile inserire entrambi i parametri negli algoritmi di clustering; viene scelto il parametro Diabetes che include anche il diabete di tipo 1.

5.3.2 Principal Component Analysis

Nei problemi di clustering e classificazione ci sono spesso troppi fattori sulla base dei quali viene sviluppato l'algoritmo di apprendimento automatico.

Questi fattori sono variabili chiamate feature ovvero caratteristiche. Maggiore è il numero di queste variabili e più diventa difficile visualizzare il set di allenamento.

Inoltre, la grande dimensionalità del problema potrebbe portare alla creazione di clustering del tutto sbilanciati.

Le feature da utilizzare per poter sviluppare l'algoritmo di clustering nello studio in esame, dopo l'ultimo passaggio appena spiegato, sono 18, e, in particolare, sono le seguenti:

- hypertension,
- age_of_65',
- diabetes,
- smoking,
- BMI_30,
- dyslipidemia,
- metabolic_Syndrome,
- family_Coronary_Artery_Disease,
- pulmonary_disease,
- cardiac_disease,
- kidney_disease_non_dialisi,
- liver_disease',
- immune_system_disease,
- known_Tumor,
- metabolic_disease,
- social_isolation,
- not_self_sufficient,
- health_operator.

Per poter confrontare diversi risultati e arrivare a scegliere il miglior algoritmo di raggruppamento, si è eseguita una Principal Component Analysis (PCA). Essa ha lo scopo di ridurre la dimensionalità del problema; è un processo, quindi, di riduzione del numero di variabili causali in esame, per cui si ottiene un insieme di variabili principali che semplificano l'analisi del problema di studio.

Più nel dettaglio, la PCA è un metodo che rientra nei problemi di trasformazione lineare e permette di trovare le direzioni della massima varianza dei dati ad alta dimensione e proiettarle su un nuovo sottospazio con dimensioni inferiori a quello originale. Le componenti principali sono le combinazioni lineari delle variabili originali, il vettore dei pesi in questa combinazione è in realtà l'autovettore trovato che a sua volta soddisfa il principio dei minimi quadrati.

La libreria ScikitLearn permette di implementare la PCA in modo intuitivo e rapido: basta definire il numero di componenti principali da ottenere e, dopo aver fornito in input l'insieme di caratteristiche da voler ridimensionare, si ha in output il vettore che racchiude le componenti principali il cui numero è, ovviamente, inferiore al numero delle variabili date in input.

Oltre a ciò vengono salvati i valori della varianza; per poterli visualizzare in modo da essere compresi maggiormente viene creato un grafico bidimensionale a istogramma, con asse verticale che misura la varianza, e asse orizzontale che rappresenta le componenti principali.

I grafici della varianza rispetto alla PCA a due componenti e quella a tre componenti sono riportati rispettivamente in Figura 5.2 e in Figura 5.3 .

Come si può facilmente notare, la varianza diminuisce man mano che ci si sposta dalla prima alla terza componente, da qui anche la riduzione dell'importanza di esse.

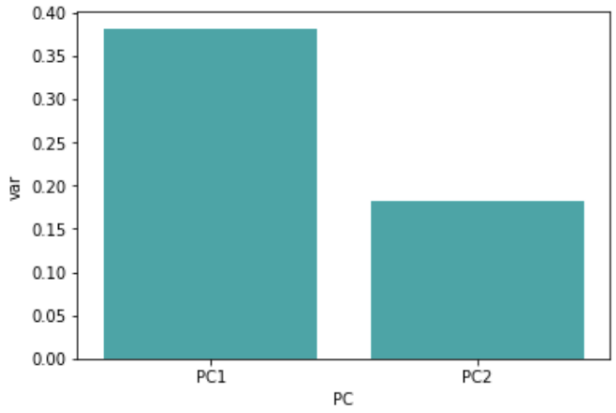


Figura 5.2. Istogramma che rappresenta la varianza delle componenti principali rispetto alla PCA a due componenti

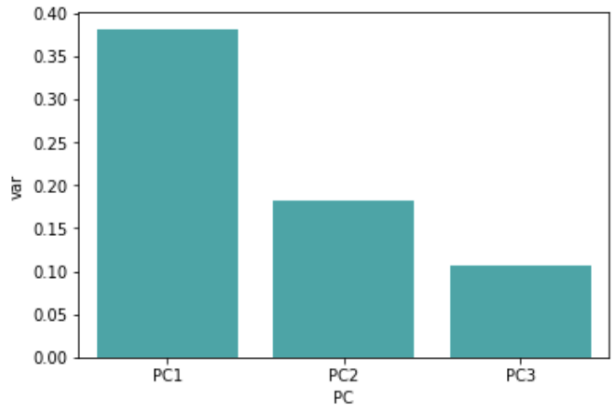


Figura 5.3. Istogramma che rappresenta la varianza delle componenti principali rispetto alla PCA a tre componenti

Nelle Tabelli 5.1-5.3 sottostanti sono riportate le prime righe dei DataFrame da dare in input agli algoritmi di clustering. La Tabella 5.1 è relativa al DataFrame con 18 variabili; la Tabella 5.2 è relativa alla PCA con due componenti principali; infine, la Tabella 5.3 è derivante dalla PCA con tre componenti.

Hypertension	Age_of_65	Diabetes	Smoking	BMI_30	Dyslipidemia	Metabolic_ Syndrome	Family_ Artery_Disease	Coronary_	Pulmonary_ disease	Cardiac_ disease	Kidney_disease_ non_dialisi	Liver_ disease	Immune_ system_disease	Known_ Tumor	Metabolic_ disease	Social_ isolation	Not_self _sufficient	Health _operator
0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	1		0	0	0	0	0	0	0	0	0	0

Tabella 5.1. DataFrame con 18 features

principal component 1	principal component 2
-1,12316	-0,12238
-1,12316	-0,12238
-1,12316	-0,12238
-0,19123	-0,41533
2,48535	-0,75772
2,48535	-0,75772

Tabella 5.2. DataFrame derivante dalla PCA con due componenti principali

principal component 1	principal component 2	principal component 3
-1,12316	-0,12238	-0,08746
-1,12316	-0,12238	-0,08746
-1,12316	-0,12238	-0,08746
-0,19123	-0,41533	-0,15658
2,48535	-0,75772	0,41527
2,48535	-0,75772	0,41527

Tabella 5.3. DataFrame derivante dalla PCA con tre componenti principali

6

Estrazione di conoscenza

Nel capitolo corrente verrà mostrato il fulcro del progetto, ovvero l'estrazione di conoscenza attraverso l'implementazione della tecnica di clustering al fine di raggruppare i pazienti sulla base delle considerazioni effettuate precedentemente; dopo aver illustrato le varie prove svolte approfondendo i limiti e il modo con cui superarli, si mostreranno l'algoritmo scelto, le motivazioni che hanno portato ad esso e i risultati ottenuti

6.1 Clustering

L'obiettivo di clustering è quello di raggruppare oggetti in cluster con un certo grado di omogeneità. Si hanno, infatti, collezioni di oggetti simili rispetto a ciascun oggetto nello stesso cluster e dissimili rispetto agli oggetti in altri cluster.

Il clustering è anche definito come *unsupervised classification*; come per la classificazione, lo scopo è segmentare i dati senza, però, assegnare etichette di classe; di conseguenza, non si hanno classi predefinite ma, ogni cluster può essere interpretato come una classe di oggetti simili.

Dopo aver analizzato e manipolato il dataset, si è posta l'attenzione sulla scelta dell'algoritmo di clustering più adatto allo studio in esame. Le prove svolte prevedono l'utilizzo dei due tra i più famosi algoritmi in letteratura, ovvero K-Means e DBSCAN.

Essi sono entrambi algoritmi partizionali che creano una partizione delle osservazioni minimizzando una certa funzione di costo, quindi determinano il partizionamento dei dati in cluster in modo da ridurre il più possibile la dispersione all'interno del singolo cluster e, contemporaneamente, da aumentare la dispersione tra i cluster.

Entrambi sono stati applicati tre volte; in particolare sono stati dati in input il DataFrame composto da tutte le 18 feature scelte in precedenza, il DataFrame composto dalle due componenti principali derivanti dalla PCA a due componenti e, infine, il DataFrame composto dalle tre componenti principali derivanti dalla PCA a tre componenti. In tutti i tre casi, ovviamente, i DataFrame sono stati ricreati con i rispettivi valori per ogni paziente e non per ogni misurazione, in modo tale da creare gruppi di pazienti patologicamente simili.

6.1.1 Criteri di scelta

Per poter scegliere il miglior raggruppamento si è calcolato l'indice di Silhouette.

Per la generica unità *i*, l'indice di silhouette *s(i)* è un indice che assume valori nell'intervallo [-1,1]. Indicando con *a(i)* la dissimilarità media dell'unità i calcolata rispetto a tutte le altre unità assegnate allo stesso cluster, e con *b(i)* la minima dissimilarità media dell'unità *i* calcolata rispetto a tutti gli altri cluster, si ha la seguente formula:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

La libreria Scikit-learn mette a disposizione la funzione *sklearn.metrics.silhouette_score* che restituisce il coefficiente medio di silhouette su tutti i campioni. Il valore migliore è 1 e il valore peggiore è -1. I valori vicini allo 0 indicano cluster sovrapposti. I valori negativi generalmente indicano che un campione è stato assegnato al cluster sbagliato.

Inoltre, grazie alla collaborazione attiva con l'equipe medica, è sembrato più opportuno creare all'incirca dieci gruppi, in modo tale da poter caratterizzare i pazienti in base a diverse patologie preferendo più cluster meno numerosi rispetto a pochi cluster che integrassero soggetti più eterogenei.

Infine, ovviamente, la scelta del miglior raggruppamento è definita anche in base alla numerosità dei cluster e al significato medico che ognuno di essi può avere. Quest'ultimo obiettivo è stato concordato con i medici coinvolti.

6.2 K-Means

K-Means è un algoritmo di apprendimento non supervisionato che trova un numero fisso di cluster in un insieme di dati.

I cluster rappresentano i gruppi che dividono gli oggetti a seconda della presenza o meno di una certa somiglianza tra di loro, e vengono scelti a priori, prima dell'esecuzione dell'algoritmo. Ognuno di questi cluster raggruppa un particolare insieme di oggetti, che vengono definiti data points. L'insieme dei data points analizzati definisce il set di dati, che rappresenta l'insieme di tutte le istanze analizzate dall'algoritmo.

Quando si utilizza un algoritmo K-Means, per ogni cluster si definisce un centroide, ossia un punto al centro di un cluster. K-Means è un algoritmo iterativo che prevede i seguenti step:

- *inizializzazione*: si definiscono i parametri di input per eseguire l'algoritmo;
- *assegnazione del cluster*: ogni data point viene assegnato al cluster (o centroide) più vicino;
- *aggiornamento della posizione del centroide*: si ricalcola il punto esatto del centroide e di conseguenza se ne modifica la sua posizione.

Questo algoritmo ha il vantaggio di essere molto veloce in quanto sono richiesti pochi calcoli e, di conseguenza, poco tempo di elaborazione per il calcolo delle distanze tra i data point e i centroidi ad ogni iterazione. D'altra parte, bisogna

selezionare il numero dei gruppi da visualizzare a priori e ciò, per il caso in esame, non è banale.

Per ovviare a quest' ultimo problema, il metodo più semplice ed utilizzato è l'elbow method. Esso consiste nell'iterazione del K-Means per diversi valori di k, ovvero del numero di gruppi. Ogni volta si calcola la somma delle distanze al quadrato tra ogni centroide ed i punti del proprio cluster.

Graficando i valori di k sull'asse orizzontale, e i valori della somma delle distanze al quadrato sull'asse verticale, si ottiene un grafico in cui si isola il punto di "gomito", ovvero il punto in cui la curva tende ad avere un cambiamento evidente rispetto alla sua pendenza. Si seleziona, quindi, il k relativo al punto di gomito e si sceglie come numero di cluster.

6.2.1 K-Means applicato a tutte le variabili

Dopo aver importato K-Means da Sklearn, viene caricato il DataFrame di tutte le variabili visualizzato, in parte, nella sezione precedente dell'elaborato.

In questo caso il grafico del elbow method è riportato in Figura 6.1.

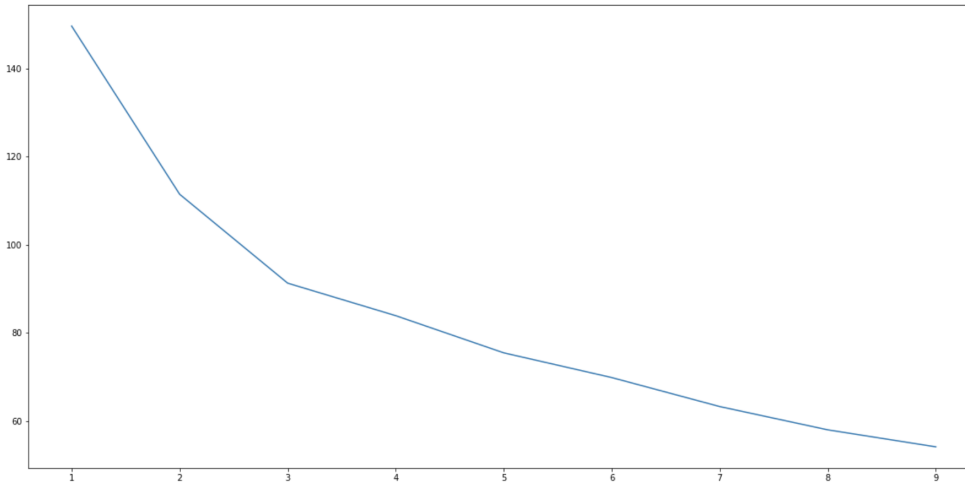


Figura 6.1. Elbow method relativo al caso di tutte le variabili in esame

Come è evidente, il gomito si ha per k=3; le altre variazioni di pendenza sono minime. Di conseguenza, si sceglie di avere tre cluster.

Una volta implementato l'algoritmo, si salva il vettore dei cluster e lo stesso si riassegna al dataset originale in cui i dati sono raggruppati per paziente; dopo si conta la numerosità di ognuno di essi.

Si hanno, quindi, tre diversi gruppi con le seguenti numerosità:

- *cluster 1:* 74 pazienti;
- *cluster 2:* 34 pazienti;
- *cluster 3*: 20 pazienti.

Viene, inoltre, calcolato l'indice di Silhouette che è 0.412.

6.2.2 K-Means applicato al DataFrame relativo alla PCA con due componenti principali

Come nel passaggio precedente, anche nel caso dell'applicazione del K-Means viene considerato l'elbow method per poter scegliere in modo adeguato il numero di cluster da assegnare in input all'algoritmo. In Figura 6.2 è riportato il grafico di cui sopra.

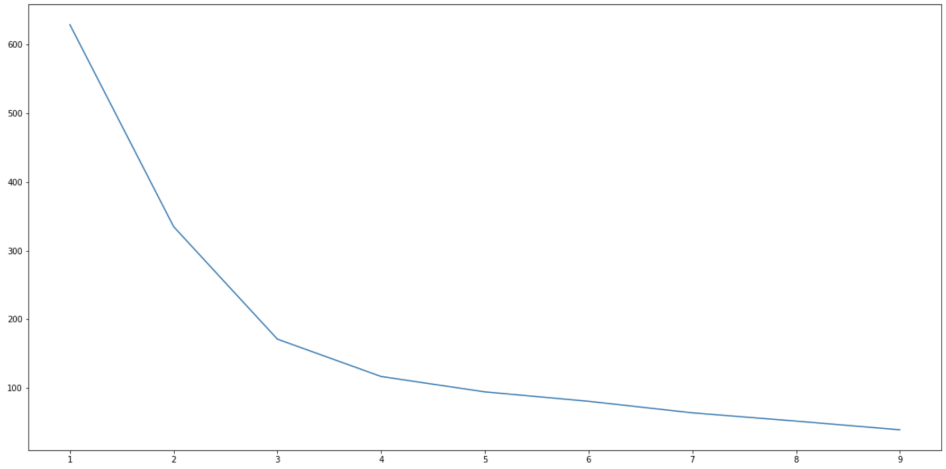


Figura 6.2. Elbow method relativo al caso del DataFrame derivante da PCA con due componenti principali

Il gomito è presente per k=2, k=3 e k=4. Per provare ad ottenere cluster più caratterizzanti, viene scelto il numero più alto di k; si ottengono, quindi, 4 diversi cluster con le seguenti numerosità:

- *cluster 1*: 11 pazienti;
- *cluster 2*: 91 pazienti;
- *cluster 3:* 3 pazienti.
- *cluster 4*: 23 pazienti.

Avendo solo due componenti, è possibile visualizzare il grafico a dispersione dei vari pazienti in base ai cluster; ciò grazie all'utilizzo della libreria *Seaborn* di cui si invoca il metodo *Scatterplot*.

Quest'ultimo è riportato in Figura 6.3 ; dall'esame di tale figura emere in modo evidente lo sbilanciamento dei vari gruppi.

6.2.3 K-Means applicato al DataFrame relativo alla PCA con tre componenti principali

Come nei casi precedenti, viene innanzitutto creato il grafico dell'elbow method in modo da poter scegliere a priori il numero di cluster. Esso è riportato in Figura 6.4.

Il gomito individuato è per k=5, in cui la curva assume una pendenza differente.

Si hanno, di conseguenza, 5 cluster con le seguenti numerosità:

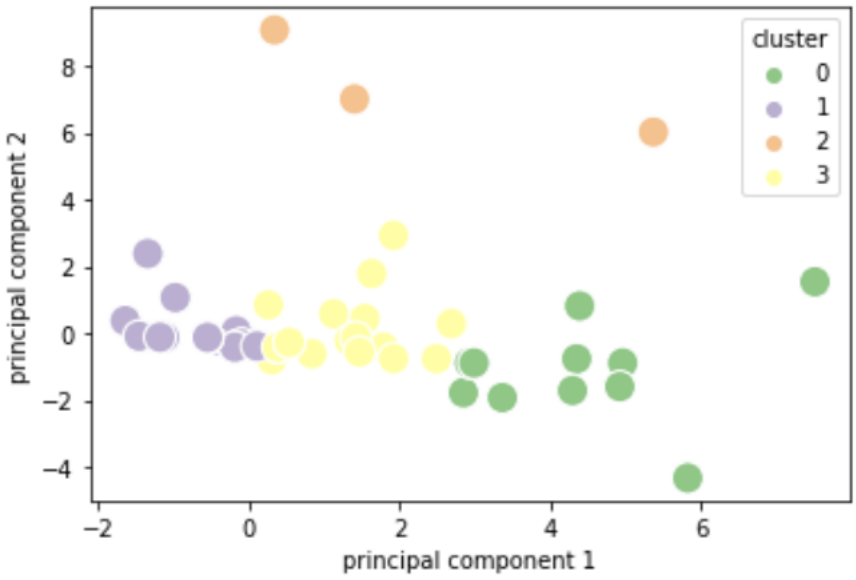


Figura 6.3. Grafico a dispersione rispettivo alla distribuzione spaziale dei pazienti di ogni cluster derivante dall'applicazione dell'algoritmo K-Mens applicato a 2 componenti principali

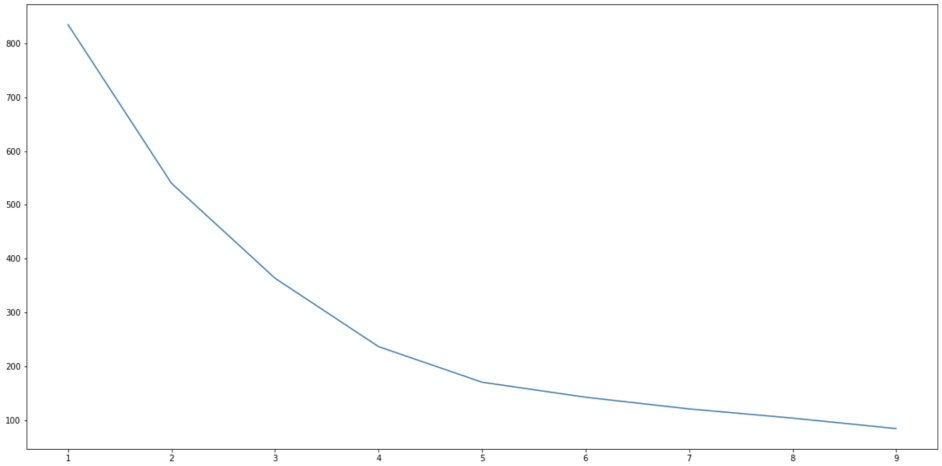


Figura 6.4. Elbow method relativo al caso del DataFrame derivante da PCA con tre componenti principali

- *cluster 1*: 102 pazienti;
- *cluster 2*: 17 pazienti;
- *cluster 3*: 5 pazienti;
- *cluster 4:* 2 pazienti;
- *cluster 5:* 2 pazienti.

Risulta, in questo caso, ancor più evidente lo sbilanciamento dei gruppi; il primo, infatti, racchiude la maggior parte dei pazienti. Viene meno, quindi, la possibilità di caratterizzarli in base ai diversi sintomi.

Viene, inoltre, calcolato l'indice di Silhouette che è 0.409.

6.3 Density-Based Spatial Clustering of Applications with Noise

Il Density-Based Spatial Clustering of Applications with Noise (DBSCAN) è un metodo di clustering proposto nel 1996 da Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu ed è l'algoritmo di clustering più utilizzato e citato in letteratura.

Esso è basato sul concetto di densità: tiene conto, infatti, delle differenze di densità tra le osservazioni nello spazio delle feature. DBSCAN, oltre a clusterizzare campioni vicini tra loro, è in grado anche di individuare eventuali outlier, cioè punti che si discostano particolarmente dalle altre osservazioni, identificati come rumore o noise.

L'algoritmo DBSCAN accetta 2 parametri: epsilon (eps), che è il raggio dei punti centrali, e il numero minimo di punti dati nel cluster (minPts).

Anch'esso, come K-Means, è un algoritmo iterativo: si comincia con un punto casuale che non è stato ancora visitato; viene calcolato il suo eps-vicinato e se contiene un numero sufficiente di punti, viene creato un nuovo cluster. Se ciò non avviene, il punto viene etichettato come rumore; tuttavia, successivamente, potrebbe essere ritrovato in un eps-vicinato sufficientemente grande, riconducibile ad un punto differente entrando a far parte di un cluster.

Se un punto è associato ad un cluster anche i punti del suo eps-vicinato sono parte del cluster. Conseguentemente, tutti i punti trovati all'interno del suo eps-vicinato sono aggiunti al cluster, così come i loro eps-vicinati.

Questo processo continua fino a quando il cluster viene completato; il ciclo termina quando tutti i punti sono stati visitati.

DBSCAN presenta i seguenti vantaggi:

- non richiede di conoscere il numero di cluster a priori, al contrario dell'algoritmo K-Means;
- può trovare cluster di forme arbitrarie;
- possiede la nozione di rumore.

D'altra parte, la qualità del clustering dipende dalla sua misura della distanza che è riconducibile alla scelta della tipologia della distanza stessa, che, generalmente, è quella euclidea. Inoltre, tale algoritmo non è in grado di classificare insiemi di dati

con grande differenze nella densità, dato che la combinazione minPts-epsilon non può essere scelta in modo appropriato per tutti i cluster.

Tuttavia, è possibile ragionare in modo da individuare la migliore combinazione degli iperparametri per il problema in esame. Ciò è possibile grazie all'utilizzo del *grid-Search* che permette di visualizzare una heatmap dalla quale, in base a valori di epsilon e numero di componenti di ogni cluster, si individua il numero ottimale di gruppi.

Nello studio in esame si è posta, anche, l'attenzione sul numero di gruppi suggerito dall'equipe medica, che si aggirava intorno a 10.

6.3.1 DBSCAN applicato a tutte le variabili

La prima prova riguarda l'applicazione DBSCAN a tutte le feature individuate precedentemente.

Dopo aver impostato i parametri grazie al *grid-Search*, l'implementazione dell'algoritmo risulta molto semplice; bisogna istanziare il modello tramite *Sklearn*, impostare i parametri eps e minPts e, infine, utilizzare la funzione *.fit()* per avviare il fitting. In output si ottiene un vettore che mostra le etichette dei cluster identificati. Tra questi è presente anche la label -1, che corrisponde ai punti di rumore.

GridSearch , per questa prima prova, restituisce la heatmap in Figura 6.5

Sull'asse delle x è presente il numero minimo di componenti minimo per ogni cluster; mentre, sull'asse delle y è riportata la distanza minima tra ogni componente facente parte dello stesso gruppo; all'interno di ogni campo è presente il numero di cluster derivante dalla combinazione di epsilon e minPts.

In questo caso, scegliendo un numero di cluster pari a 11, in linea con i suggerimenti dei medici, si avranno eps=0.5 e minPts=2.

Si ottengono, di conseguenza, 10 cluster con le seguenti numerosità:

- *cluster 1:* 48 pazienti;
- *cluster 2*: 6 pazienti;
- *cluster 3*: 2 pazienti;
- *cluster 4*: 19 pazienti;
- *cluster 5*: 5 pazienti.
- *cluster 6*: 2 pazienti;
- *cluster 7:* 2 pazienti;
- *cluster 8*: 3 pazienti;
- *cluster 9*: 5 pazienti;
- *cluster 10*: 2 pazienti.

Il numero di noise point è pari a 34.

6.3.2 DBSCAN applicato al DataFrame relativo alla PCA con due componenti principali

La seconda prova riguarda l'applicazione di DBSCAN al DataFrame relativo alla PCA con due componenti principali. *GridSearch* restituisce la heatmap in Figura 6.6

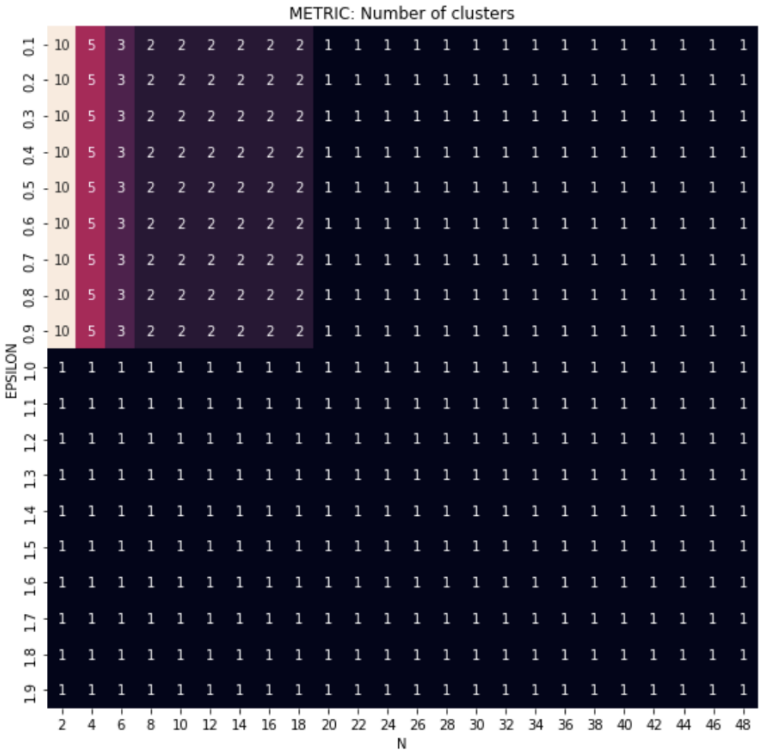


Figura 6.5. Heatmap derivante da *grid-Search* per ottimizzare la scelta degli iperparametri di DBSCAN applicato a tutte le feature in esame

In questo caso, scegliendo un numero di cluster pari a 11, in linea con i suggerimenti dei medici, si avranno eps=0.2 e minPts=2.

Si ottengono, di conseguenza, 11 cluster con le seguenti numerosità:

- *cluster 1*: 53 pazienti;
- *cluster 2*: 10 pazienti;
- *cluster 3*: 2 pazienti;
- *cluster 4*: 19 pazienti;
- *cluster 5*: 2 pazienti.
- *cluster 6*: 3 pazienti;
- *cluster 7*: 2 pazienti;
- *cluster 8*: 5 pazienti;
- *cluster 9:* 2 pazienti;
- *cluster 10*: 2 pazienti.
- *cluster 11*: 2 pazienti.

Il numero di noise point, invece, è pari a 26.

Studiando soltanto due componenti, è possibile visualizzare il grafico a dispersione dei vari pazienti in base ai cluster. Quest'ultimo è riportato in Figura 6.7.

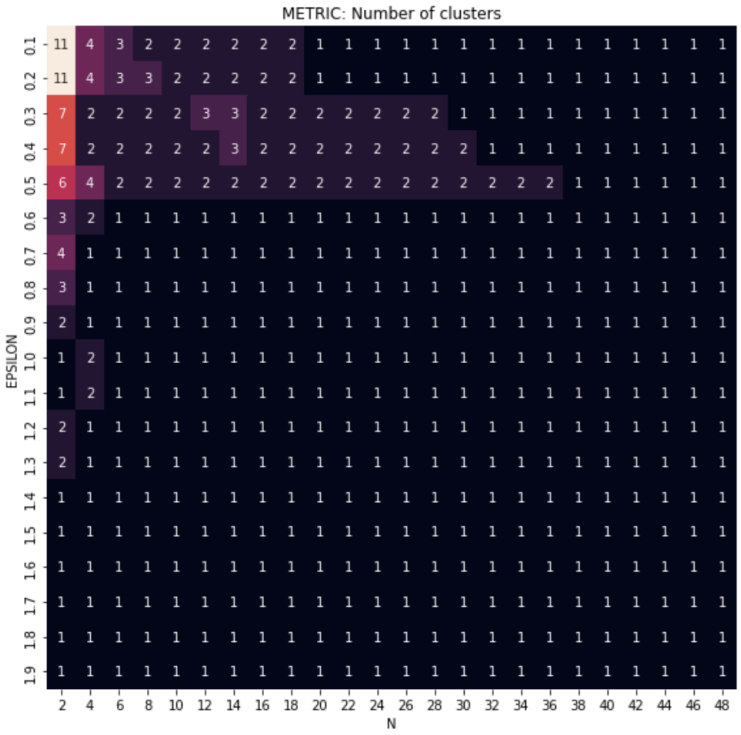


Figura 6.6. Heatmap derivante dal *grid-Search* per ottimizzare la scelta degli iperparametri del DBSCAN applicato al DataFrame relativo alla PCA con due componenti principali

Ovviamente, essendo molti i cluster, non è molto evidente la separazione di ogni gruppo. Da notare la disposizione dei noise point, rappresentati nel grafico tramite punti in nero.

Essi potrannno, in futuro, essere studiati considerando un'analisi degli outlier.

6.3.3 DBSCAN applicato al DataFrame relativo alla PCA con tre componenti principali

La terza e ultima prova riguarda l'applicazione di DBSCAN al DataFrame relativo alla PCA con tre componenti principali.

Il gridSearch restituisce la heatmap in Figura 6.8

In questo caso, scegliendo un numero di cluster pari a 10, si avranno eps=0.4 e minPts=2.

Si ottengono, di conseguenza, 10 cluster con le seguenti numerosità:

- *cluster 1*: 53 pazienti;
- *cluster 2:* 27 pazienti;
- *cluster 3:* 2 pazienti;
- *cluster 4:* 2 pazienti;
- *cluster 5*: 2 pazienti.

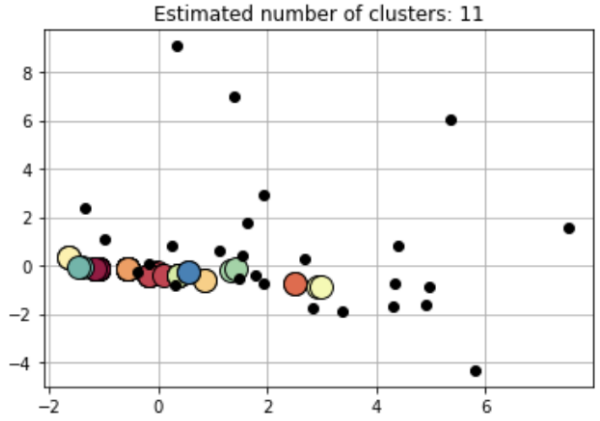


Figura 6.7. Grafico a dispersione relativo alla distribuzione spaziale dei pazienti di ogni cluster derivante dall'applicazione dell'algoritmo DBASCAM applicato a 2 componenti principali

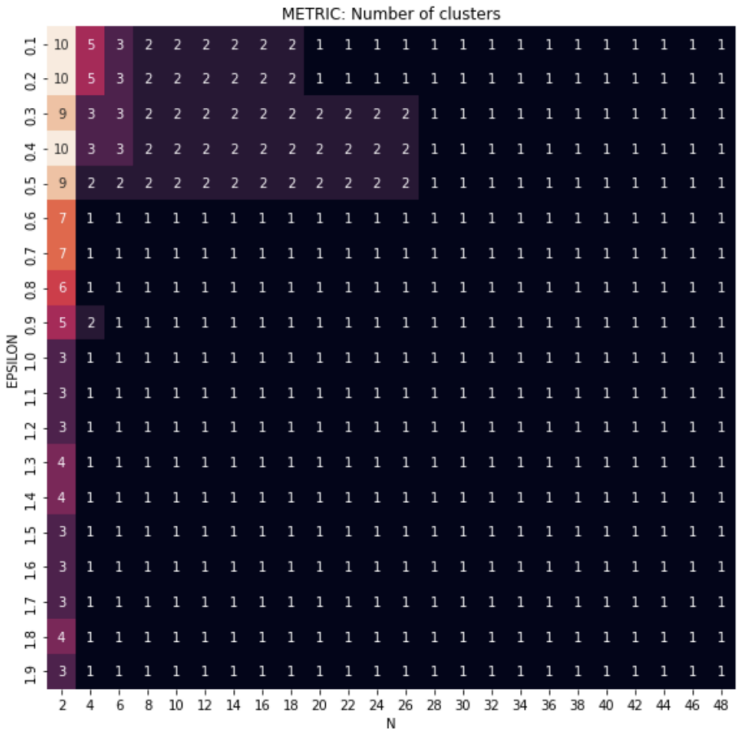


Figura 6.8. Heatmap derivante dal *grid-Search* per ottimizzare la scelta degli iperparametri di DBSCAN applicato al DataFrame relativo alla PCA con tre componenti principali

- *cluster 6*: 3 pazienti;
- *cluster 7*: 2 pazienti;
- *cluster 8*: 7 pazienti;
- *cluster 9*: 2 pazienti;
- *cluster 10*: 2 pazienti.

Il numero di noise point, invece, è pari a 26.

6.4 Algoritmo scelto

Per poter scegliere l'algoritmo più adatto al problema, sono stati valutati diversi aspetti; per questo sono state create delle tabelle per poter riassumere i diversi risultati in modo sintetico.

Innanzitutto, si è posta l'attenzione sul numero di cluster: come è possibile notare nella Tabella 6.1, il numero dei gruppi relativi all'algorimo K-Means è al massimo pari a 5. Ciascuno di essi è composto da molti elementi; questo non permette una caratterizzazione puntuale dei pazienti in base alle diverse feature; in altre parole, non consente di svolgere un'analisi rispetto alle comorbilità e ai fattori di rischio.

	K-Means				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
senza PCA	74	34	20		
PCA 2	11	94	3	23	
PCA 3	102	17	5	2	2

Tabella 6.1. Tabella riassuntiva dei risultati dell'algoritmo K-Means

Di conseguenza, il clustering creato utilizzando quest'ultimo algoritmo viene scartato in tutte le tre varianti di DataFrame di input provati.

Ci si è, allora, concentrati sui risultati ottenuti dall'implementazione di DBSCAN riportati in Tabella 6.2.

	DBASCAN											
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	**Noise**
senza PCA	48	6	2	19	5	2	2	3	5	2		**34**
PCA 2	53	10	2	19	2	3	2	5	2	2	2	**26**
PCA 3	53	27	2	2	2	3	2	7	2	2		**26**

Tabella 6.2. Tabella riassuntiva dei risultati dell'algoritmo DBSCAN

In questo caso, il numero di gruppi è stato forzato dalla scelta degli iperparametri; infatti, vengono rispettati i requisiti consigliati dai medici.

Per poter scegliere quale delle tre varianti analizzare in modo più approfondito, è stato considerato il numero dei Noise Point, ovvero il numero di pazienti non

appartenenti a nessuna categoria. Viene, quindi, immediatamente scartato il caso di DBSCAN senza l'utilizzo della PCA.

A questo punto, si è passati a valutare il "significato medico" di ogni cluster analizzandoli uno ad uno e ponendo l'attenzione sui sintomi simili riportati.

Ovviamente, per poter esaminare quest'ultimo aspetto, il contributo dei medici si è rivelato fondamentale.

Quindi, in uno degli incontri multidisciplinari, sono stati mostrati i risultati sotto forma di tabelle e istogrammi che potessero facilitare e velocizzare la comprensione del problema e, in accordo con tutta l'equipe, è stato scelto il clustering ricavato dall'implementazione di DBSCAN applicato al DataFrame derivante dall'applicazione della PCA a due componenti. In Figura 6.9 sono mostrati tutti gli istogrammi del caso scelto: per ogni cluster vengono presentate le percentuali di pazienti con una determinata comorbilità o che presentano i fattori di rischio esaminati.

In Figura 6.10 è riportata la tabella relativa ai pazienti non facenti parte di nessun gruppo e che, quindi, hanno label -1 nel vettore risultante dall'implementazione del clustering.

Da questi risultati è stato possibile intensificare lo studio approfondendo il significato medico di ogni cluster.

Inizialmente, per poter assegnare gli outlier a determinati gruppi e cercare di accorpare i cluster meno numerosi, era stato implementato un algoritmo di clustering di tipo gerarchico che, però, per la copiosità delle feature e il ridotto numero dei pazienti, non ha portato a nessun risultato soddisfacente.

Di conseguenza, avendo pochi pazienti su cui ragionare, l'accorpamento dei gruppi e l'assegnazione di alcuni dei noise point, sono stati svolti manualmente sulla base della caratterizzazione sintomatologica di ogni cluster.

Sono stati, quindi, creati 8 diversi gruppi con le seguenti caratteristiche:

- *gruppo 1*: pazienti senza patologie e senza alcun fattore di rischio con meno di 65 anni;
- *gruppo 2*: pazienti con ipertensione con meno di 65 anni;
- *gruppo 3*: pazienti senza patologie e senza alcun fattore di rischio con più di 65 anni;
- *gruppo 4*: pazienti affetti da malattie polmonari;
- *gruppo 5*: pazienti con ipertensione con più di 65 anni;
- *gruppo 6:* pazienti affetti da malattie cardiache;
- *gruppo 7:* pazienti affetti da dislipidemia;
- *gruppo 8:* pazienti con più di una patologia.

In Tabella 6.3 è riportato il clustering finale, risultato dell'implementazione di DBSCAN applicato al DataFrame derivante dall'applicazione della PCA a due componenti e della manipolazione manuale di questi ultimi sulla base delle considerazioni mediche; per ogni gruppo vengono, anche, mostrati i cluster da cui derivano e gli identificativi dei pazienti che ne fanno parte.

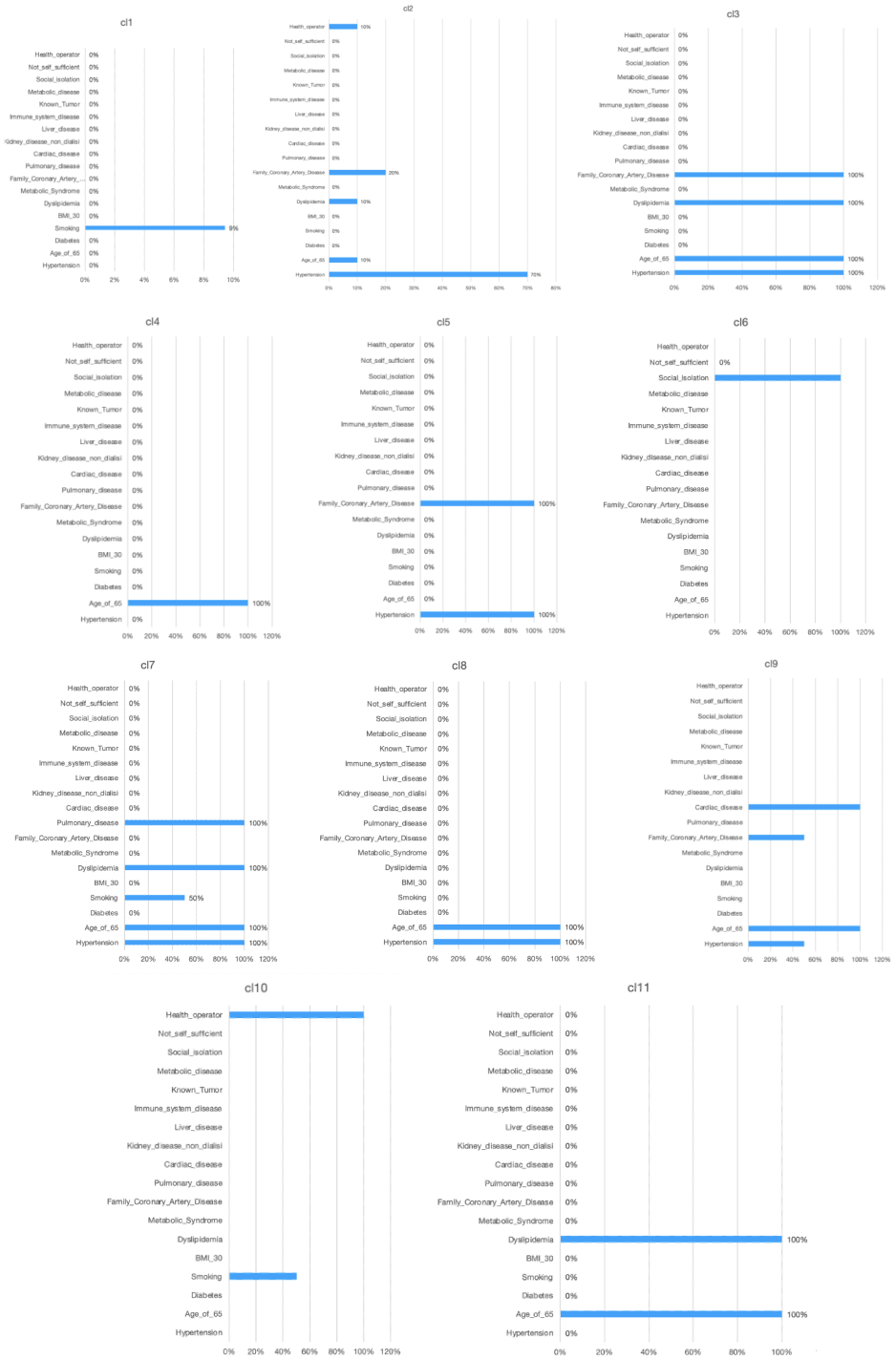
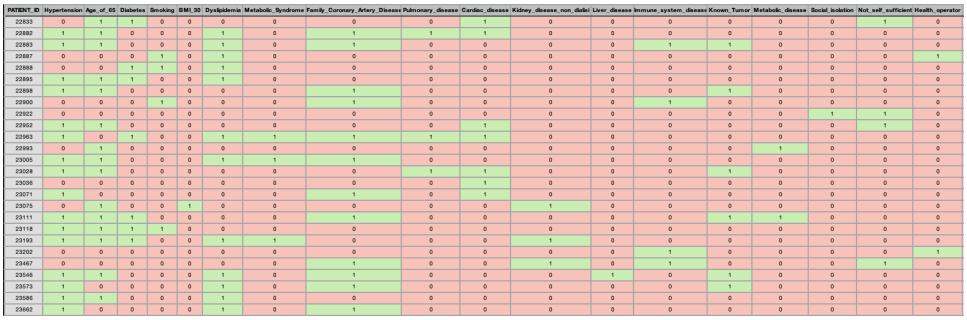


Figura 6.9. Istogrammi relativi a tutti i cluster risultanti dell'implementazione di DBSCAN applicato al DataFrame derivante dall'applicazione della PCA a due componenti



PATIENT_ID	Hypertension	Age_of_65	Diabetes	Smoking	BMI_30	Dyslipidemia	Metabolic_Syndrome	Family_Coronary_Artery_Disease	Pulmonary_disease	Cardiac_disease	Kidney_disease_non_dialisi	Liver_disease	Immune_system_disease	Known_Tumor	Metabolic_disease	Social_isolation	Not_self_sufficient	Health_operator
22833	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
22882	1	1	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0
22883	1	1	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0
22887	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1
22888	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
22895	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
22898	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
22900	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0
22922	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
22952	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
22963	1	0	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0
22993	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
23005	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
23028	1	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0
23036	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
23071	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
23075	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
23111	1	1	1	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0
23118	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23193	1	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0
23202	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
23467	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0
23546	1	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0
23573	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0
23586	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
23662	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0

Figura 6.10. Focus riguardante gli outliers

Gruppo 1	Gruppo 2	Gruppo 3	Gruppo 4	Gruppo 5	Gruppo 6	Gruppo 7	Gruppo 8
Pazienti senza patologie e fattori di rischio con meno di 65 anni	Pazienti ipertesi con meno di 65 anni	Pazienti senza patologie e fattori di rischio con più di 65 anni	Pazienti con problemi polmonari	Pazienti ipertesi con più di 65 anni	Pazienti con problemi cardiaci	Pazienti con dislipidemia	Pazienti con più di una patologia
16861	22729	22814	23019	23057	23108	23876	22833
16982	22942	22880	23060	23059	23575	23877	22882
20329	23020	22881	Cluster 7	23194	22952	22887	22883
22885	23112	22884		23196	Cluster 9	Cluster 11	22888
22897	23244	22886		23576			22895
22918	23253	22896		22739			22898
22937	22939	22906		22740			22900
22943	23866	22964		Cluster 3			22963
22954	Cluster 5	23018					22993
22955		23024					23005
23003		23058					23028
23004		23093					23036
23006		23243	Cluster 4				23075
23007		23587					23111
23008		23888					23118
23011		23889					23193
23038		23890					23202
23039		23991					23467
23040							23546
23054							23573
23055							23586
23056							23662
23070							23407
23072							22940
23073							22941
23080							23114
23081							Noise Points
23094							
23095							
23098							
23099							
23100							
23101							
23113							
23117							
23128							
23132							
23245							
23250							
23251							
23252							
23260							
23408							
23508							
23517							
23585							
23628							
23629							
23630							
23878							
23883							
24003							
22922							
24005							
23010							
23084	Cluster 6						
23089							
23201	Cluster 10						
23203							

Tabella 6.3. Tabella riassuntiva dei gruppi finali dopo la manipolazione dei risultati

7

Discussione dei risultati

In questo capitolo saranno illustrati i risultati delle analisi svolte precedentemente; in particolare, verrà approfondito l'evolversi della sintomatologia da un punto di vista globale basandosi sul raggruppamento effettuato.

7.1 Introduzione

Il lavoro svolto ed illustrato nelle sezioni precedenti aveva lo scopo di creare gruppi di pazienti che hanno caratteristiche comuni riguardanti l'anamnesi di ogni paziente.

Lo step successivo ha permesso di valutare diversi aspetti sull'andamento della malattia derivante da Covid-19.

Alcune delle analisi svolte sono considerazioni generali riguardanti uno specifico aspetto della malattia e non considerano, quindi, la suddivisione in cluster. Altre, invece, ripongono l'attenzione sull'evoluzione dei parametri in seguito all'esito positivo del tampone o all'insorgere dei primi sintomi, classificando i pazienti in base al raggruppamento svolto.

Viene, di seguito, riproposto un riassunto della suddivisione in gruppi dei pazienti insieme alle corrispettive caratteristiche, in modo da poter contestualizzare le analisi successive:

- *gruppo 1*: pazienti senza patologie e senza alcun fattore di rischio con meno di 65 anni;
- *gruppo 2*: pazienti con ipertensione con meno di 65 anni;
- *gruppo 3*: pazienti senza patologie e senza alcun fattore di rischio con più di 65 anni;
- *gruppo 4*: pazienti affetti da malattie polmonari;
- *gruppo 5*: pazienti con ipertensione con più di 65 anni;
- *gruppo 6:* pazienti affetti da malattie cardiache;
- *gruppo 7:* pazienti affetti da dislipidemia;
- *gruppo 8:* pazienti con più di una patologia.

Sulla base della quantità di dati disponibili e su suggerimento dell'equipe medica, vengono considerati, per la maggior parte delle analisi, i seguenti parametri:

- *sintomi minori*: fatica, gola infiammata, mal di testa, dolore muscolare, congestione nasale e perdita dell'olfatto;
- *sintomi maggiori*: tosse e febbre;
- *sintomi allarme*: saturazione dell'ossigeno, fiato corto e coscienza alterata;
- *sintomi e terapia*: sintomi persistenti e risposta alla terapia standard.

7.2 Sintomatologia per gruppi

Nella Tabella 7.1 sono riportati i risultati derivanti da un'analisi svolta riguardante la sintomatologia per gruppi. Quest'ultima è stata creata estrapolando, per ogni gruppo, tutti i pazienti che avevano i parametri considerati posti a 1; per facilitare la comprensione, ogni campo è colorato in base alla percentuale dei valori secondo una scala che prevede il verde per le percentuali più basse, fino al rosso per quelle più alte.

Da notare che per le colonne relative a tosse e fiato corto bisogna considerare le etichette sottostanti; quindi, tutte le percentuali molto alte nel caso di tosse con etichetta 0, stanno a significare che la maggior parte dei pazienti hanno un livello del sintomo minimo.

Considerando la globalità del problema, è evidente che i sintomi che più si ripetono nei pazienti affetti da Covid-19, indipendentemente dall'appartenenza ad uno specifico gruppo, sono la febbre e l'alterazione della saturazione di ossigeno. Inoltre, la totalità dei pazienti risponde alla terapia standard iniziale.

Per quanto riguarda l'analisi per gruppi, invece, è da sottolineare che i pazienti appartenenti al primo gruppo, ovvero coloro che non hanno patologie e fattori di rischio e hanno meno di 65 anni presentano delle percentuali medie della presenza dei sintomi; lo stesso accade per il gruppo 8, composto da pazienti con più di una patologia. Questo, però, è da contestualizzare con la numerosità dei gruppi; essi sono quelli che racchiudono più pazienti.

Risulta interessante, infine, porre l'attenzione sul gruppo 3 e sul gruppo 5 che includono i pazienti con più di 65 anni, nel primo senza patologie e nel secondo con ipertensione; essi presentano un valore della saturazione di ossigeno alterato e febbre per la maggior parte dei casi; tali parametri sono, quindi, da monitorare costantemente.

	Num. Pazienti	Sintomi minori							Sintomi maggiori					Sintomi allarme									Sintomi e terapia	
		Fatica	*Gola Infiammata*	*Mal di testa*	*Dolore muscolare*	*Congestione nasale*	*Perdita dell'olfatto*	*Perdita del gusto*	*Tosse* 0	*Tosse* 1	*Tosse* 2	*Tosse* 3	*Febbre*	*Saturazione Ossigeno* 0	*Saturazione Ossigeno* 1	*Saturazione Ossigeno* 2	*Saturazione Ossigeno* 3	*Coscienza alterata*	*Fiato corto* 0	*Fiato corto* 1	*Fiato corto* 2	*Fiato corto* 3	*Sintomi persistenti*	*Non risponde alla terapia*
Gruppo 1	59	16,95%	5,08%	6,78%	8,47%	6,78%	1,69%	1,69%	88,14%	10,17%	0%	1,69%	83,05%	3,39%	6,78%	45,76%	44,07%	0%	91,53%	5,08%	3,39%	0%	8,47%	0%
Gruppo 2	8	37,50%	0%	12,50%	0%	12,50%	0%	0%	75,00%	25,00%	0%	0%	75,00%	0%	0%	62,50%	37,50%	0%	100%	0%	0%	0%	0%	0%
Gruppo 3	19	15,79%	5,26%	0%	21,05%	0%	0%	0%	89,48%	5,26%	5,26%	0%	57,89%	0%	0%	52,63%	47,37%	0%	100%	0%	0%	0%	5,26%	0%
Gruppo 4	2	50,00%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	50,00%	0%	0%	50,00%	50,00%	0%	0%	50,00%	0%	50,00%	50%	0%
Gruppo 5	7	0%	0%	0%	0%	14,29%	0%	0%	85,71%	14,29%	0%	0%	71,43%	0%	0%	57,14%	42,86%	0%	100%	0%	0%	0%	14,29%	0%
Gruppo 6	3	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	66,67%	33,34%	0%	33,33%	33,33%	0%	100%	0%	0%	0%	0%	0%
Gruppo 7	3	0,00%	0%	0%	0%	0%	0%	0%	66,67%	33,33%	0%	0%	66,67%	0%	0%	66,67%	33,33%	0%	100%	0%	0%	0%	0%	0%
Gruppo 8	27	14,81%	3,70%	3,70%	11,11%	0%	0%	0%	85,19%	11,11%	3,70%	0%	70,37%	0%	0%	40,74%	59,26%	7,41%	88,89%	0%	7,41%	3,70%	7,41%	0%
TOTALE	128	16,41%	3,91%	4,69%	9,38%	4,69%	0,78%	0,78%	85,16%	12,50%	1,56%	0,78%	74,22%	2,34%	3,12%	46,88%	47,66%	1,56%	92,20%	3,12%	3,12%	1,56%	7,81%	0%

Tabella 7.1. Percentuale di pazienti per ogni gruppo che è affetto dai sintomi in esame

7.3 Andamento temporale della sintomatologia

Un'altra analisi svolta, considerando il raggruppamento dei pazienti, si riferisce all'andamento temporale dell'insorgere dei sintomi elencati precedentemente. Di conseguenza, per ogni gruppo viene creata una tabella che distingue giorno per giorno, per la prima settimana di misurazioni, la comparsa dei sintomi. Nella Tabella 7.2 è presente il caso relativo al primo gruppo, in cui è evidente che le percentuali si mantengono all'incirca costanti se non per alcune eccezioni.

I risultati di questo esempio sono in linea con tutti gli altri gruppi rispetto alla comparsa di febbre e alterazione della saturazione di ossigeno nel sangue sin dai primi giorni.

		Sintomi minori							Sintomi maggiori		Sintomi allarme			Sintomi e terapia	
	Giorno	*Fatica*	*Gola Infiammata*	*Mal di testa*	*Dolore muscolare*	*Congestione nasale*	*Perdita dell'olfatto*	*Perdita del gusto*	*Tosse*	*Febbre*	*Saturazione Ossigeno*	*Coscienza alterata*	*Fiato corto*	*Sintomi persistenti*	*Non risponde alla terapia standard iniziale*
Gruppo 1	1	9%	4%	2%	4%	2%	0%	0%	6%	43%	70%	0%	2%	8%	0%
	2	9%	2%	2%	0%	4%	0%	0%	6%	43%	64%	0%	4%	9%	0%
	3	11%	2%	2%	0%	4%	0%	0%	6%	34%	64%	0%	4%	9%	0%
	4	11%	2%	4%	2%	4%	0%	0%	8%	47%	74%	0%	4%	8%	0%
	5	12%	2%	4%	2%	4%	2%	2%	6%	54%	65%	0%	4%	8%	0%
	6	12%	2%	2%	2%	4%	0%	0%	8%	46%	74%	0%	4%	8%	0%
	7	12%	4%	2%	2%	4%	0%	0%	6%	43%	76%	0%	4%	8%	0%

Tabella 7.2. Percentuale di pazienti del primo gruppo che presentano i sintomi per i primi 7 giorni dall'inizio della malattia

Anche in questo caso, come nell'analisi precedente, si ha che i gruppi 1 e 8 sono quelli che presentano la maggior quantità di dati e, di conseguenza, la maggior variabilità di questi nel tempo.

Nei gruppi 2, 4 e 5 sono presenti sin da subito anche tosse e congestione nasale. Il primo sintomo è d'allarme soprattutto per il gruppo 4 che è composto da pazienti con malattie polmonari e che dal primo giorno di misurazione presentano, per la totalità dei casi, fiato corto.

Diversamente dalle aspettative, invece, la maggior parte dei gruppi non presenta perdita del gusto e dell'olfatto durante i primi giorni della malattia.

Infine, è stato evidenziato il fatto che in tutti i gruppi, uno dei primi sintomi a comparire è la variazione di saturazione di ossigeno nel sangue.

7.3.1 Primo sintomo

Dall'ultima osservazione elaborata, è stato approfondito il tema del primo sintomo.

Per questo, sono stati valutati tutti i pazienti e ci si è chiesto quale fosse il primo sintomo a comparire.

In Figura 7.1 è presente un istogramma rappresentativo del problema. Per il 79% dei casi è la variazione di saturazione di ossigeno del sangue il primo parametro registrato, a prescindere dall'appartenenza ai vari gruppi; di conseguenza, esso deve essere costantemente monitorato.

Oltre a questo, in molti casi è la febbre a manifestarsi come primo sintomo; ciò rispetta le aspettative.

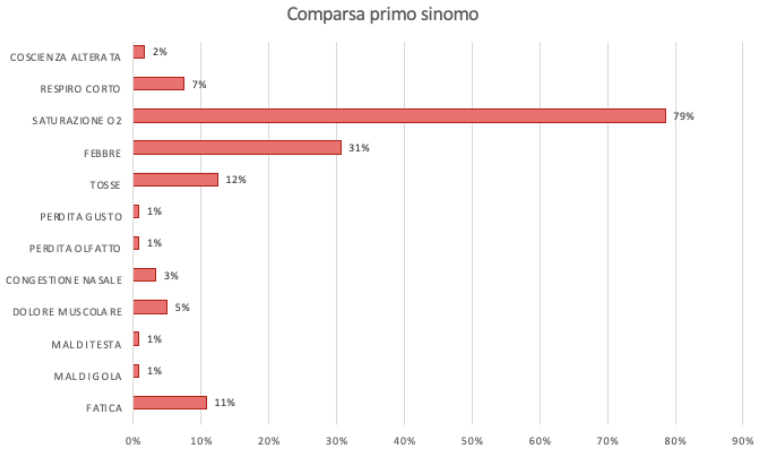


Figura 7.1. Istogramma di comparsa del primo sintomo

7.4 Analisi statistiche generali

Nell'ultimo step del progetto sono state svolte delle analisi statistiche generali.

Nella Tabella 7.3 sono presenti osservazioni inerenti alla durata della malattia, intesa come il numero di giorni trascorsi dal primo tampone positivo, al primo tampone negativo. In generale, i risultati delle analisi statistiche globali sono in linea con le informazioni presenti in letteratura.

Per quanto riguarda, invece, le analisi per gruppi, purtroppo, non per tutti i pazienti si hanno dati riferibili a queste due informazioni; in particolare è stato possibile analizzare questo tema solo per i pazienti facenti parte dei gruppi 1 e 2.

Tendenzialmente, si può affermare che la durata media della malattia è maggiore per pazienti con più di 65 anni; tuttavia, nel campione studiato, le durate minima e massima della malattia sono minori per i pazienti più anziani.

Altra anali svolta è inerente alla comparsa della febbre rispetto all'insorgere degli altri sintomi. Nella Tabella 7.4 sono presenti i risultati ottenuti.

Questo aspetto viene valutato contando i giorni trascorsi tra la prima misurazione e la data in cui il parametro *Body_Temperature* supera i 37°.

Globalmente, la durata media è di sei giorni, quella massima e minima sono, rispettivamente, di 19 e 1 giorno. I risultati rispecchiano ciò che i medici si aspettavano.

Diversamente dall'analisi discussa in precedenza, in cui le informazioni erano mancanti, in questo caso le misurazioni svolte hanno permesso di valutare il problema sia globalmente che per sette gruppi su otto.

Si può, certamente, affermare che i pazienti del gruppo 5, con ipertensione e più di 65 anni, hanno valori nell'analisi che superano di molto quelli globali; questo sta a significare che, in media, la febbre compare prima ed è persistente.

Al contrario, i pazienti senza patologie e senza alcun fattore di rischio con meno di 65 anni, del gruppo 1, hanno valori inferiori a quelli globali; quindi, tendenzialmente, la febbre compare più tardi e dura meno.

DURATA MEDIA (giorni)
20

DURATA MASSIMA (giorni)
38

DURATA MEDIANA (giorni)
19

DURATA MINIMA (giorni)
7

Gruppi	DURATA MEDIA (giorni)
1	21
2	18,67

Gruppi	DURATA MASSIMA (giorni)
1	37
2	38

Gruppi	DURATA MEDIANA (giorni)
1	20
2	10

Gruppi	DURATA MINIMA (giorni)
1	7
2	8

Tabella 7.3. Statistiche globali (a sinistra) e statistiche per gruppi (a destra) riguardanti la durata della malattia

Gli altri gruppi non hanno valori che non si distaccano molto dalla media totale.

L'ultima analisi svolta riguarda, infine, il decorso della malattia per quei pazienti che avevano sintomi persistenti per più di tre giorni, ovvero coloro che presentano la variabile *Persistent_symptoms_3_days* posta a 1.

Essa è stata svolta considerando il numero di misurazioni con lo stesso sintomo per tutti i pazienti con sintomi persistenti.

Questi ultimi, però, sono un campione limitato; di conseguenza, sono state svolte osservazioni globali e non basate sui raggruppamenti.

Globalmente si può dire che, come è emerso dagli altri risultati, la saturazione dell'ossigeno è il sintomo da monitorare maggiormente in quanto persistente.

Per molti pazienti del campione, anche la tosse e il fiato corto hanno una durata maggiore rispetto agli altri sintomi.

DURATA MEDIA (giorni)
6,03

DURATA MASSIMA (giorni)
19

DURATA MEDIANA (giorni)
5

DURATA MINIMA (giorni)
1

Gruppi	DURATA MEDIA (giorni)
1	4,67
2	7,50
3	7,67
5	10,33
7	6,50
8	4,33

Gruppi	DURATA MASSIMA (giorni)
1	14
2	12
3	19
5	16
7	8
8	13

Gruppi	DURATA MEDIANA (giorni)
1	3,5
2	7,5
3	7
5	12
7	6,5
8	3

Gruppi	DURATA MINIMA (giorni)
1	1
2	3
3	2
5	3
7	5
8	1

Tabella 7.4. Statistiche globali (a sinistra) e statistiche per gruppi (a destra) riguardanti i giorni trascorsi tra la prima misurazione e l'insorgere della febbre

		Sintomi minori						Sintomi maggiori		Sintomi allarme		
Paziente	**Gruppo**	*Fatica*	*Gola Infiammata*	*Mal di testa*	*Dolore muscolare*	*Congestione nasale*	*Perdita dell'olfatto*	*Perdita del gusto*	*Tosse*	*Febbre*	*Saturazione Ossigeno*	*Fiato corto*	*Coscienza alterata*
22922	1	3	0	0	0	0	0	0	0	0	20	6	0
23080	1	0	0	0	0	0	0	0	0	50	180	72	0
23081	1	0	0	0	0	0	0	0	74	0	41	74	0
23098	1	21	0	0	0	0	0	0	21	4	16	0	0
23132	1	11	0	0	41	0	0	0	0	29	67	0	0
22964	4	0	21	0	21	0	0	0	21	10	89	0	0
23060	4	46	0	0	0	0	0	0	1	0	36	47	0
23059	5	0	0	0	0	0	0	0	0	7	16	0	0
22833	8	2	0	0	0	0	0	0	2	0	1	2	2
23075	8	0	0	0	0	0	0	0	25	6	25	25	0

Tabella 7.5. Numero di misurazioni con lo stesso sintomo per tutti i pazienti con sintomi persistenti

8

Conclusioni

In questo lavoro di tesi sono stati analizzati i dati relativi al progetto RicovAi-19. In una prima fase, la collaborazione con il team di medici e ingegneri coinvolti è stata marginale in quanto lo scopo iniziale era quello di raccogliere più dati possibili, reclutando la maggior parte dei pazienti malati di Covid-19 nel comune di Offagna e creando una campagna di informazione nella città dello studio pilota.

Durante questa prima fase, sono stati condivisi i primi dati; essi sono stati studiati, modificati, corretti e manipolati per renderli adatti alle analisi svolte successivamente.

Sin dai primi momenti è stata indispensabile una comunicazione diretta con l'equipe medica per poter comprendere al meglio il "significato medico" dei parametri coinvolti e per scegliere la strada più adatta allo studio della sintomatologia.

Lavorare con dati biomedici reali, seppur stimolante e interessante, è stato piuttosto complesso; gli errori, spesso di distrazione, hanno rallentato le analisi e, in alcuni casi, hanno reso impossibile creare soluzioni per poter salvare le informazioni; talvolta, quindi, la mancanza di alcuni dati, durante la fase progettuale, non ha permesso di realizzare alcune idee, frutto di ragionamenti svolti considerando il dataset senza l'errore umano.

Il problema principale, inoltre, è stato la mancanza di una mole di dati adeguata per poter applicare in maniera ottimale gli algoritmi di Intelligenza Artificiale.

I risultati, comunque, sono stati in linea con ciò che riporta la letteratura, seppur minima, rispetto all'andamento dei sintomi di pazienti affetti da Covid-19. Gli algoritmi utilizzati e l'idea di base risultano, comunque, interessanti; essi sono stati scritti in modo standard e certamente, con una quantità di dati più elevata e con la giusta attenzione alla pulizia degli stessi durante la fase di raccolta, i risultati potranno essere più accurati e le considerazioni più interessanti.

Per quanto riguarda il progetto, lo studio pilota nel comune di Offagna ha avuto un gran successo; la maggior parte dei pazienti e dei medici coinvolti ha saputo apprezzare, seppur con una iniziale difficoltà all'adattamento, i benefici della telemedicina.

Riferimenti bibliografici

1. Report Contagiati Per Comune. https://www.regione.marche.it/Regione-Utile/Salute/Coronavirus/Report-contagiati-per-comune/, 2020.
2. Progetto RicovAi-19. https://www.regione.marche.it/News-ed-Eventi7, 2021.
3. Covid, Almawave porta l'Intelligenza Artificiale nella lotta alla pandemia, 2021. https://www.adnkronos.com/covid-almawave-porta-lintelligenza-artificiale-nella-lotta-alla-pandemia_1eFOIe6tRt69OnNi7OJHnF
4. Pazienti Covid, l'intelligenza artificiale per limitare i ricoveri. https://www.ilrestodelcarlino.it/ancona/cronaca/pazienti-covid-ricoveri-intelligenza-artificiale-1.6223393, 2021.
5. Rapporto ISS COVID-19 n. 12/2020 - Indicazioni ad interim per servizi assistenziali di telemedicina durante l'emergenza sanitaria COVID-19. Versione del 13 aprile 2020. https://www.iss.it/rapporti-covid-19/-/asset_publisher/btw1J82wtYzH/content/rapporto-iss-covid-19-n.-12-2020-indicazioni-ad-interim-per-servizi-assistenziali-di-telemedicina-durante-l-emergenza-sanitaria-covid-19.-versione-del-13-aprile-2020.
6. Colab: Overview of Colaboratory Features. https://colab.research.google.com/notebooks/basic_features_overview.ipynb.
7. Scikit-learn. https//scikit-learn.org/stable/.
8. Numpy. https://numpy.org.
9. Seaborn. https://seaborn.pydata.org/generated/seaborn.heatmap.html.
10. PCA sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.
11. DBSCAN sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html.
12. K-Means sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.
13. McKinney, Wes. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc., 2017.
14. Rogel-Salazar, Jesus. *Data Science and Analytics with Python*. Chapman & Hall/CRC, 2017.
15. Ilyas, Ihab F. and Chu, Xu. *Data Cleaning*. Association for Computing Machinery, 2018.
16. Kelleher, John D. and Tierney, Brendan *Data science*. The MIT Press, 2018.
17. Lutz, Mark. *Python*. Ed. O'Reilly, 2014.
18. Lutz, Mark. *Programming Python*. Ed. O'Reilly, 2011.
19. Mckinney, Wes. *Python for data analysis: data wrangling with pandas, numpy, and ipython*. Ed. O'Reilly Media, 2017.

20. Nelli, Fabio. *Python Data Analytics: With Pandas, NumPy, and Matplotlib.* Ed. O'Reilly, 2018.
21. Heydt, Michael. *Learning pandas: high perfomance data manipulation and analysis using Python.* Ed. O'Reilly, 2018.

Ringraziamenti

Il primo ringraziamento va alla mia famiglia che mi ha accompagnata stringendomi la mano passo dopo passo, per questa intensa salita, dai momenti più difficili a quelli di estremo entusiasmo; questo traguardo raggiunto è dedicato unicamente a loro.

Ringrazio il Professore Domenico Ursino per l'empatia e la disponibilità mostrate durante tutto il corso di studi, dalle prime lezioni alla firma della tesi e per aver creduto in me facendo in modo che anche io partecipassi al progetto di tirocinio rappresentando l'Università.

Ringrazio anche Gianluca Bonifazi, che mi ha accompagnata nel lavoro di tirocinio, per aver avuto la pazienza di confrontarsi con me costantemente e per avermi permesso di collaborare con l'intero team senza mai sentirmi fuori posto.

Vorrei ringraziare, infine, i miei amici, quelli di sempre che mi hanno sostenuta da lontano; e quelli che, in punta di piedi, in pochissimo tempo, sono diventati la mia seconda famiglia ad Ancona, con cui ho condiviso momenti di studio intensi e, contemporaneamente, di leggerezza indimenticabile.