POLITECNICO DI TORINO

Laurea Magistrale in Ingegneria Informatica

Approccio ibrido per la classificazione gerarchica automatica di oggetti di contratti della Pubblica Amministrazione Italiana

Relatori

Prof. Antonio VETRÒ

Prof Juan Carlos DE MARTIN

Dott. Davide ALLAVENA

Candidato

Domenico Flavio AMATO

Aprile 2022

"Where children shivered in the cold,
the sun will shine."
Fermi Paradox, Avenged Sevenfold

Ringraziamenti

In genere durante i ringraziamenti si parla di quanto sia stato difficile e pieno di ostacoli il percorso affrontato.
Il mio non fa eccezione, ci sono stati periodi difficili, ma sono riuscito a superarli anche grazie al supporto di chi mi è stato vicino sia dal punto di vista umano che da quello didattico e professionale.

Vorrei ringraziare il mio relatore, il prof. Vetrò per essere stato molto presente durante la stesura dell'elaborato e per avermi permesso di conoscere la realtà di Synapta. Un gran ringraziamento va a Giulio Carducci, sempre pronto a soddisfare le richieste di informazioni sul classificatore di primo livello e al team di Synapta, che rende l'esperienza in azienda sempre stimolante e degna di nota. C'è qualcosa da imparare da ognuno di voi, grazie per avermi accolto e messo a mio agio.
Bisogna evidenziare l'enorme apporto e la costante presenza di Davide Allavena, per avermi guidato durante lo sviluppo. Grazie per la pazienza, per i meme, per il tempo concessomi e per i confronti che abbiamo avuto, nulla di tutto ciò è mai stato dato per scontato.

Ringrazio Domenico, saggio consigliere e amico, per essere stato presente durante la fase preparatoria della tesi, con suggerimenti acuti e mirati, Gianluca per le consulenze legali, i ragazzi di via Fieramosca, Pasquale, Salvatore (rigorosamente secondo), Giuseppe(x2), Traspadano e Benedetto, con cui ho condiviso gran parte della mia carriera universitaria, Calogero e Davide, amici da una vita, e Francesco, inesauribile fonte di idee, di ispirazione e di motivazioni.

Ringrazio mia zia Rossana, Miriam, Angelo - con cui condivido il *difetto di fabbrica* - e Anna per avermi dato spensieratezza nei momenti più delicati. Grazie ai miei genitori per il sostegno continuo, per la fiducia e per avermi finanziato, ai miei fratelli, Totò e Flavio, per la franchezza, per condividere

con me paure, sogni e ambizioni, e a Valentina, per avermi sempre ascoltato, supportato e sopportato. Non sarei la persona che sono senza di voi.
Infine, vorrei ringraziare coloro i quali sono troppo lontani perché possiamo sentirne la voce, ma - sono certo - se fossero con noi oggi, sarebbero felici.

"Frangar, non flectar",
D.

Indice

Elenco delle figure

Capitolo 1

Introduzione

La materia degli appalti pubblici è un aspetto centrale all'interno dello Stato italiano. Con l'avvento della tecnologia e della digitalizzazione, alcune leggi hanno obbligato le pubbliche amministrazioni a pubblicare i dati riguardanti il loro appalti in formato aperto, gli *Open Government Data*.

Questi dati, pubblicati all'interno della sezione *Amministrazione Trasparente* dei siti web delle stazioni appaltanti, sono accessibili da chiunque e senza autenticazione; vengono raccolti e indicizzati da Synapta per offrire uno strumento di *business intelligence* alle aziende interessate a monitorare gli appalti pubblicati o alle stazioni appaltanti interessate a monitorare le aree in cui converge la propria spesa.

Gli schemi per la pubblicazione dei contratti pubblici, però, non prevedono nessun campo che indichi la categoria merceologica di appartenenza dell'appalto richiesto dall'ente aggiudicatore, quindi per offrire uno strumento adatto alle esigenze di mercato, viene aggiunto al contratto il codice CPV, un vocabolario sviluppato dall'Unione Europea per fornire agli stati membri un sistema unico di classificazione, un albero gerarchico.

Lo scopo di questo lavoro di tesi è quello di sollecitare una riflessione sulle possibilità di classificazione automatiche dei dati relativi ai contratti pubblici

tramite approccio ibrido.

In Synapta viene utilizzato un classificatore random forest per inferire la classe di primo livello a cui il contratto appartiene; i contratti con poche parole, tuttavia, non possono essere classificati, in quanto non contengono le informazioni minime richieste per poterli classificare.

A tale scopo, quindi, è stato implementato un sistema di regole basato su regex per identificare il contratto esaminato tramite corrispondenza tra la sottostringa di ricerca della *regular expression* e l'oggetto del contratto analizzato.
In parallelo alle regex sono stati implementati due filtri: uno riguardante la copertura della regola sull'oggetto del contratto e uno relativo all'importo previsto per l'appalto erogato.
Dato che le classi di primo livello sono 45, la scrittura di regole è stata circoscritta alle divisioni 33 e 72, che identificano rispettivamente le forniture medicali e i servizi connessi all'informatica.

I contratti che non sono stati taggati dal sistema di regole vengono successivamente processati dal classificatore *random forest*.
Synapta ha già sviluppato un classificatore di primo livello, il nuovo classificatore si pone in cascata a quello in produzione per discriminare tra le sottoclassi della divisione fornita dal classificatore di primo livello: se il classificatore fornisce in uscita, ad esempio, la classe 45, allora la discrimazione riguarderà le sottoclassi della divisione 45.

Il fine di questa ulteriore classificazione è di scendere di almeno due livelli nella gerarchia dei CPV. Tramite questo procedimento è possibile migliorare la trasparenza amministrativa e fornire filtri più granulari alle imprese interessate all'analisi di mercato.

Capitolo 2

Evoluzione sulla materia dei contratti pubblici e problemi aperti

2.1 Cenni sull'evoluzione in materia di trasparenza e contratti pubblici in Italia

La materia degli appalti pubblici, prima che intervenissero le disposizioni europee a integrare la normativa nazionale, è stata disciplinata in momenti diversi della storia italiana post-unitaria.

Il primo provvedimento atto a regolamentare tale materia è il Regio Decreto del 18 Novembre 1923 il quale, nel IV capitolo *"delle entrate dello Stato"*, stabilisce che gli uffici amministrativi sono tenuti a comunicare alla Ragioneria di Stato i provvedimenti di qualsivoglia natura da cui possano derivare impegni di spesa [1]. Contestualmente, alla Ragioneria bisognerà fornire anche l'importo di tali impegni, sanzionando la mancata comunicazione dell'impegno di spesa alla Ragioneria.

Il decreto regola altresì la modalità di pagamento da parte delle pubbliche amministrazioni e stabilisce che *«il ministro delle finanze può provvedere ad ispezioni per riconoscere l'esistenza presso i funzionari delegati alle somme prelevate e la regolarità dei pagamenti disposti o effettuati»* [2] [Legge 2440/1923, art. 58].

Questa norma intendeva *«garantire l'interesse pubblico attraverso la definizione della scelta del miglior offerente, sempre nel caso di contratti pubblici, così da garantire gli interessi di economicità e di efficacia dei lavori proposti dal singolo operatore che vi partecipava»*[3][Tulino, 2013].

La pubblicazione sulla Gazzetta Ufficiale di *«atti e comunicati che interessino la generalità dei cittadini e la cui pubblicità risponda ad esigenze di carattere informativo»* [4][Legge 839/1984, art. 3] è stabilita dalla legge n. 839 dell'11 Dicembre 1984.

Durante gli anni '90, una serie di norme e direttive emesse dall'Unione Europea si affiancano alla preesistente regolamentazione del settore in campo nazionale, dando impulso alla materia legislativa riguardante gli appalti pubblici.

La legge n. 241 del 7 Agosto 1990 comincia a disciplinare la trasparenza e l'accesso agli atti amministrativi, statuendo che le pubbliche amministrazioni sono tenute a pubblicare gli atti dei provvedimenti adottati.

Essa chiarisce, altresì, che un documento amministrativo è *«ogni rappresentazione grafica, fotocinematografica, elettromagnetica o di qualunque altra specie del contenuto di atti, anche interni, formati dalle pubbliche amministrazioni o, comunque, utilizzati ai fini dell'attività amministrativa»* [5][Legge 241/1990, articolo 2].

Ancora, la norma garantisce il diritto di accesso, al fine di assicurare la trasparenza dell'attività amministrativa - *«a chiunque vi abbia interesse per la tutela di situazioni giuridicamente rilevanti»* [5][Legge 241/1990, articolo 2] - assicurando al contempo lo svolgimento imparziale dell'attività

amministrativa.

Sul punto va però osservato che non tutto ciò che riguarda i pubblici uffici può essere oggetto di pubblicazione e di trasparenza: il diritto di accesso, infatti, decade nel momento in cui bisogna salvaguardare *«la sicurezza, la difesa nazionale e le relazioni internazionali; la politica monetaria e valutaria; l'ordine pubblico e la prevenzione e repressione della criminalità; la riservatezza di terzi»* [5][Legge 241/1990, art. 24].

Nel 2012, la Legge Anticorruzione istituisce l'ANAC – Autorità Nazionale Anticorruzione – i cui compiti, tra gli altri, sono quelli di vigilare sulla applicazione della legge in materia di appalti e di svolgere *«attività di controllo, di prevenzione e di contrasto della corruzione e dell'illegalità nella pubblica amministrazione»* [6][Legge 190/2012, art. 1] e di analizzarne le cause e i fattori.

Il decreto legislativo n. 33 del 14 Marzo 2013 modifica parzialmente la legge del 7 Agosto 1990, estendendo l'accesso civico. Mentre la legge consentiva l'accesso ai documenti previa presentazione di richiesta motivata, il nuovo provvedimento rimuove ogni limitazione alla richiesta di accesso.
Il decreto legislativo 33/2013, inoltre, impone alla pubblica amministrazione di inserire nella *home page* dei siti istituzionali una sezione chiamata *«Amministrazione trasparente»* che conterrà dati, informazioni e documenti pubblicati. È fatto obbligo di pubblicare i dati in modo che questi possano essere indicizzati dai motori di ricerca web.

Con cadenza semestrale, le pubbliche amministrazioni sono obbligate a pubblicare e aggiornare gli elenchi dei provvedimenti adottati, in particolare i provvedimenti che includono *«autorizzazione o concessione; scelta del contraente per l'affidamento di lavori, forniture e servizi»* [7][D. Lgs. 33/2013, art. 23] con relative mansioni assegnate e l'eventuale impegno di spesa. Sono soggetti a pubblicazione anche gli *«accordi stipulati[...] con soggetti privati o altre amministrazioni pubbliche»* [6][Legge 190/2012, art. 1].

L'accesso civico viene esteso e generalizzato con il Freedom of Information Act (FOIA), tramite cui si riconosce *«la libertà di accedere alle informazioni in possesso delle pubbliche amministrazioni come diritto fondamentale [...] così da svolgere un ruolo attivo di controllo sulle attività delle pubbliche amministrazioni»* [8] [Dipartimento della funzione pubblica, FOIA].
A differenza del decreto legislativo 33/2013, il FOIA consente a chiunque di accedere a tutti i dati di cui le pubbliche amministrazioni sono in possesso, riconoscendo l'accesso alle informazioni delle pubbliche amministrazioni come un diritto del cittadino.

Il Decreto Legislativo n. 50 del 18 Aprile 2016, conosciuto anche come Codice dei contratti pubblici, *«disciplina i contratti di appalto e di concessione delle amministrazioni aggiudicatrici e degli enti aggiudicatori aventi ad oggetto l'acquisizione di servizi, forniture, lavori e opere»* [9][D. Lgs. 50/2016, art. 1]. Con il codice dei contratti si passa *«da gare formali, in cui si premia l'offerta confezionata meglio, a gare sostanziali ove trionfa l'offerta migliore; il passaggio da criteri meramente economistici [...] ad approcci qualitativi; [...] il passaggio da un'Autorità di vigilanza con poteri limitati a un'ANAC che [...] assomma, in modo collaborativo e sinergico, poteri di regolazione, di vigilanza, di sanzione e di consulenza»* [10] [Caringella - Manuale di diritto amministrativo, p. 1523, 2018].

2.2 ContrattiPubblici.org e i problemi connessi ai dati delle pubbliche amministrazioni

ContrattiPubblici.org è un motore di ricerca che permette la navigazione tra i contratti delle pubbliche amministrazioni. È consentito l'accesso sia al privato cittadino desideroso di conoscere il modo in cui la PA spende i fondi che le vengono erogati, sia ai fornitori di servizi, opere o materiale di varie

categorie.

Quello dei contratti pubblici è un mercato di grandissime dimensioni sia economiche che sociali, stimato nell'11% [11] [Pisanu, 2021] del PIL : si tratta di un valore superiore ai 100 miliardi di euro[1]. Essendo presenti nello Stato italiano quasi 8000 comuni e più di 12'000 pubbliche amministrazioni, è facilmente intuibile che reperire i vari appalti delle PA sia uno sforzo considerevole, in quanto, sebbene sia stata creata la sezione trasparenza all'interno dei siti web dei comuni, sarebbe comunque necessario consultare singolarmente i siti della pubblica amministrazioni: una ricerca di questo tipo risulterebbe onerosa e poco esplicativa.

Alla vasta legiferazione in materia non sempre segue una corretta applicazione della stessa da parte della pubblica amministrazione: nonostante ANAC abbia dettato le linee guida circa la pubblicazione dei dati in formato XML, vengono spesso commessi errori in fase di compilazione da parte della pubblica amministrazione.
Errori ricorrenti nei file pubblicati sono la mancanza quasi sistematica delle voci di data di inizio e ultimazione del servizio o della fornitura, sebbene *«i tempi di completamento dell'opera sono espressamente previsti dalla legge 190/2012 come una serie di informazioni che le stazioni appaltanti sono tenute a pubblicare»* [12] [ContrattiPubblici.org].

In altri casi, come quello dei raggruppamenti temporanei di imprese, vengono adoperati dei tag errati che poi vengono esportati compromettendo la correttezza delle informazioni riportate. Nel momento in cui questi dati "scorretti" vengono processati automaticamente, ci si potrebbe trovare di fronte a un'informazione potenzialmente fuorviante, soprattutto se alcuni sistemi effettuano predizioni basate su quei dati. *«Pubblicare i dati senza controllarne appropriatamente la qualità ne può compromettere il riutilizzo*

[1]Stima PIL 2020 della World Bank. A partire dal dato della nota precedente, insieme a quello riportato dalla World Bank, si estrapola il valore effettivo di mercato.

e incidere negativamente sulla partecipazione civile»[11] [Pisanu, 2021]: gli errori interni ai dati possono essere disambiguati soltanto dall'attività umana. Ciò rende il processamento dei dati molto più complicato e oneroso e potrebbe portare a un errore nel momento in cui i dati vengono riutilizzati e vengono adoperati degli strumenti di decisione automatica.

2.3 I dati trasmessi dalla pubblica amministrazione: gli Open Government Data

Oltre alla regolamentazione delle modalità di pubblicazione dei dati e la pubblicità dei bandi pubblici, la legge 190/2012 individua le specifiche tecniche che i dati devono avere nel momento in cui vengono pubblicati e sono resi disponibili all'utenza.
Secondo quanto stabilito dal Codice dell'Amministrazione Digitale, i documenti delle pubbliche amministrazioni devono essere in *«formato aperto: un formato di dati reso pubblico, documentato esaustivamente e neutro rispetto agli strumenti tecnologici necessari per la fruizione dei dati stessi»* [13][Codice dell'Amministrazione digitale, art. 68]. Anche i dati devono essere aperti, bisogna pubblicarne tramite documentazione, la sintassi, la semantica, il contesto operativo e le modalità di uso [14][AGID, formati aperti].

«Per gli Open (Government) Data sono state svolte principalmente attività in grado di favorire la collezione dei dati. [...] Molto del lavoro fatto dalle PA sugli Open Government Data è stato soprattutto per fare divulgazione e sensibilizzare i vari soggetti sul tema e costruire infrastrutture e portali in grado di mettere a disposizione il dato in versione aperta. [...] Così facendo i cittadini possono riutilizzare gli open data per attività rivolte alla trasparenza e fornitori e Pubbliche Amministrazioni per fare analisi strategiche del mercato e trarre vantaggio da dati che sono potenzialmente sempre stati a disposizione di tutti». [15][Morando, 2020]

«I dati aperti sono dati che possono essere liberamente utilizzati, riutilizzati e ridistribuiti da chiunque, soggetti eventualmente alla necessità di citarne la fonte e di condividerli con lo stesso tipo di licenza con cui sono stati originariamente rilasciati» [16][Open Data Handook] a condizione, però, di indicare la provenienza del dataset. È possibile anche utilizzare i dati per degli scopi diversi da quelli per cui sono stati pubblicati. Di fatto, i dati raccolti da Synapta per alimentare la piattaforma *«vengono messi a disposizione [...] per finalità anticorruzione [...] e hanno libertà di riutilizzo che permette un'analisi [...] per aiutare una stazione appaltante a trovare il giusto fornitore [...] o un fornitore a individuare nuove opportunità di business»* [15] [ContrattiPubblici.org, 2020].

I dati dei contratti, inoltre, possono essere utilizzati per fini di trasparenza, perché *«se i cittadini sapessero delle azioni dei loro governi, potrebbero prendere decisioni meglio informate e chiedere servizi migliori. Gli open data possono aiutare i governi a rimanere all'erta e intraprendere migliori azioni politiche per la società, l'economia e l'ambiente»* [17] dunque, per quanto riguarda la consapevolezza circa l'uso dei fondi delle pubbliche amministrazioni e i settori economici più critici che necessitano intervento, gli open data sono *«buoni per la democrazia»* [17].

Tramite questo portale, ContrattiPubblici.org, il privato cittadino può avere più contezza delle azioni intraprese dalle pubbliche amministrazioni. Queste ultime, dal canto loro, possono avere una migliore visione di insieme al fine di perfezionare la spesa pubblica. I dati possono essere utilizzati come strumento di lotta alla corruzione. Ulteriormente, le aziende eroganti servizi e forniture possono utilizzare il portale nell'ottica di *business intelligence* e per effettuare analisi di mercato, valutando l'entità del mercato per periodo e area geografica e l'importo dovuto dalle PA alle aziende vincitrici della gara.

Attraverso la piattaforma ContrattiPubblici.org, è possibile filtrare i contratti per importo erogato alla fornitura o al servizio, analizzando dei pattern ricorrenti all'interno dei bandi della pubblica amministrazione. Si può notare come alcune tipologie di contratti siano ripetutamente affidati a dei fornitori. In particolare, sono contratti con importo sotto soglia, che possono essere affidati in modo più discrezionale da parte della stazione appaltante e per cui è prevista la possibilità di affidamento diretto. Un sistema di questo tipo si «nutre» dei dati condivisi dalle pubbliche amministrazioni in formato aperto.

I "*linked data*" costituiscono il miglior modo in assoluto per pubblicare e collegare dati strutturati. I *linked data* si realizzano concretamente nella creazione di collegamenti tra dati di diverse sorgenti, come database mantenuti da fonti diverse o, più semplicemente, sorgenti eterogenee che non hanno interoperabilità tra i dati. La pubblicazione di dataset come *linked data* deve attenersi alle linee guida che seguono[2]:

1. assegnare URI alle entità descritte dal dataset e usare questi URI HTTP in modo tale che questi oggetti possano essere referenziati e cercati da persone e user agent;
2. fornire link RDF ad altri datasource sul web, per agevolare la navigazione all'interno dei dataset collegati;
3. fornire metadati sui dati pubblicati, in modo che sia facile verificare la veridicità delle informazioni riportate;
4. usare gli URI per indicare i dati, cosicché altri utenti possano puntare ai dati di partenza;
5. collegare i dati ad altri dati per fornire contesto.

La proposta di Tim Berners-Lee basata sulla classificazione a 5 stelle, però, non prende in carico tutti i problemi connessi ai dati, in quanto focalizza

[2]L'elenco è tratto da [18][Bizer et al., 2009].

l'attenzione soltanto sul formato utilizzato per la pubblicazione dei dati. Dunque non è possibile stimare la qualità complessiva dei dati; potrebbe accadere di avere un dataset aperto, in formato non proprietario, con tutti i metadati inseriti correttamente e le celle tutte presenti, alcune delle quali prive di un contenuto.
Nonostante le numerose lacune intrinseche, un dataset del genere potrebbe essere considerato valevole delle 5 stelle per la completezza formale pur non essendo né interpretabile, né riutilizzabile perché non comunica nulla.
Appare ovvia, quindi, la necessità di introdurre altre metriche di valutazione della qualità dei dati che prendano in considerazione altri parametri, oltre al formato utilizzato per la pubblicazione.

Nel 2007, un gruppo di esperti di Internet e open data riuniti sotto il nome di "*Open Data government working group*" ha pubblicato una serie di attributi di cui i dati aperti dovrebbero godere. I dati devono essere completi, primari - ovvero presi dalla fonte così come sono - , accessibili, processabili dalle macchine, aggiornati, non discriminatori in quanto è possibile accedervi anche senza autenticazione e in formato non proprietario, i dati devono essere documentati, sicuri da aprire e gratuiti [19][Vetrò et al., 2016].

2.4 La qualità dei dati

La qualità dei dati contenuti all'interno di un dataset può impattare in modo significativo e su molteplici aspetti le attività delle aziende e la trasparenza nei confronti dei cittadini[3]:

- Processo di decisione automatica: migliore è la qualità dei dati, maggiore sarà l'affidabilità sui risultati generati;

[3]L'elenco è tratto da [20][Moreno, 2017].

- Produttività: dati di buona qualità permettono ai tecnici di essere più produttivi, non è necessario sprecare tempo per validare i risultati e «aggiustare» i dati in ingresso rendendoli processabili;
- Secondo il «Global CEO Outlook» di KPMG del 2016, l'84% dei CEO è preoccupato circa la qualità dei dati su cui vengono basati i processi di decisione automatica;
- dati di bassa qualità possono portare alla perdita di guadagni per l'azienda, in quanto quest'ultima non è in grado di fornire output affidabili, quindi oltre a perdere potenziali clienti, si avrebbe un danno reputazionale.

Allo scopo di stimare la qualità dei dati l'ISO - International Organization for Standardization - propone due framework, l'ISO 25012 e 25024.

Lo standard ISO 25024 fornisce delle formule e delle metriche per effettuare le misurazioni degli indici di qualità dei dati. Queste misurazioni si basano su alcuni aspetti dei dati individuati nello standard ISO 25012, il cui scopo è quello di definire un modello sulla qualità dei dati, «per supportare le aziende ad acquisire, manipolare e usare dati con le necessarie caratteristiche qualitative nell'ottica di perseguire i propri obiettivi». [21][standard ISO 25012]

Gli indici di qualità - ad alto livello - indicati in questo standard possono essere utili per molteplici scopi, tra cui quello di identificare i requisiti che i dati devono possedere nel momento in cui vengono acquisiti o processati.

Vengono individuati sei requisiti funzionali; la funzionalità è definita come «*la capacità dei dati di andare incontro alle esigenze degli utenti e ai loro obiettivi*»[20][Moreno, 2017].
Più in generale, questi requisiti afferiscono alle funzioni per cui i dati in questione possono tornare utili.

Alcune caratteristiche per la valutazione della qualità dei dati sono:

- *consistency*: la consistenza si riferisce all'assenza di contraddizioni all'interno dei dati;

- *currency*: è la misura in cui il dato è aggiornato. È possibile misurarla rispetto ai dati reali e al lavoro che si sta effettuando, si tratta di un aspetto critico per dati altamente volatili o che si aggiornano con alta frequenza;

- *completeness*: la misura in cui tutti i valori necessari sono stati assegnati e memorizzati. Si riferisce sia ai diversi record presenti nel dataset, sia alle singole tuple presenti all'interno del record stesso. Dal punto di vista di un utente, la completezza indica - in modo quantitativo - quanto è possibile soddisfare le esigenze dell'utente a partire dai dati. Si tratta, anche, di un attributo grazie al quale i dati stessi sono in grado di fornire contesto alle osservazioni effettuate da utenti;

- *accuracy*: indica la conformità dei valori dei dati ai valori effettivi. Vengono individuati due tipi di accuratezza, sintattica e semantica. La prima indica la vicinanza del valore del dato a un insieme di valori considerati sintatticamente corretti; si parla di inaccuratezza sintattica quando vengono memorizzati valori contenenti refusi, e.g. scrivere «softyware» anziché «software». L'accuratezza semantica è, invece, la vicinanza dei valori dei dati a un insieme di valori semanticamente corretti; una bassa accuratezza semantica si ha nel momento in cui si sbaglia a memorizzare un valore che è nello stesso insieme semantico di quello che dovrebbe essere immagazzinato, e.g. supponiamo di volere indicare un sistema operativo, ma anziché «Windows» viene inserito «iOS»; i due valori sono sintatticamente accurati, perché hanno come riferimento lo stesso dominio, tuttavia i sistemi in questione sono diversi;

- *understandability*: indica quanto è facile la comprensione delle informazioni riportate all'interno del dataset. Concorrono nella valutazione di questo indicatore anche le unità di misura;

- *portability*: indica l'interoperabilità dei dati, cioè la capacità di essere spostati da una piattaforma all'altra senza però perdere di efficacia nell'utilizzo delle informazioni [19][estratto da Vetrò et al., 2016].

Oltre a definire le caratteristiche appena citate, ISO/IEC 25012 *«postula che queste caratteristiche possono essere valutate oggettivamente e usate per determinare l'aderenza»*[21][ISO 25012] allo standard; lo scopo è quello di fornire linee guida per la qualità dei dati. Tutto questo, a sua volta, si riflette sulla destinazione d'uso delle informazioni all'interno del dataset e sulle possibilità di interoperabilità e cooperazione aventi gli stessi dati come comune denominatore.

Nonostante l'ISO definisca le metriche per la qualità dei dati e queste vengano motivate, rimangono semplicemente delle best practices abbastanza fini a se stesse.

Uno dei primi effetti del miglioramento della qualità degli Open Government Data concerne la trasparenza amministrativa, poiché tutte le informazioni riportate nel dataset sono accurate, aggiornate e affidabili: *«quando una Pubblica Amministrazione è in grado di sfruttare e analizzare i Big Data a sua disposizione, vede subito miglioramenti nella gestione dei servizi pubblici»*[22][ContrattiPubblici.org].

Con il Piano Triennale per l'Informatica 2017-2019 è stato introdotto il DAF, Data & Analytics Framework, il cui scopo è quello di promuovere la trasformazione digitale delle Pubbliche Amministrazioni, di sensibilizzare e valorizzare il patrimonio informativo nazionale. Il DAF offre strumenti di gestione e analisi dei dati, risolve tematiche di Data Governance, standardizzazione, integrazione, gestione delle API e degli Open Data. Il sistema di gestione e analisi dei dati, in questo modo, viene centralizzato e l'onere della computazione non ricade più sulle singole pubbliche amministrazioni, con conseguente economizzazione risorse.

Tuttavia, per quanto riguarda la creazione dei dataset *«l'Italia è caratterizzata da un'elevata decentralizzazione amministrativa che fa sì che il ruolo delle PA regionali/locali sia particolarmente rilevante nel processo di innovazione tecnologica»*[23][Piano Triennale ICT - Sommario, 2020] , quindi la realizzazione delle innovazioni ricade quasi interamente sulle Pubbliche Amministrazioni perché si punta al miglioramento della qualità dei dati, per monitorare la quale sono state introdotte apposite attività.

Tra i temi trattati dal DAF sono presenti dati e interoperabilità. I risultati attesi concernenti l'interoperabilità includono l'incremento del numero di API e l'ampliamento del bacino di utenti - cittadini e imprese - registrate come fruitori delle API [24] [Piano Triennale ICT - Interoperabilità: Obiettivi e risultati attesi, 2020], quindi l'obiettivo è quello di convogliare l'attenzione dei cittadini su questi dataset nell'ottica di avere una maggiore partecipazione attiva nella cosa pubblica

Alcuni obiettivi individuati per migliorare la qualità dei dati riguardano [25][Piano Triennale ICT - Dati: Obiettivi e risultati attesi, 2020] :

- aumento del numero di basi di dati che espongono API coerenti con modelli interoperabilità di dati nazionali ed europei;
- aumento del numero di dataset aperti dinamici;
- aumento del numero di dataset con metadati di qualità conformi agli standard di riferimento europei;
- aumento del numero di dataset aperti conformi con caratteristiche di qualità dello standard ISO/IEC 25012 e pubblicati con licenza aperta.

La lettura degli obiettivi fissati dal piano triennale per ICT, permette di desumere che la qualità dei dati condivisi dalle Pubbliche Amministrazioni, soprattutto a livello più decentralizzato, come i comuni, è tendenzialmente bassa e non soddisfacente sotto molti aspetti.

Avere dati qualitativamente «poveri» potrebbe inficiare la trasparenza amministativa, rendendo difficile la comprensibilità delle informazioni e inutili gli sforzi legislativi profusi per garantire al cittadino accessibilità e comprensione complessiva di quanto riportato nei dataset.

2.5 Studi correlati alla qualità dei dati delle Pubbliche Amministrazioni italiane

Per comprendere pienamente la qualità dei dati che i vari enti pubblici rendono disponibili online, bisognerebbe esaminare alcuni contributi in torno a questo tema.

Doversi interfacciare con una quantità di dati così grande e renderli qualitativamente fruibili è un problema nevralgico all'interno dei flussi di informazione condivisi dalla pubblica amministrazione: di fatto dati non buoni potrebbero alterare la vita «naturale» degli *Open Government Data*, il loro processamento, i loro eventuali usi futuri e, conseguentemente, i *linked data* che da questi si originano.

Il primo studio considerato, *Open data quality measurement framework: Definition and application to Open Government Data* [19][Vetrò et al., 2016], analizza OpenCoesione, che contiene dati riguardanti gli investimenti e i progetti delle regioni e del governo finanziati da fondi europei: in OpenCoesione si trovano 900'000 progetti per un ammontare di oltre 90 miliardi di euro di finanziamenti erogati tra il 2007 e il 2013. Ulteriormente, lo studio analizza i dati pubblicati da enti decentralizzati - alcuni comuni italiani - per permettere un confronto sulla qualità dei dati a diversi livelli di centralità.

Il proposito dell'articolo è quello di individuare le criticità esistenti e le buone pratiche acquisite nella pubblicazione degli open data, focalizzandosi sulla qualità intrinseca dei dati.

Un primo controllo su un campione di circa 70 mila contratti distribuiti su sei diverse regioni ha portato a 18 mila progetti scartati e circa 110 mila warnings generati; molti errori erano dovuti all'inconsistenza dei dati riportati, ad esempio indirizzi errati e dati mancanti.

Si sono confrontati i dataset pubblicati da OpenCoesione e dalle municipalità tramite le misure riguardanti la qualità dei dati pubblicati dalle stazioni appaltanti. Il dataset OpenCoesione - centralizzato - mostra una qualità di dati maggiore rispetto a quello delle amministrazioni decentralizzate; ciò potrebbe essere legato alla presenza di controlli di qualità nel processo di creazione del dataset a livello centrale.

Una criticità comune alle due tipologie di dataset risiede nella mancanza di aggiornamenti: le informazioni riguardanti la creazione del dataset sono spesso presenti, ma non c'è nulla che indichi aggiornamenti e/o modifiche di dati. D'altro canto, nei dataset decentralizzati non sono quasi mai presenti i metadati, assenza che di fatto inficia la comprensibilità dell'informazione riportata. In OpenCoesione si modificano e si inseriscono nuovi dati con cadenza bimestrale, pertanto le informazioni riportate sono molto più aggiornate rispetto ai dataset di amministrazioni decentralizzate, che – di contro - pubblicano all'incirca una volta all'anno.

Un secondo studio, *Preserving the Benefits of Open Government Data by Measuring and Improving Their Quality: An Empirical Study* [26] [Vetrò et al., 2017] , analizza alcuni aspetti qualitativi dei dati pubblicati da dodici università italiane: se ne valutano accuratezza - sintattica e semantica-, completezza e consistenza delle informazioni pubblicate.

L'assunto di base è quello secondo cui la pubblicazione dei dati in formato aperto non sia sufficiente a garantire un «efficiente ecosistema di riutilizzo», per cui bisogna valutare la qualità dei dati secondo alcune metriche definite – inizialmente e ad alto livello – tramite lo standard ISO 25012 e successivamente con lo standard ISO 25024. Quest'ultimo, di fatto, rende operativo il primo standard ed è fondamentale per la valutazione quantitativa della

qualità dei dati.

Una problematica rilevata dallo studio sui dataset analizzati è la mancanza del CIG - Codice Identificativo Gara - nei bandi pubblicati dalle università. Questa mancanza incide particolarmente sulla trasparenza, poiché questo campo identifica in maniera univoca la gara d'appalto. Altra situazione deficitaria si rileva nella elencazione dei criteri di scelta del contraente.

Tout court, un cittadino non può ritenere attendibili le informazioni pubblicate da alcune università italiane: le misurazioni effettuate sui dati invalidano circa il 30% dei dati esaminati dallo studio. Qualsiasi riutilizzo di questa porzione di dati porterebbe a dei risultati inconsistenti e/o incompleti, eludendo - praticamente - l'obiettivo per cui gli Open Government Data sono stati pensati [26][Vetrò et al., 2017].

Capitolo 3

I dati aperti sui contratti pubblici

Prima di passare alla descrizione e all'analisi della classificazione ibrida dei contratti pubblici, sarebbe utile avere una panoramica di informazioni relative ai contratti e ai campi di cui sono composti.

3.1 Lo standard definito da ANAC

Il Decreto trasparenza stabilisce, oltre alle tempistiche, il formato e la modalità di pubblicazione dei file relativi ai bandi o contratti pubblici prodotti dalle stazioni appaltanti.

Con un documento, *Specifiche tecniche per la pubblicazione dei dati ai sensi dell'art. 1 comma 32 Legge n. 190/2012* [27][ANAC, 2016], l'Autorità Nazionale Anticorruzione [ANAC] intende stabilire *«le specifiche tecniche a cui la SA [stazione appaltante] deve fare riferimento per adempiere agli obblighi previsti»* [27][ANAC, 2016, p.2]. In particolare, vengono descritte in dettaglio *«le modalità con cui la SA deve comunicare all'Autorità l'avvenuta pubblicazione dei dati sul proprio sito web istituzionale [...] e le*

strutture dati che la stazione appaltante deve utilizzare per la pubblicazione delle informazioni in formato digitale standard aperto sul proprio sito web istituzionale»[27][ANAC, 2016].

La stazione appaltante - o ente aggiudicatore, che affida i contratti pubblici - deve fornire all'Autorità link diretti a «dataset in formato digitale standard aperto contenente i dati per l'anno di riferimento e a un dataset indice in formato digitale standard aperto contenente una collezione di link, che puntano ai singoli dataset in formato digitale standard aperto contenenti i dati per l'anno di riferimento»[27][ANAC, 2016].

Le pubbliche amministrazioni devono, altresì, pubblicare le informazioni riguardanti gli appalti concessi nella sezione Amministrazione Trasparente, della cui obbligatorietà si è avanti discusso, e devono ulteriormente provvedere all'invio dei link diretti ai dataset entro il 31 Gennaio, con cadenza annuale.

La dimensione massima del dataset è di 5 MB, è comunque possibile produrne più di uno nel caso in cui la stazione appaltante abbia un numero di gare tale da sforare la soglia massima prevista. La presenza di più dataset deve essere, anch'essa, riportata tra i riferimenti del dataset indice.

«Le Specifiche ANAC rappresentano [...] il documento di riferimento sulle modalità tecnico-operative che le stazioni appaltanti devono seguire per la pubblicazione dei dati; in particolare, le stazioni appaltanti sono tenute a strutturare le informazioni impiegando il formato standard XML grazie anche al supporto degli schemi di definizione XSD descritti da ANAC stessa»[27][ANAC, 2016]. Tra gli elementi che compongono l'XML [12][ContrattiPubblici.org, 2019] sono presenti la data di pubblicazione del dataset e dell'ultima modifica dello stesso, il CIG (Codice Identificativo Gara) che indica in modo univoco una gara d'appalto, le informazioni sul committente, i partecipanti e gli aggiudicatari della gara insieme con la data di inizio e ultimazione dei lavori e la somma liquidata.

«Per "Data di ultimazione lavori, servizi, forniture" deve intendersi la

data di ultimazione contrattualmente prevista ed eventualmente prorogata o posticipata in virtù di successivi atti contrattuali»[27][ANAC, 2016, p.13]. La possibilità di prorogare o posticipare la data di ultimazione dei lavori, servizi o forniture implica il dover aggiornare questo campo all'interno del dataset già pubblicato.

«Il dovere di aggiornamento è infatti rinvenibile all'interno Codice dei Contratti pubblici, art. 29.1 (Principi in materia di trasparenza): "Tutti gli atti delle amministrazioni aggiudicatrici e degli enti aggiudicatori relativi [...] alle procedure per l'affidamento di appalti pubblici di servizi, forniture, lavori e opere [...] devono essere pubblicati e aggiornati sul profilo del committente". [L'articolo 6 del] Decreto Trasparenza incorpora il rispetto dei doveri di aggiornamento all'interno del più ampio dovere di garantire la qualità delle informazioni riportate sui siti-web istituzionali delle Pubbliche Amministrazioni»[28][ContrattiPubblici.org, 2019].

Tuttavia, nonostante la presenza dei tag di data inizio e fine della fornitura, servizio o lavoro, è stata rinvenuta l'omissione sistematica di queste voci da parte di alcune stazioni appaltanti. *«I tag "dataInizio" e "dataUltimazione" sono accompagnati dall'attributo minOccurs="0", quindi la natura di tali tag»*[12][ContrattiPubblici.org, 2019] è opzionale.

In breve, la gestione amministrativa non sempre si attiene pedissequamente alle prescrizioni della normativa vigente impedendone la piena attuazione, motivo per il quale spesso i dataset risultano incompleti. Capita, talora, che i pubblici funzionari utilizzino dei tag XML per scopi diversi da quelli per cui sono stati ideati, per cui l'informazione riportata non ha correlazione alcuna con il rispettivo campo: *«una pubblicazione di dati tecnicamente lacunosa potrebbe quindi essere ricca di incongruenze. Quando i dati sono predisposti per venire processati in forma automatica (machine-readable data), un'esposizione non corretta finirebbe in sostanza con il restituire e veicolare un'informazione scorretta o solo parzialmente corretta»*[12][ContrattiPubblici.org, 2019] che solo un essere umano potrebbe disambiguare.

3.2 Il portale ContrattiPubblici.org

All'interno di questo ecosistema, si inserisce Synapta con la piattaforma ContrattiPubblici.org. Si tratta di un portale che indicizza e aggrega i contratti della pubblica amministrazione. I contratti vengono aggregati a partire dai dataset che la stazione appaltante rende pubblici e - dato che questi dati sono aperti - ne è concesso il riutilizzo anche per finalità diverse da quelle per cui i dati originali sono stati pubblicati. Gli open data hanno un ruolo cruciale per quanto riguarda la possibilità di uso dei dati. *«Aperto significa che chiunque può liberamente accedere, usare, modificare e condividere per qualsiasi scopo»*[29].

«Questa definizione sottintende che spesso i dati siano raccolti per uno scopo specifico. L'esercizio di renderli disponibili serve proprio ad abilitare riusi dei dati diversi da quelli iniziali per cui erano stati inizialmente raccolti. Per fare un esempio pratico: nel caso di ContrattiPubblici.org, una parte importante dei dati, che Synapta raccoglie come open data, sono dati messi a disposizione per delle finalità differenti da quelle che esistono all'interno della piattaforma. I dati sui contratti pubblici, ad esempio, vengono messi a disposizione dalle Pubbliche Amministrazioni per finalità di anticorruzione, in ottemperanza alla Legge 190/2012. In particolare, i dati sui contratti delle PP. AA., hanno una certa libertà di riutilizzo che permette un'analisi non solo per fini di trasparenza, ma anche per guidare una stazione appaltante a trovare il giusto fornitore per una gara d'appalto, o un fornitore a individuare nuove opportunità di business.
Il tema del riutilizzo degli open data ha anche un aspetto legato al fatto che più dati standardizzati sono messi a disposizione e più è ampia la disponibilità, più è semplice trovare occasioni di riutilizzo che possano trovare una loro autonoma sostenibilità economica»[15][Morando, 2020]. Inoltre *«i dati aperti sono un bene comune che fornisce descrizioni condivise della realtà, stimola dibattiti, livella asimmetrie informative, riduce le barriere all'ingresso per startup e PMI (Piccole e Medie Imprese) innovative, incoraggia responsabilità*

e trasparenza»[30][Morando, 2019].

Sebbene si possa pensare che la piattaforma sia utile unicamente ai fini della trasparenza, sarebbe riduttivo rispetto alle possibilità che Contratti-Pubblici.org offre. Si tratta, infatti, di uno strumento di business intelligence pensato per l'ottimizzazione della spesa pubblica e per la capillarizzazione delle informazioni riguardanti le gare indette e/o espletate.

«ContrattiPubblici.org può essere considerato uno strumento di Open Source Intelligence (OSINT) [...] si tratta di una disciplina dell'intelligence che si occupa della ricerca, raccolta e analisi di dati e di notizie d'interesse pubblico attraverso fonti aperte. L'OSINT utilizza diverse fonti di informazioni fra cui [...] mezzi di comunicazione, come giornali, riviste, televisione, radio e siti web [...][e] dati pubblici, come rapporti dei governi, piani finanziari, dati demografici e conferenze stampa. In questo contesto, ContrattiPubblici.org si rivela uno strumento fondamentale per chiunque voglia fare fact checking. [...] La piattaforma raccoglie dei dati accessibili – ovvero gli XML pubblicati dalle Pubbliche Amministrazioni in ottemperanza della Legge 190/2012 – e li rende fruibili a coloro che vogliono fare attività di Open Source Intelligence. I soggetti a cui ci riferiamo potrebbero essere, ad esempio, giornalisti legati al data driven journalism, ma potrebbero anche essere semplicemente degli utenti che vogliono approfondire le notizie che hanno letto» [31][Zangarini, 2019].

Quindi, ContrattiPubblici.org si rivela utile a coloro i quali vorrebbero informazioni afferenti a servizi, lavori e/o opere affidati con appalti pubblici, tramite un motore di ricerca che può filtrare i contratti e i bandi pubblici per oggetto, per stazione appaltante e per ente aggiudicatore.
Ancora, il portale può essere utilizzato dalle pubbliche amministrazioni per avere una panoramica sui settori in cui converge la propria spesa ai fini dell'ottimizzazione della stessa; inoltre offre ai fornitori la possibilità di creare dei mercati aventi una distribuzione geografica e una porzione di mercato di interesse, al fine di avere una notifica nel momento in cui dovesse

essere indetta una gara pubblica inerente al mercato scelto.

Tra le testimonianze riguardanti l'uso e l'utilità della piattaforma, è significativo riportare le opinioni di Daniela Galletti, *Marketing e Customer Satisfaction Specialist* della direzione Marketing e Innovazione di Reekep e di Anna Cavallo, responsabile direzione marketing e PMO di CSI Piemonte (Consorzio per il Sistema Informativo). I pareri espressi, seppure giungano da posizioni che nel mondo degli appalti pubblici sono opposte – quella di un ente fornitore, la prima, e quella di una stazione appaltante, la seconda – concordano nella valutazione positiva del servizio offerto da ContrattiPubblici.org.

«ContrattiPubblici.org [...] consente di avere una visione d'insieme sull'offerta e la domanda di servizi nei settori di nostro interesse del mercato della PA, per rispondere alle esigenze di oggi e anticipare per tempo le necessità del futuro»[32][Synapta, 2021]. ContrattiPubblici *consente di monitorare e conoscere in anticipo le scadenze dei contratti dei nostri competitor e [di] servir[si] dei dati raccolti, filtrati ed estrapolati dalla piattaforma al fine di supportare l'attività di pianificazione commerciale [...] [è possibile] conoscere in anticipo le future necessità delle pubbliche amministrazioni locali e nazionali mediante l'analisi delle programmazioni di spesa di ogni stazione appaltante [...] [risulta utile la presenza di] grafici semplificati per visualizzare in un colpo d'occhio la distribuzione geografica dei volumi di mercato e importi medi dei contratti, tabelle personalizzabili di riepilogo dei principali acquirenti di un determinato servizio, possibilità di estrarre i dati dalla piattaforma»*[33][Synapta, 2021].

Avere dei dati di alta qualità, saperli comprendere e saper estrarre conoscenza da essi risulta cruciale in questo contesto: *«la disponibilità di dati sui fornitori di opere analoghe è essenziale per poter individuare i partecipanti più adatti ad assicurare un lavoro di qualità»* [34][Synapta, 2022].
Con la legiferazione in materia di appalti pubblici si è intervenuti su molteplici aspetti. *«Il codice dei contratti del 2016 ha sancito [...] il passaggio da gare formali, in cui si premia l'offerta confezionata meglio, a gare sostanziali*

ove trionfa l'offerta migliore [e] il passaggio da criteri [...] che privilegiano i prezzi più bassi ad approcci qualitativi sedotti dall'utilità sostanziale e dalla sostenibilità effettiva dell'offerta»[10][Caringella, 2018, pp. 1522/1523].

3.3 Rappresentazione dei contratti su ContrattiPubblici.org

ContrattiPubblici.org indicizza più di 58 milioni[1] tra contratti, bandi e gare della pubblica amministrazione. Attraverso la navigazione all'interno di questo motore di ricerca è possibile ispezionare i contratti disponibili e visualizzare alcune informazioni correlate al pubblico appalto.

Aprendo la pagina relativa al contratto, la prima informazione disponibile riguarda l'oggetto del contratto, ovvero *«sia il bene, inteso concretamente, in relazione al quale si svolge l'operazione economica sottesa al contratto, sia la stessa operazione economica»*[35].
Di seguito si trovano le informazioni concernenti la stazione appaltante, *«un'amministrazione aggiudicatrice, un ente aggiudicatore o un soggetto aggiudicatore che affida a un operatore economico un contratto pubblico di appalto o di concessione avente per oggetto l'acquisizione di servizi o forniture oppure l'esecuzione di lavori o opere»*[36].

Quindi, vengono mostrati i dati relativi al contratto: data di inizio e di fine dei lavori, delle opere o delle forniture - se presenti nel dataset - , importo concordato e quello effettivamente erogato, tipologia di scelta del contraente e la modalità operativa secondo cui la stazione appaltante seleziona *«l'operatore economico al quale verrà affidata l'esecuzione di un contratto pubblico [...] serve a regolamentare le fasi di aggiudicazione di un contratto pubblico secondo le modalità previste dal codice dei contratti pubblici*

[1]ContrattiPubblici.org, dato risalente al 30 Gennaio 2022.

per ogni procedura di scelta del contraente»[37].
È possibile suddividere le procedure di scelta del contraente in: procedure ad affidamento diretto, in cui non vi è necessità di una gara; e *«procedure competitive ovvero dove è richiesta una fase competitiva tra più operatori economici per selezionare l'offerta migliore»*[38], suddivise a loro volta in procedure aperte, negoziate, ristrette e sistema dinamico di acquisizione:

- procedura aperta: *«qualsiasi operatore economico dotato dei requisiti di partecipazione può presentare un'offerta in risposta a un avviso di indizione di gara. La stazione appaltante pubblica un bando di gara [...] gli operatori economici presentano le proprie offerte che vengono valutate in base al criterio di aggiudicazione prescelto [...][che]può essere tecnico/economico oppure solo economico. Una serie di fasi di verifica porta infine all'aggiudicazione definitiva e alla stipula effettiva del contratto»*[39];

- procedura negoziata: *«procedure di scelta del contraente in cui una stazione appaltante negozia con gli operatori economici prescelti le condizioni dell'appalto [...] attivabili unicamente quando ricorrono gli specifici presupposti indicati dal Nuovo Codice dei Contratti Pubblici agli artt. 59, comma 2, e 63 (settori ordinari) e all'art. 125 (settori speciali)»*[40];

- procedure ristrette: *« procedure di scelta del contraente in cui solo gli operatori economici prescelti da una stazione appaltante possono partecipare al bando di gara»*[41];

- sistema dinamico di acquisizione: *«procedura di acquisizione interamente elettronica con cui le stazioni appaltanti acquistano beni e servizi di uso corrente, generalmente disponibili sul mercato»*[42].

Sempre nella sezione riguardante i dati dell'appalto è possibile trovare il Codice Identificativo Gara, un codice alfanumerico di dieci caratteri che identifica in maniera univoca la gara d'appalto ed è utile, inoltre, *«per*

tracciare le movimentazioni finanziarie degli affidamenti di lavori, servizi o forniture»[43] . Il CIG, rilasciato da SIMOG, il Sistema Informativo di Monitoraggio delle Gare dell'ANAC, è obbligatorio per ogni tipologia di appalto pubblico, anche i contratti assegnati con affidamento diretto devono esserne provvisti, sebbene in questo caso non ci sia nessuna gara e va richiesto indipendentemente dall'importo erogato previsto. *«Ai contratti sotto soglia si applica un CIG semplificato (detto anche Smart CIG), che può essere richiesto fornendo un numero ridotto di informazioni e documenti rispetto alla procedura ordinaria. Il CIG smart si applica anche ad altre tipologie di contratti, come stabilito dalle leggi che disciplinano la tracciabilità dei flussi finanziari»*[43]. Chiude il raggruppamento dei dati del contratto il codice CPV, che identifica la categoria merceologica del contratto, codice di cui si dirà di seguito.

Infine, è possibile trovare le informazioni riguardo i partecipanti alla gara, l'ente aggiudicatario ed eventuali allegati, se presenti.

Tutti questi elementi - insieme - descrivono il contratto o bando di gara su ContrattiPubblici.org. Le informazioni riportate rendono la procedura degli appalti trasparente e accessibile, permettendo all'utente di fruire di una rappresentazione accurata del contratto, attraverso un'interfaccia grafica che rende il contenuto fruibile a qualsiasi tipo di utente a differenza di un file XML pubblicato dalla stazione appaltante.

3.4 I codici CPV

ContrattiPubblici.org è uno strumento di business intelligence, permette alle imprese interessate a concorrere per aggiudicarsi gli appalti pubblici di definire un mercato personalizzato.

All'interno del portale, nella sezione "Mercati" è possibile creare il proprio mercato di interesse inserendo all'interno del form che si apre le parole chiave

relative alla porzione di mercato desiderata. È possibile definire un'area operativa, in termini di comuni, province o regioni in cui si desidera agire o monitorare le richieste delle diverse stazioni appaltanti presenti all'interno del territorio selezionato.

Per poter indicizzare i contratti in modo significativo per le aziende bisognerebbe aggregarli per categoria merceologica; tuttavia, tra le linee guida per gli XML pubblicate da ANAC non è presente nessuna indicazione che riguardi la categoria merceologica del contratto di appartenenza. Da qui scaturisce la necessità di dover aggiungere ai campi del contratto un codice in grado di identificare in maniera univoca la categoria merceologica a cui il contratto appartiene, ovvero il codice CPV.

Con CPV - Common Procurement Vocabulary - si intende *«il vocabolario comune per gli appalti pubblici»*[44], un sistema di classificazione che identifica in modo univoco una determinata categoria merceologica presente nell'oggetto di un bando pubblico .

«È stato introdotto a livello europeo con il Regolamento (CE) n. 2195/2002 e successivamente fatto oggetto di modifica con il Regolamento (CE) n. 213/2008. È composto da un vocabolario principale a nove cifre, per la descrizione di forniture, lavori o servizi oggetto dell'appalto, e da un vocabolario supplementare da sei caratteri alfanumerici, che può essere utilizzato per completare la descrizione dell'oggetto fornendo dettagli aggiuntivi sulla natura o sulla destinazione del bene da acquistare»[44].

«Il proposito del CPV è di standardizzare, tramite un singolo sistema di classificazione per gli appalti pubblici, i termini usati da autorità ed enti appaltanti per descrivere il soggetto del contratto offrendo uno strumento appropriato a potenziali utenti [...] l'uso di codici standard [...] aumenta la trasparenza nelle pubbliche procure, rende [possibile] identificare le opportunità di business»[45][p.3] per le aziende, inoltre semplifica per le pubbliche amministrazioni la possibilità di monitorare il proprio mercato estraendo statistiche quantitative, permettendo anche di individuare le aree in cui sono

effettuate le spese della stazione appaltante.

Precedentemente al vocabolario CPV la classificazione proposta per monitorare gli acquisti a livello mondiale era il CPC - *Central Product Classification* - e la sua controparte europea era la *European Classification of Economic Activities*, NACE. *«Queste due classificazioni [NACE e CPC] possono essere considerate come le basi su cui è stato costruito CPA [Classification of Products by activity]»*[45][p.3] . L'ultima versione di CPA risale al 1992, mentre il CPV compare nella sua prima stesura l'anno successivo ed è stato - fino al 2008 - costantemente aggiornato nel tempo perché soddisfacesse le richieste di mercato dei tempi.

La versione attualmente in uso del vocabolario degli appalti pubblici risale al 2008, sono presenti nove cifre, di cui una di controllo, anziché le iniziali sei. Come è possibile notare dalla figura 3.1, la prima cifra del codice CPV individua la divisione, la seconda il gruppo, la terza la classe, la quarta classifica la categoria e dalla quinta si identificano le sottocategorie di appartenenza.

Division	35000000-4	Other transport equipment
Group	35100000-5	Ships and boats
	...	...
Class	35110000-8	Ships
Category	35112000-2	Ships and similar vessels for the transport of persons or goods
Sub-category	35112100-3	Cruise ships, ferry boats and the like, primarily designed for the transport of persons
	35112110-6	Ferry boats
	35112180-7	Cruise or excursion boats n.e.c.
	35112200-4	Tankers

Figura 3.1: Livelli delle classi CPV[45][p.5]

«L'ente appaltante dovrebbe provare a trovare un codice che soddisfa i propri bisogni in modo quanto più accurato possibile. Ovviamente più di un codice potrebbe essere usato nei form standard di pubblicazione»[45][p. 9], ma se nessun codice dovesse soddisfare i requisiti sarebbe possibile assegnarne uno gerarchicamente più in alto (con più cifre a 0). Nel caso in cui si avesse un contratto con più categorie merceologiche aggregate, allora si potrebbe assegnare anche il codice preponderante in termini di importo.

3.4.1 Il mercato su ContrattiPubblici.org e i CPV

Su ContrattiPubblici.org, dei circa sessanta milioni di contratti indicizzati, soltanto sette, il 12,2%, hanno un codice CPV allegato. Alcuni contratti[46][Martiello et al, 2019] sono esenti dalla pubblicazione del CIG, tra questi possiamo annoverare *«i contratti di lavoro conclusi dalle stazioni appaltanti con i propri dipendenti[...] il trasferimento di fondi da parte delle amministrazioni dello Stato in favore di soggetti pubblici, se relativi alla copertura di costi per le attività istituzionali espletate dall'ente[...] gli affidamenti diretti a società in house»*[46][Martiello et al, 2019].

Inoltre, per alcuni contratti è possibile richiedere ad ANAC lo smartCIG.

«Lo smart-CIG è un CIG con procedura semplificata, realizzato unicamente per garantire la tracciabilità degli appalti di importo inferiore alla soglia di sottoposizione degli obblighi informativi e contributivi verso l'ANAC, per i quali cioè non è richiesto il CIG, nonché per le fattispecie che non rientrano in questi obblighi comunicativi e contributivi così come definiti dai vari Comunicati dell'Autorità. Esso è stato pensato, quindi, principalmente per individuare gli appalti che altrimenti sarebbero stati privi di identificazione specifica ai fini della tracciabilità dei flussi finanziari»[47][ANAC, 2020].

Quindi *«lo smart CIG è stato introdotto per semplificare la procedura di ottenimento del codice degli appalti di minor interesse per l'Autorità (i "micro-contratti"), ed in particolare [...] per i contratti di lavori, servizi e forniture,*

inclusi i contratti di cui agli artt. 17 (Esclusioni specifiche per contratti di appalto e concessione di servizi) e 19 (Contratti di sponsorizzazione) e [...] di importo inferiore a 40.000 euro [...] [e] per i contratti di cui agli articoli 7 (Appalti e concessioni aggiudicati ad un'impresa collegata), 16 (Contratti e concorsi di progettazione aggiudicati o organizzati in base a norme internazionali) e 162 (Contratti secretati) del Codice dei contratti, indipendentemente dall'importo. In questi casi il codice CIG serve solo ai fini della tracciabilità, quindi andrà riportato su tutte le lettere, bandi, mandati ecc., mentre non sarà necessaria alcuna ulteriore comunicazione all'ANAC»[48][Ruggiero, 2021]. Sebbene la soglia è riportata a 40.000 euro, questa è soggetta a variazioni in seguito a modifiche di legge, di fatto è stata alzata durante la pandemia dovuta al coronavirus per snellire la burocrazia degli appalti pubblici.

Per riuscire a dare un'idea di quanti siano i contratti sprovvisti di CPV, basti pensare che la categoria merceologica può non essere indicata nei contratti con SMART CIG - il 33-50% degli importi liquidati dalle stazioni appaltanti- e che per i contratti assegnati tramite procedura per affidamento diretto - l'83% - non è presente un campo relativo al codice di categoria merceologica.
Ulteriormente è possibile che qualche funzionario alleghi al bando pubblicato dall'ente appaltante un codice dal vocabolario CPV, tuttavia, si tratta davvero di inezie rispetto al volume d'affari collegato alle pubbliche amministrazioni.

Dato che la spesa pubblica convoglia una quantità sempre maggiore di capitale nel corso degli anni, come dimostra il grafico sul mercato delle pubbliche procure in Italia in Figura 3.2, il monitoraggio delle aree di spesa potrebbe costituire un punto di interesse per gli enti pubblici ai fini dell'ottimizzazione della spesa.
Essendo dati aperti, il mercato delle pubbliche amministrazioni è uno dei pochi che si presta a fare *business intelligence* in modo trasparente, confrontando i prezzi degli appalti precedenti.

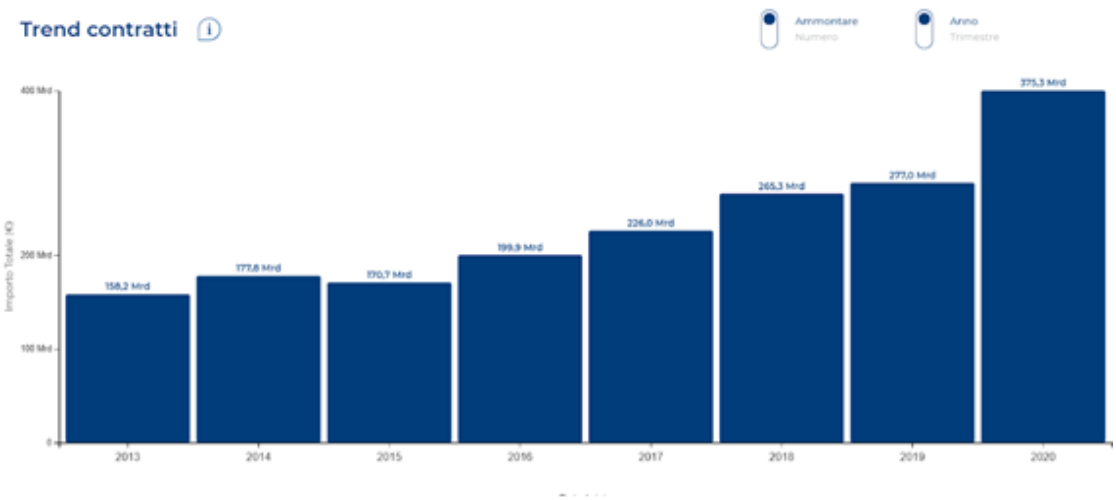


Figura 3.2: Importi erogati per i contratti annualmente, ContrattiPubblici.org

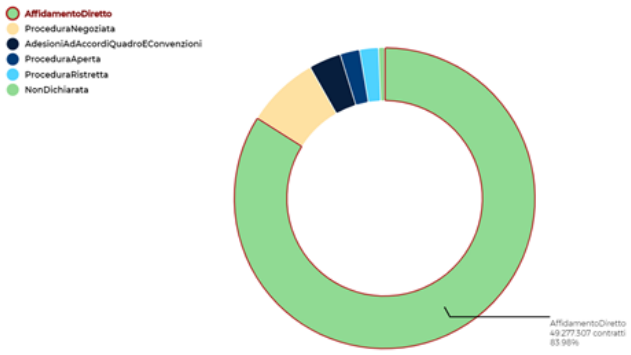


Figura 3.3: Percentuale affidamenti per numero di contratti, ContrattiPubblici.org

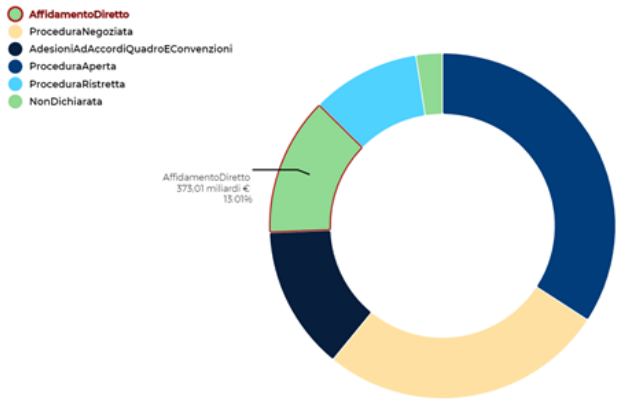


Figura 3.4: Percentuale affidamenti contratti per importo erogato, ContrattiPubblici.org

La componente di contratti assegnati tramite affidamento diretto rende difficile l'aggregazione dei contratti per CPV, in quanto per questi contratti non è previsto un Codice Identificativo di Gara.

Questo si traduce in prima approssimazione in un'assenza sistematica del CIG e, successivamente, nell'assenza del codice CPV, con cui Synapta aggrega i contratti della pubblica amministrazione per utilizzare proficuamente i dati pubblicati per il fine della trasparenza.

Inoltre, è stato notato che i codici definiti dal vocabolario CPV vengono generalmente assegnati dai funzionari delle stazioni appaltanti nel momento in cui l'importo erogato di un contratto diventa considerevole.
A partire delle Figure 3.3 e 3.4, è possibile notare come i contratti assegnati tramite affidamento diretto costituiscano l'83,98% del numero di contratti totali, ma rappresentino soltanto il 13,01% in termini di importo liquidato dalle stazioni appaltanti; inoltre l'importo medio dei contratti con CPV in ContrattiPubblici.org è 317mila euro, mentre l'importo medio dei contratti senza CPV è 11mila euro. Per questi ultimi, in genere, viene scelto l'affidamento diretto come tipologia di gara.

I contratti che hanno altre procedure di affidamento, in genere prevedono un importo erogato maggiore di quello richiesto nel caso dell'affidamento diretto.

Capitolo 4

Classificazione ibrida in ContrattiPubblici.org

Prima di illustrare lo studio, bisogna investigare - seppure ad alto livello - alcune problematiche attualmente presenti in ContrattiPubblici.org che potrebbe - in maniera appropriata - essere definito come "il Google dei contratti pubblici", sebbene aggiunga delle altre funzionalità al motore di ricerca.

Quasi nessun contratto stipulato dalla stazione appaltante ha in allegato un CPV. È di tutta evidenza, però, che per rendere possibile la *business intelligence* è necessario aggregare i contratti per categoria merceologica di appartenenza.

4.1 Business Intelligence

«La business intelligence (BI) si riferisce alle capacità che consentono alle organizzazioni di prendere decisioni migliori, intraprendere azioni informate e implementare processi aziendali più efficienti.

Le potenzialità di BI consentono di:

- *Raccogliere dati aggiornati dalla propria organizzazione*;
- *Presentare i dati in formati di facile comprensione (come tabelle e grafici)*;
- *Fornire i dati in modo tempestivo ai dipendenti della propria organizzazione; La BI mantiene [...] [l'] organizzazione al corrente degli sviluppi e il successo dipende in gran parte dalla conoscenza di chi, cosa, dove, quando, perché e come degli eventi del mercato. Quanto sono popolari i tuoi prodotti e i tuoi servizi tra i consumatori? Cosa stanno facendo i tuoi concorrenti? Perché i consumatori scelgono un brand anziché un altro? Come e quando cambierà il mercato? Quali sono le tendenze per il futuro?*»[49][Oracle].

Il centro nevralgico della business intelligence è rappresentato dai dati: è di tutta evidenza che bisogna comprendere i dati per estrarre delle informazioni da questi ultimi. In breve, *«prima che i dati possano davvero essere utili bisogna spesso prendersi carico di una serie di costose operazioni preliminari. Una metafora utile a capire quali siano i passaggi per trarre veramente valore dagli open data prende spunto dalla piramide di Maslow»*. [15][Morando, 2020].

Sviluppata dall'omonimo psicologo statunitense, la piramide di Maslow ha l'obiettivo di classificare i bisogni intrinseci alla natura umana, a partire da quelli fisiologici - che si trovano alla base della piramide - all'autorealizzazione che, invece, si trova in cima a essa.

«Anche per i dati, è difficile fare analisi avanzate o generare nuova conoscenza, se prima non ci si prende cura dei "bisogni" basilari dei dati. [...] i dati vanno prima raccolti e puliti. Si potrebbe rappresentare, quindi, la piramide del valore dei dati in questo modo:

- *Conoscenza: l'illuminazione alla cima della piramide, in cui i dati riescono a dare veramente un valore per chi li sta analizzando*;
- *Apprendimento e ottimizzazione: i dati diventano informazioni "azionabili"*;
- *Arricchimento: i dati vengono etichettati, e arricchiti in caso di mancanze*;
- *Gestione della qualità: i dati vengono gestiti in termini di qualità e pulizia*;
- *Gestione dei flussi e Collezione: infrastruttura minima per acquisire e spostare in modo affidabile ed automatizzato i dati*»[15][Morando, 2020].

La piramide di Maslow, quindi, se applicata ai dati, potrebbe essere rappresentata come in Figura 4.1.

Figura 4.1: Piramide di Maslow, ContrattiPubblici.org

Per provare a quantificare l'abnormità di dati generati, basti pensare che nel 2018 sono stati generati giornalmente circa 2.5 quintilioni (1030) di byte di dati[50].

«La BI rappresenta il cuore di ogni impresa data-driven, il che la rende l'epicentro della trasformazione. Aumentare l'impatto di un'organizzazione e renderla più efficiente sono gli obiettivi finali dell'implementazione di un nuovo strumento di BI; tuttavia, con la giusta tecnologia BI, è possibile ottenere anche numerosi vantaggi aggiuntivi [come] migliorare la precisione dei dati, prendere decisioni migliori più rapidamente [...] eliminare sprechi, frodi e abusi [e] migliorare la produttività»[49][Oracle].

4.2 La classificazione in ContrattiPubblici.org

Nel caso di ContrattiPubblici.org, i contratti e i bandi vengono analizzati tramite un algoritmo di machine learning che li classifica.
L'algoritmo di machine learning attualmente in produzione è costruito tramite un modello basato su bag of word, che viene costruita a tramimite le parole più frequenti e rimuovendo le *stopwords*, ovvero parole semanticamente inutili alla classificazione. Si tratta di termini che non contribuiscono ad aggiungere un significato al contratto che viene preso in esame; in altre parole le *stopwords* sono parole ricorrenti a molti contratti o bandi che - però - non sono utili a determinare la tipologia del servizio, dell'opera, o della fornitura richiesta dalla stazione appaltante.
Ad esempio, le parole «bando», «lotto» o «gara», seppure sovente presenti all'interno del corpus del contratto, non concorrono a determinare la categoria di appartenenza merceologica del contratto.

Inoltre, le parole più corte di quattro caratteri vengono scartate e assimilate alle stopwords: non concorrono alla determinazione della classe di appartenenza del contratto.
Moltissimi bandivengono classificati grazie alla verbosità considerevole e

all'interno del loro *corpus* sono presenti molte parole che lo descrivono; di contro, i contratti vengono scartati più spesso poiché non contengono un'elevata espressività e può capitare che nel *corpus* di un contratto siano presenti poche parole e corte.

4.3 Perché approccio ibrido

In questo contesto si parla di approccio ibrido di classificazione perché al classificatore di machine learning viene affiancato un insieme di *regular expression* utile a classificare quei contratti altamente specifici per una classe.

Le regex - regular expression - servono a individuare univocamente la categoria merceologica a cui appartiene un contratto. Sono come sottostringhe che devono essere individuate all'interno dell'oggetto di un contratto.
Risultano utili perché il classificatore in produzione non riesce a classificare alcuni contratti, quelli con poche parole. Lo scopo del sistema di regole è quello di taggare i contratti con il loro codice CPV, scendendo verticalmente nella gerarchia ad albero del sistema di tassonomia.

Se c'è un hit, ovvero se la sottostringa argomento della regex trova un match con l'oggetto del contratto, il contratto viene classificato secondo la categoria merceologica che viene assegnata dalla regex.
Diversamente, il contratto non viene classificato e quindi è necessario passare dal classificatore per ottenere un codice CPV, sempre che ne sussistano le condizioni.

Viene altresì implementato un classificatore in grado di assegnare le classi di secondo e di terzo livello agli appalti delle pubbliche amministrazioni.
In cascata al classificatore attualmente in produzione, che assegna ai bandi la classe di primo livello dei CPV, viene posto il nuovo classificatore sviluppato, il cui compito è quello di individuare la classe di secondo o terzo livello da assegnare al contratto, discriminando a partire dal valore prodotto in uscita

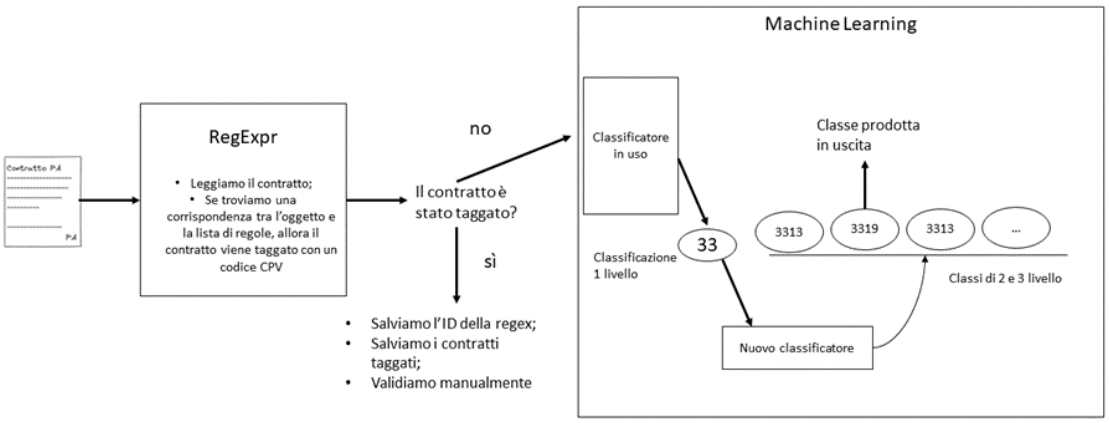


Figura 4.2: Classificatore ibrido

dal classificatore di primo livello.
Con questa classificazione si vogliono minimizzare i *falsi positivi*, incrementando la *precision*. Il focus è quello di classificare gli appalti nel modo più accurato possibile, sacrificando tuttavia il numero di contratti classificati.
Un approccio del genere risulta utile perché permette *«di utilizzare un classificatore statistico, basato su machine learning, e un classificatore deterministico»*[1], basato sulle regex.
La presenza di un dizionario di regole, inoltre, permette di poter avere una panoramica più dettagliata e granulare delle aree di interesse.

[1]D. Allavena, corrispondenza personale.

Capitolo 5

Le Regular Expression

«Una regular expression - spesso abbreviata come regex o regexp - è una sequenza di caratteri che specificano un pattern di ricerca all'interno di un testo. Di solito questi pattern sono usati da algoritmi di ricerca per stringhe al fine di trovare e [opzionalmente] rimpiazzare la sottostringa ricercata o per la validazione dell'input». Le regex, ad esempio, sono spesso utilizzate nel momento in cui è necessario validare il formato di una mail, quindi bisogna controllare la presenza di alcuni caratteri come '@'; o ancora la validazione delle password, dato che spesso è richiesta la presenza di lettere maiuscole, minuscole, segni di interpunzione o simboli più in generale.
In breve, viene verificato che il pattern in argomento alla regola sia contenuto nella stringa esaminata.

La ricerca di questi caratteri speciali all'interno, per esempio, di un form viene effettuata tramite le regex.
Concettualizzate negli anni '50 dal matematico americano S.C. Kleene che formalizzò la descrizione di un linguaggio regolare, le regex si avvalgono di due diverse sintassi: il POSIX standard e la sintassi Perl.

I caratteri all'interno delle stringhe che compongono una regex possono assumere un significato letterale o - se combinati con alcuni simboli - possono rappresentare dei metacaratteri.

È utile riportare alcuni metacaratteri che è possibile usare con le regular expression:

- '?': indica la presenza opzionale di una lettera o di un gruppo. La stringa 'colou?r' permette di trovare una corrispondenza sia con la stringa 'color' che con 'colour';

- '|': indica una funzione 'or' booleana, e.g. 'casa|case' permette di trovare una corrispondenza con il sostantivo sia al singolare sia al plurale;

- '[]': le lettere presenti all'interno delle parentesi quadre sono in or tra loro, con la regex 'cas[ae]' otteniamo un comportamento identico a quello di sopra;

- '.': è il jolly all'interno del mondo delle regex, qualsiasi carattere viene individuato come appartenente ad una regex in presenza di un punto;

- '+': indica una o più occorrenze della lettera che precede il simbolo, la regex 'e+' troverà tre corrispondenze con la parola 'veemenza', cioè le prime due vocali e, successivamente, la terza;

- '^': indica l'operatore not;

- '-': in combinazione con alcuni caratteri, può indicare un insieme di lettere o numeri, '[a-z]' permette di trovare tutte le lettere minuscole.

Inoltre, è possibile settare alcuni flag, perché la regex individui, ad esempio, sia lettere maiuscole che minuscole oppure la sottostringa di ricerca su più righe.

5.1 Vantaggi delle regex nella classificazione ibrida

Nel contesto di una classificazione ibrida, le regex hanno un ruolo rilevante: queste permettono di superare i vincoli posti dall'algoritmo di intelligenza artificiale.
Alcuni testi non contenfono informazioni e/o parole sufficienti per poter essere classificate tramite machine learning. L'insieme di regole permette, inoltre, di poter creare campagne di classificazione ad-hoc per gli specifici clienti riguardo gli ambiti merceologii in cui abbondano termini tecnici, commerciali, o parole ricorrenti specifiche.

Nel caso di ContrattiPubblici.org, il classificatore, prima di processare il contratto o il bando, effettua alcuni controlli riguardanti le parole che compongono il *corpus* del contratto. Vengono rimosse le parole che contengono meno di quattro lettere e le le *stopwords*, ovvero parole che non aggiungono significato semantico al contratto.
Se il contratto così epurato avrà almeno quattro parole sarà processato dal classificatore.

Le *stopwords* sono delle parole presenti in una stoplist o dizionario negativo che vengono scartate prima di processare i dati testuali. Non esiste una lista di stopword applicabile a tutte le lingue, ragione per cui ogni parola può essere scelta come stopword all'interno di un problema di classificazione. Alcune stopword appartenenti alla lingua italiana potrebbero essere gli articoli e le preposizioni, in quanto sebbene portino con sé un significato, non sono utili al fine di determinare la classificazione del testo, un discorso analogo potrebbe essere fatto con i mesi dell'anno: se una fornitura di prodotti alimentari è presente nell'oggetto del contratto, allora questa fornitura rimarrà tale indipendentemente dal mese in cui viene erogata e verrà assegnato al contratto un codice di categoria merceologica uguale a '15000000-8'. Allo stesso modo, se un ente aggiudicatario vince un bando pubblicato da una stazione

appaltante, è trascurabile - sempre ai fini della classificazione - sapere se la pubblica amministrazione ha affidato il suo contratto tramite procedura ristretta o affidamento diretto, in quanto il punto nevralgico dello studio è classificare il CPV di appartenenza del contratto.

In breve, sono molteplici le ragioni per cui un contratto può non essere processato. In questi casi, è molto utile avere un sistema di regex in grado di classificare in modo univoco il contratto.

Effettuando una ricerca su ContrattiPubblici.org, con la stringa "acquisto medicinali", vengono restituiti più di 16mila contratti che hanno come oggetto queste due parole. Questi contratti - se sprovvisti di un codice di categoria merceologica - non vengono classificati dalla parte di machine learning, in quanto il contratto non presenta il numero minimo di parole richieste per poter essere processato.

L'approccio basato su regex tenta di superare i limiti del machine learning, fornendo una classificazione verticale dei contratti tramite il match con il pattern della regular expression.
Ciò permette di migliorare le performance di classificazione: si può intervenire sui contratti che non è possibile analizzare con il machine learning.
Nel caso di ContrattiPubblici.org, attualmente si perde la possibilità di classificare circa il 23,6% dei contratti; la percentuale scende all'1.6% in riferimento ai bandi.

5.2 Osservazione della piattaforma e definizione di un raggio di azione

Il primo, propedeutico, passo prima di scrivere le regex consiste nell'osservazione della piattaforma e delle categorie CPV presenti. Ci si è resi conto che la scrittura di un insieme di regole che coinvolge tutte le categorie individuate dal vocabolario dei CPV sarebbe stata impossibile per due principali motivi:

1. Le categorie CPV individuano 45 divisioni, ovvero classi di primo livello. Scendendo in verticale nella classificazione, includendo i gruppi, i sottogruppi, si contano un totale di più di 9mila classi. Questo renderebbe il lavoro di scrittura delle regole mastodontico, sia in termini di sforzo, sia in termini di tempo necessario per eseguirlo;

2. La seconda motivazione, puramente pratica, risiede nella mancanza di conoscenze e competenze che abbracci tutte le classi: i codici di categoria merceologica si estendono dalla fornitura di cibo ai servizi fognari, passando per servizi di consulenza, acquisto di software e realizzazione di lavori. Appare di tutta evidenza che una competenza così ampia non può essere attribuita né posseduta da un singolo individuo.

La definizione di un perimetro di azione è stato un passaggio obbligato. A seguito di un'analisi della piattaforma attraverso il CPV sunburst è emerso come la maggior parte dei contratti con codice CPV appartenesse alla classe 33000000-0, *Apparecchiature mediche, prodotti farmaceutici e per la cura personale.*

Quindi è stato deciso di concentrarsi sulla divisione 33 e sulla 72, riguardanti rispettivamente *apparecchiature mediche, prodotti farmaceutici e per la cura personale* e *servizi informatici: consulenza, sviluppo di software, Internet e supporto.*
La prima classe è stata selezionata perché abbondantemente rappresentanta nella piattaforma, come si può notare dalla Figura 5.1; la seconda classe si è scelta per la conoscenza della materia abbastanza buona da parte del candidato e poiché il mercato relativo ai servizi informatici è in forte espansione e costituisce una parte importante delle spese effettuate dalle stazioni appaltanti. Secondo le stime AGID [51][AGID, 2021] la spesa ICT della pubblica amministrazione è cresciuta del 7% in ogni anno tra il 2018 e il 2020, passando da 230 a 270 miliardi di euro e secondo le stime Confindustria

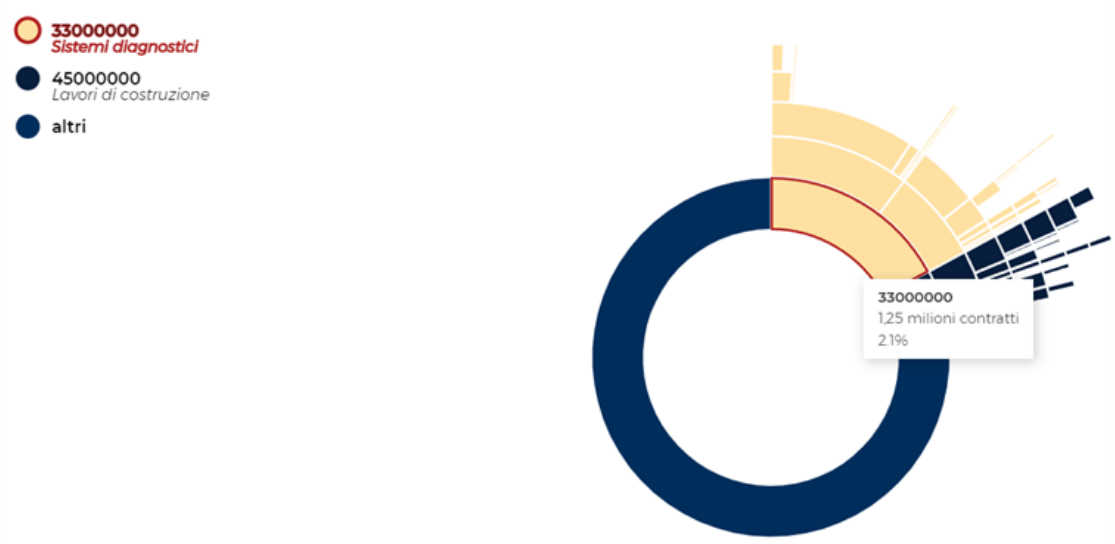


Figura 5.1: Distribuzione contratti per codice CPV, ContrattiPubblici.org

in [52] i servizi informatici rappresenteranno una porzione di mercato sempre crescente per il triennio 2021-2024.

È bene ricordare che dei circa 60 milioni di contratti presenti su ContrattiPubblici.org, solo 7 presentano un CPV allegato: circa l'88% dei contratti è sprovvisto di codice CPV.

Nel momento in cui si effettua una ricerca sulla piattaforma, tra le informazioni riportate è presente la tag cloud, che riporta le informazioni riguardanti la distribuzione delle parole presenti nei contratti per i quali è stata eseguita la ricerca.

Esiste la possibilità di visualizzare la tag cloud sia per la frequenza di parole nel corpus del contratto, sia per l'ammontare dello stesso: *«la dimensione del carattere rappresenta il numero di contratti relativi alla tematica, presenti nella ricerca corrente [...] [è possibile] passare dalla rappresentazione fatta*

in base al numero dei contratti, a quello in base alla somma degli importi»[1]. In questo modo è possibile analizzare i contratti scendendo verticalmente sulla gerarchia dei CPV per trovare quei contratti - sprovvisti di codice CPV - rappresentativi di una ricerca e/o di una specifica.

Dalle due tag cloud riportate, le quali si riferiscono alla classe 3369, è possibile individuare alcune parole che potrebbero tornare utili ai fini della classificazione e alcune che - invece - non danno alcun contributo semantico in relazione alla classificazione del contratto.
A partire dalle tag cloud, sono state esplorate le parole maggiormente significative e rappresentate all'interno della ricerca al fine di individuare dei pattern ricorrenti nei contratti.

Figura 5.2: Tag cloud per il codice CPV 3369, ContrattiPubblici.org

Ovviamente, non è possibile analizzare con una copertura completa una

[1]Descrizione tag cloud da ContrattiPubblici.org

classe; preliminarmente è necessario individuare dei pattern ricorrenti all'interno di una determinata classe, dopodiché bisogna capire se il numero di contratti associati al pattern è considerevole: se una stringa, per esempio, rappresenta 50 contratti, il suo apporto all'analisi in questione è praticamente trascurabile.

D'altro canto, se sulla piattaforma si effettua una ricerca impostando il codice di categoria merceologica "33600000-6 Prodotti farmaceutici" e si inserisce come query di ricerca la stringa "fornitura farmaci", sono restituiti quasi 78mila risultati. Rimuovendo il filtro sul CPV, i risultati aumentano a 306 mila contratti.

Da questo dato si deduce la presenza di 228 mila contratti senza un codice CPV; poiché si tratta di un numero considerevole di contratti, allora sarebbe vantaggioso che quella stringa di ricerca diventasse una regex.

5.3 Scrittura delle regole

Dopo l'osservazione della piattaforma e delle due divisioni da analizzare, si è passati alla stesura delle regole.

Il primo nodo da sciogliere è quello della struttura dati da utilizzare. L'informazione che sarebbe tornata utile ai fini della classificazione è stata individuata nel codice CPV di appartenenza della regular expression e la relativa descrizione. Questo dizionario - relativo a una singola regex - viene accorpato ad altri tramite una lista; quindi, l'oggetto su cui viene effettuata l'iterazione è una lista di dizionari.

La prima struttura dati implementata è quindi composta come mostrato in Figura 5.3.

Con questa struttura, però, non si riesce a discernere se l'oggetto del

```
rules = [

    {
        "ID": "0",
        "regexValue": [
            re.compile("(set|kit)(?:.+)(c[ /]?pap(?:[^a-z]))", re.IGNORECASE),
            re.compile("(acquist[oi]|fornitur[ae])(?:.+)(set|kit)(?:.+)(c[ /]?pap(?:[^a-z]))",
               re.IGNORECASE),
        ],
        "cpvCode": "33157200-7",
        "codeName": "Kit respiratori",
        "cpvName": "Kit respiratori",
    },

…

]
```

Figura 5.3: Prima versione di un elemento del dizionario di regole

contratto è pienamente inerente al pattern compilato dalla regex.
Se fosse presente un contratto che aggrega il casco cpap, i medicinali e altre apparecchiature mediche, non potremmo discriminare l'effettivo CPV che dovrebbe essere assegnato al contratto. Inoltre è stato rilevato come in alcuni insiemi di contratti, la voce "importo liquidato" riesca a discriminare tra due tipologie di forniture.

Per cui, è stato necessario aggiungere alcuni filtri che rendessero il match con i contratti esaminati più accurato. Il primo filtro implementato è quello relativo alla copertura della regex rispetto all'oggetto del contratto esaminato.

Prima di poter implementare la copertura, bisognava processare il testo del contratto rimuovendo alcune stopwords. Le stopwords che si è cercato di eliminare sono parole inerenti al mese in cui viene erogata la fornitura e la durata della fornitura, il lavoro o il servizio, gli articoli, le preposizioni - sia semplici che articolate -, le indicazioni riguardanti la stazione appaltante (e.g. "comune", "azienda", "ospedaliera") e alcune parole ricorrenti per ogni bando (e.g. "lotto/i", "contratto").

A questo punto, il contratto così modificato deve essere duplicato: è necessario effettuare due diverse operazioni per ottenere il punteggio di copertura: bisogna ottenere due oggetti diversi, l'oggetto depurato e il contratto depurato. Per ottenere il denominatore, il contratto depurato, sono state rimosse le stopwords e i segni di interpunzione dal contratto in ingresso.

Per ottenere l'oggetto depurato, invece, sono stati sottratti i gruppi matchati dalle regex.

La copertura è stata definita come:

$$\left| \text{copertura} = \quad 1 - \frac{len(oggetto\_depurato)}{len(contratto\_depurato)} \right|$$

È stato necessario trasformare le sottostringhe nel pattern delle regex in gruppi, la prima regular expression mostrata in 5.3 precedente viene trasformata come mostrato in Figura 5.4:

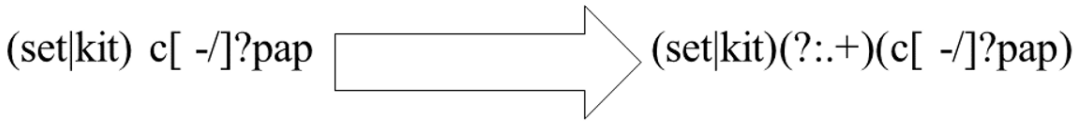


Figura 5.4: Trasformazione regex

nel caso in cui tra 'set' o 'kit' ci fosse stata qualche altra parola, la regex non avrebbe dato nessuna corrispondenza. Per cui si è intervenuti inserendo i non-capturing group, ovvero la stringa *(?:.+)*. Per spiegarne l'utilità risulta funzionale osservare e commentare le immagini in Figura 5.5 e 5.6:

Figura 5.5: Simulazione corrispondenze tra regular expression e stringhe

La prima stringa non viene matchata dalla regex, ma l'intento era quello di avere un match indipendentemente dai caratteri presenti tra le parole "set" e "cpap". I non-capturing group permettono di ignorare eventuali sottostringhe presenti tra le due ricercate dalla regex e vengono indicati con *(?:.+)*.

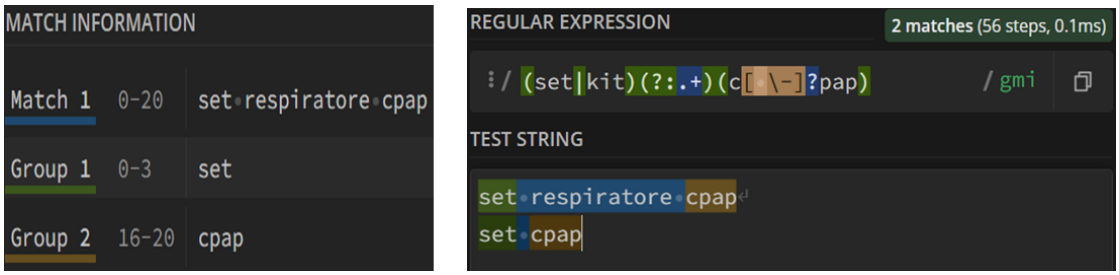


Figura 5.6: Simulazione corrispondenze tra regular expression e stringhe con non-capturing group

In questo modo, vengono trovate delle corrispondenze anche con la prima stringa, ma la parola "respiratore" non viene considerata nel computo dei gruppi matchati.

Quindi, dal contratto esaminato vengono rimossi i gruppi corrispondenti ai pattern contenuti nelle regex; successivamente esso viene *tokenizzato* per rendere possibile la rimozione delle stopword, viene applicata una *join* tra le parole rimanenti e viene - a questo punto - calcolata la lunghezza della stringa rimanente, la quale sarà il numeratore all'interno della formula per calcolare la copertura, l'*oggetto depurato*.

Si è assegnato un valore di copertura minimo inversamente proporzionale alla lunghezza della regex. La nuova struttura delle regole è riportata in Figura 5.7.

```
rules = [

    {
        "ID": "0",
        "regexValue": [
            re.compile("(set|kit)(?:.+)(c[ /]?pap(?:[^a-z]))", re.IGNORECASE),
            re.compile("(acquist[oi]|fornitur[ae])(?:.+)(set|kit)(?:.+)(c[ /]?pap(?:[^a-z]))",
               re.IGNORECASE),
        ],
        "cpvCode": "33157200-7",
        "codeName": "Kit respiratori",
        "cpvName": "Kit respiratori",
        "copertura": 0.8,
    },

…

]
```

Figura 5.7: Seconda versione di un elemento del dizionario di regole.

Le prestazioni, in termini di accuratezza della classificazione, migliorano con questa struttura dati; si è pensato di poter ulteriormente affinare la classificazione aggiungendo un intervallo di prezzo al contratto.

Tra le statistiche che ContrattiPubblici.org mostra agli utenti che effettuano una ricerca è presente la sezione Mercato, che *«mostra la distribuzione dei contratti stipulati in base al prezzo. Sull'asse orizzontale sono mappati - in*

scala logaritmica - gli importi dei contratti. Sull'asse verticale è rappresentato il numero dei contratti stipulati per ogni determinata fascia di prezzo»[2].

Attraverso il grafico in Figura 5.8, che rappresenta l'importo erogato per numero di contratti, e quello di dispersione, che mostra la correlazione tra la somma degli importi dei contratti e il numero dei contratti stipulati, si è potuto stimare l'intervallo di importo entro cui collocare i contratti esaminati per la ricerca, in modo da eliminare gli *outlier*.

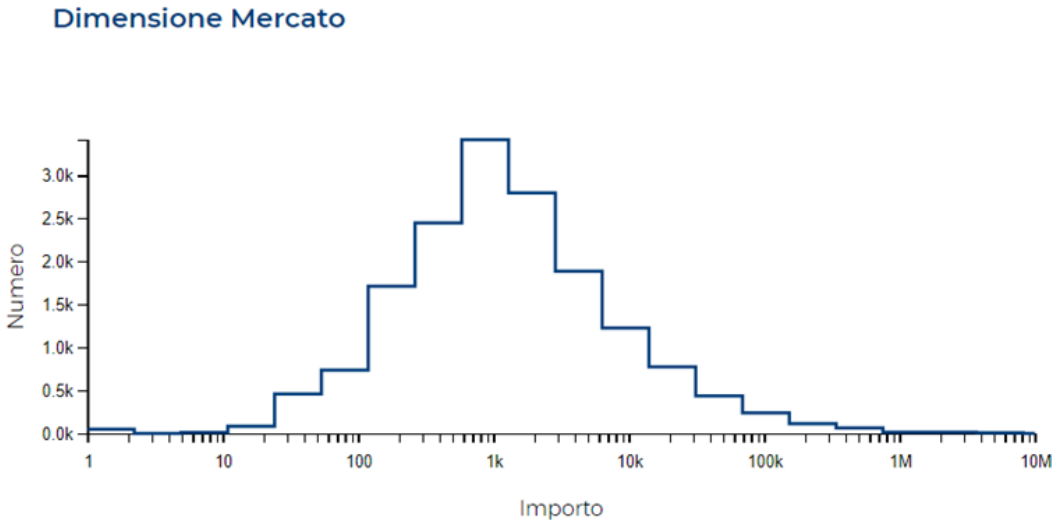


Figura 5.8: Importo relativo ai contratti per una ricerca effettuata, ContrattiPubblici.org

Dopo avere aggiunto il filtro sull'importo erogato, la struttura finale diventa come mostrato in Figura 5.9

[2]ContrattiPubblici.org, spiegazione della sezione Mercato.

```
rules = [

    {
        "ID": "0",
        "regexValue": [
            re.compile("(set|kit)(?:.+)(c[ /]?pap(?:[^a-z]))", re.IGNORECASE),
            re.compile("(acquist[oi]|fornitur[ae])(?:.+)(set|kit)(?:.+)(c[ /]?pap(?:[^a-z]))",
               re.IGNORECASE),
        ],
        "cpvCode": "33157200-7",
        "codeName": "Kit respiratori",
        "cpvName": "Kit respiratori",
        "importoA": 500,
        "importoDa": 30000,
         "copertura": 0.8,
    },

…

]
```

Figura 5.9: Terza versione di un elemento del dizionario di regole.

Con i contratti esaminati attraverso questa struttura è stata effettuata l'ultima validazione.

5.4 Problemi implementativi

Durante la fase di scrittura delle regex le prestazioni non erano soddisfacenti: i tempi di esecuzione delle regex si dilatavano; è stato perciò implementato un sistema per monitorare il tempo di processamento impiegato da una regola.

Se indichiamo con N il numero di contratti e con R il numero di regole scritte, la complessità dell'algoritmo è *O(N*R)*, perché ogni contratto viene confrontato con tutte le regole per trovare una corrispondenza tra la sottostringa ricercata e l'oggetto del contratto; ovviamente, se più di una regex trova una corrispondenza, verrà selezionata quella con il punteggio di copertura massimo.

Perciò è stato eseguito il codice su un campione casuale di mille contratti ed è stato monitorato il tempo in cui ogni singola regex veniva processata, ottenendo un tempo medio relativo a mille iterazioni.
Analizzando i tempi di esecuzione di ogni regola, è stato osservato come le regex più onerose in termini di tempo e computazionali erano quelle con molti gruppi opzionali, perché in un gruppo opzionale è considerata sia la presenza sia l'assenza del gruppo in questione: ogni gruppo opzionale può essere o meno presente all'interno di un contratto, da ciò discende che - dati N gruppi opzionali - la regex potrà avere $2^N valori$.

Supponendo di avere la regex:

$$(set|kit)?(?:.+)(c[/]?pap(?:[\hat{}a-z]))$$

la si potrebbe espandere nel seguente modo:

- $(set|kit)(?:.+)(c[/]?pap(?:[\hat{}a-z]))$;
- $(c[/]?pap(?:[\hat{}a-z]))$.

Per correggere questo comportamento e avere dei tempi di processamento soddisfacenti, le regex aventi più di 4 gruppi opzionali sono state sdoppiate; in questo modo, nel corso di un'altra simulazione - sempre su un campione di mille contratti - è stato osservato un miglioramento sostanziale delle prestazioni.

I due grafici in Figura 5.10 mostrano sull'asse orizzontale l'id della regola esaminata, lungo l'asse verticale vengono indicati i tempi di esecuzione delle regole in scala logaritmica, la linea rossa indica la media e i tempi sono riportati in microsecondi. Per poter rendere efficacemente l'idea dell'ottimizzazione raggiunta, è possibile considerare il campione di dati su cui è stata eseguita la classificazione, circa 430 mila contratti. La variazione media di tempo di esecuzione è pari a:

$$\Delta t = (2095.322 - 128.059)ms = 1967.263ms$$

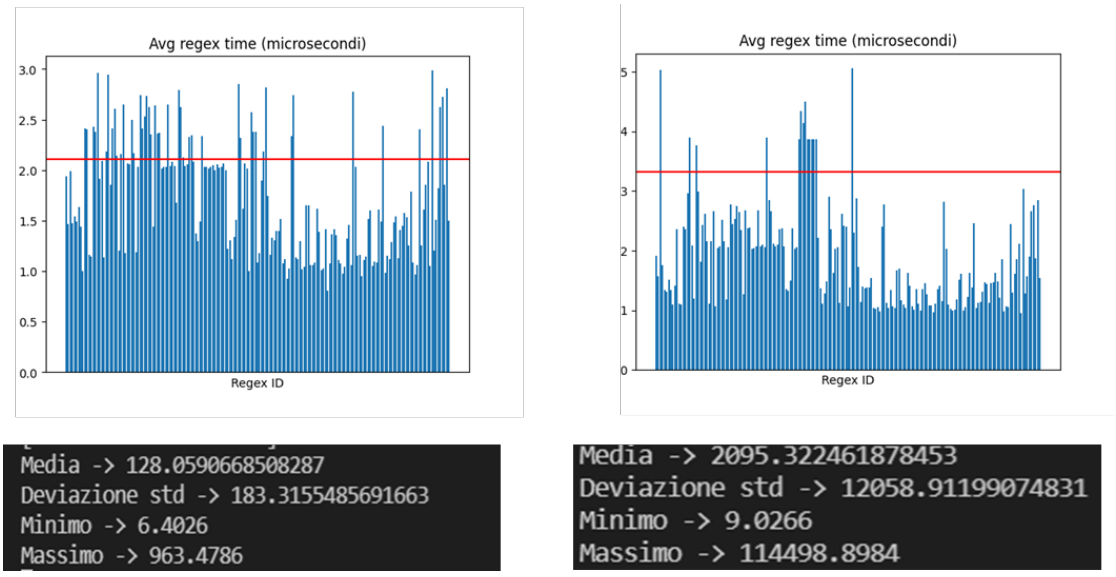


Figura 5.10: Prestazioni delle regex.

Questa variazione, moltiplicata per il numero di contratti analizzati, fa in modo che si riescano a risparmiare nove giorni circa:

$$\Delta T = 1.967263s * 430'000 = 845'923s \sim 9giorni.$$

5.5 Validazione e statistiche comparative

L'operazione di taggatura dei contratti è stata effettuata su un campione di 430 mila contratti aventi un CPV assegnato dalla stazione appaltante. Il campione estratto rappresenta la reale distribuzione dei CPV su ContrattiPubblici.org.

Come sostenuto e dimostrato da diversi studi, le informazioni pubblicate dalle stazioni appaltanti sono spesso insoddisfacenti rispetto agli standard di

qualità dei dati definiti dall'ISO, per cui si è resa necessaria la validazione manuale, che risulta inevitabile anche in relazione al fatto che:

- i CPV spesso sono assenti, in quanto non previsti dagli schemi ANAC;
- anche quando i codici di categoria merceologica sono allegati a un contratto o bando, spesso sono errati.

I problemi legati alla qualità dei dati, in questo caso, rendono obbligatoria la validazione manuale, processo oneroso in termini temporali. La validazione è stata effettuata su un campione casuale di mille contratti estratti a partire da quelli taggati tramite le espressioni regolari.

È doveroso puntualizzare che alcuni contratti sono di difficile classificazione, non è stato possibile individuare con sicurezza la loro categoria merceologica di appartenenza.
La tassonomia CPV attualmente in uso risale al 2008, ma alcuni ambiti merceologici evolvono a un ritmo molto alto, quindi è difficile mantenere la classificazione aggiornata. In alcuni casi non è stato possibile stabilire se la classificazione assegnata ai contratti dal sistema di regole fosse corretta o errata.

Il numero di contratti taggati diminuisce all'aggiunta di un nuovo filtro. Il campione di partenza è di 431 mila contratti estratti dal DB aventi la stessa distribuzione di mercato reale.

Con la prima versione del codice, quella senza nessun filtro applicato, vengono taggati 25'313 contratti, il 5,863%; l'accuratezza di classificazione è l'89,30% con un tasso di errore del 10.40%.
Aggiungendo il filtro sulla copertura vengono taggati 13'405 contratti pari al 3,107% del campione iniziale; l'accuratezza aumenta fino al 94,30% con un tasso di errore del 4,40%.
Con l'ultima versione, contenente i filtri sulla copertura del contratto e il range di importo, l'accuratezza giunge al 98,71% e la percentuale di errore

scende all'1.29%; di contro il numero dei contratti che si è riusciti a classificare è di 8'904, il 2.062%.

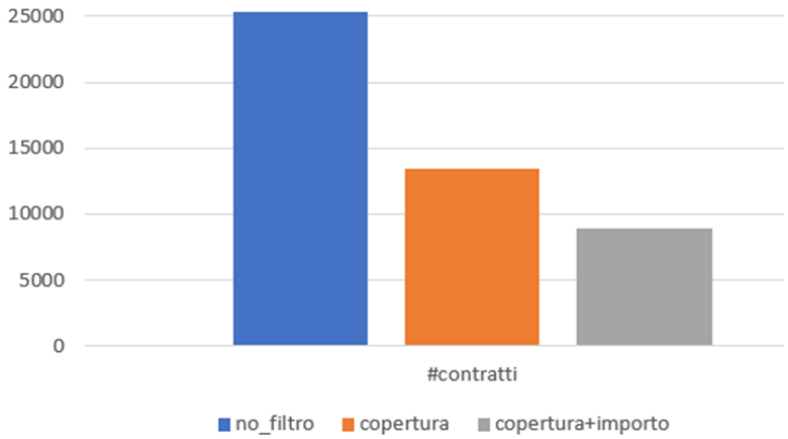


Figura 5.11: Numero di contratti taggati con i diversi filtri.

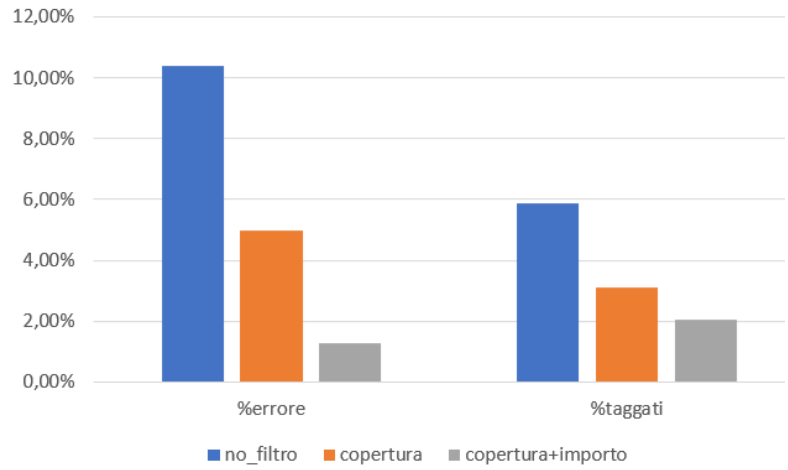


Figura 5.12: Statistiche accuratezza validazione delle regex.

Capitolo 6

Machine Learning

La dissertazione sull'algoritmo di machine learning che porta a classificare i contratti che non si è riusciti a taggare tramite le regular expression chiude lo studio.

Nel momento in cui si sceglie di processare qualcosa tramite tecniche di machine learning, preliminarmente sarebbe necessario argomentare sull'approccio scelto e dissertare dei vari approcci disponibili.

Avendo operato in un contesto aziendale, non è stato possibile mettere in discussione l'approccio, poiché un cambiamento nelle tecniche e/o nelle reti da utilizzare avrebbe vanificato anni di lavoro e di affinazione dell'algoritmo di classificazione stesso.

Tuttavia, si è pienamente consapevoli che alcune reti neurali abbastanza recenti potrebbero essere ideali per il compito di classificazione che si deve affrontare, situazione che rappresenta – ad avviso di chi scrive - la via da seguire per eventuali e possibili migliorie a venire e ulteriori sviluppi futuri.

6.1 Cos'è NLP

Con NLP, *Natural language processing*, si intende una branca dell'intelligenza

artificiale il cui obiettivo è quello di fornire l'abilità di interpretare suoni e testi prodotti dagli umani nel modo in cui vengono percepiti e successivamente interpretati dagli umani stessi.

«Il natural language processing combina la linguistica computazionale [basata su un modello a regole del linguaggio umano] [...] con statistiche, modelli di machine learning, deep learning. Insieme queste tecnologie abilitano il computer a processare il linguaggio umano tramite dati in forma di testi e suoni per comprenderne il pieno significato, con le intenzioni e i sentimenti propri del parlante o dello scrivente»[53].

La principale difficoltà che si incontra con NLP è rappresentata dal confronto con la lingua stessa; scrivere del software che comprenda appieno il linguaggio umano è un'attività difficoltosa perché bisogna interpretare molto bene *«omonimi, omofoni, idiomi, metafore, grammatiche, eccezioni e variazioni nella forma delle frasi»*[53].

La ricerca che ha come argomento il natural language processing trova applicazione in una serie di strumenti e casi d'uso, tra cui è possibile annoverare:

- *tokenizzazione*, è il processo di separare le parole presenti all'interno di un testo. Nell'italiano le parole vengono separate da spazi, quindi l'operazione risulta triviale, ma nei linguaggi che fanno uso di ideogrammi la segmentazione di testo richiede un'approfondita conoscenza sintattica e morfologica della lingua;

- *riconoscimento dello spam*, *«le migliori tecnologie di riconoscimento dello spam usano tecnologie NLP per la classificazione del testo [...] per scansionare email al fine di individuare termini che indicano spesso spam o phishing. Questi [termini] [...] includono l'uso eccessivo di termini finanziari, grammatica spesso scorretta, urgenza inappropriata, nomi di compagnia scritti male»*[53]. Quest'ambito di ricerca è considerato soddisfacente rispetto all'uso, le tecniche di riconoscimento dello spam

funzionano molto bene, sebbene la terminologia utilizzata e i modi si evolvano con il tempo;

- *stemming e lemmatizzazione.* Il primo è un processo linguistico che porta la parola alla sua forma flessa per trovare lo stemma. Il lemma, invece è la voce di citazione della parola all'interno del vocabolario, quindi con la lemmatizzazione si porta la parola al suo lemma;

- *sentiment analysis*, è il più comune strumento di classificazione del testo con lo scopo di stabilire *«se il sentimento alla base [della frase o del testo] è positivo, negativo o neutro»*[54][Gupta, 2019] . Gli usi più comuni della sentiment analysis riguardano la percezione che il pubblico ha di un brand. Tramite questo tipo di approccio è possibile anche riuscire a determinare l'opinione delle persone sugli argomenti più disparati.

Sentiment analysis e riconoscimento dello spam sono esempi di applicazioni di NLP.

6.2 I possibili approcci

La classificazione del testo diventa - quindi - un punto cruciale all'interno di NLP. La categorizzazione di flussi testuali assume rilevanza significativa per le aziende interessate a conoscere l'opinione sui propri prodotti e servizi, dal momento che fornisce la possibilità di visualizzare statistiche complesse e aggregati ad alto livello.

Se si applicasse ai contratti e ai bandi emessi dalle stazioni appaltanti, il natural language processing potrebbe aiutare a discriminare il codice di categoria merceologica da assegnare a ciascun appalto pubblico.

6.2.1 Creazione dei modelli

I *language model* costituiscono le basi su cui si poggia il natural language

processing. *Un language model nel contesto di NLP è un modello statistico che determina la probabilità che una data sequenza di parole sia presente in una frase in relazione alle N parole precedenti*[55][Chaudhuri, 2022].
Innanzitutto il testo viene *tokenizzato* per separare le parole, quindi se le parole vengono non sono nella lista di stopword, vengono stemmizzate per ridurne la cardinalità. Le macchine non riescono a comprendere i dati in formato testuale e i vettorizzatori convertono le paorle in numeri, in modo che l'informazione presente nel testo sia comprensibile.
Per la creazione dei modelli, ci si è appoggiati ai vectorizer creati da Synapta basatii su *Bag of Words*, validati da esperti nel settore.
Dopo la creazione della Bag of Words, si possono allenare i modelli.

6.2.2 Classificazione del testo

Uno studio del 2016[56][Singh et al., 2016] mette a confronto alcune tecniche di apprendimento supervisionato finalizzato alla classificazione del testo come le *reti bayesiane, logistic regression, support vector machine, reti neurali, k-nearest neighbours e random forest.* Oltre a spiegare il modo in cui funzionano questi diversi approcci, ne vengono anche evidenziati criticità e punti di forza per esplicitare la ragione per cui dovrebbero - auspicabilmente - essere utilizzati.

Le *reti bayesiane* non sono adatte a dataset con alta dimensionalità, esse tra l'altro restituiscono risultati difficilmente interpretabili. Se si pensa che le sole divisioni - le classi di primo livello - CPV sono 45, allora questo approccio non si adatta al compito di classificazione dei contratti.

La *logistic regression*, invece, è un modello che può essere facilmente aggiornato per analizzare nuovi dati: è facile da implementare e veloce, ma tende all'*overfitting*[57][Rout, 2020] in dataset ad alta dimensionalità, nonostante la possibilità di effettuare la regularization per evitarlo; tuttavia, se il numero di osservazioni è minore del numero di features, LR non dovrebbe essere usata, altrimenti potrebbe portare all'overfitting.

Le *support vector machine* (SVM) possono produrre un'accuratezza molto alta e raramente vanno in overfitting, inoltre *«con un kernel adatto, possono funzionare molto bene anche se i dati non sono linearmente separabili [...][ma] a differenza del k-Nearest-Neighbors (k-NN), l'accuratezza e le prestazioni sono indipendenti dal numero di dati, però dipendono dal numero di cicli di allenamento. [...] Hanno una buona capacità di generalizzazione e sono robusti rispetto a dati con alta dimensionalità»*[56][Singh et al., 2016]. La selezione dei parametri impatta l'accuratezza della classificazione. D'altro canto, il k-NN è computazionalmente molto costoso e la scelta del parametro k risulta di cruciale importanza per ottenere delle prestazioni accurate.

«Gli alberi di decisione sono facili da interpretare [...] gli outlier non impattano il modello. [...] Gli alberi di decisione possono gestire un'alta varietà di dati, valori mancanti e attributi ridondanti, hanno un buon grado di generalizzazione, [tuttavia] è difficile gestire i dati ad alta dimensionalità»[56][Singh et al., 2016], come nel caso di SVM e k-NN.

Il classificatore *Random Forest*, come suggerisce il nome, si compone di un gran numero di singoli alberi decisionali che agiscono come un ensemble. *«Un ensemble consiste di un insieme di classificatori allineati individualmente le cui predizioni sono combinate nella classificazione di nuove istanze»*[58][Opitz & Maclin, 1999] . La logica dietro questa scelta è rappresentata dalla cosiddetta *wisdom of the crowd*: *«un gran numero di modelli (alberi) mutuamente scorrelati si comporterà in modo migliore dei modelli costituenti. La bassa correlazione tra i modelli è la chiave [...] gli alberi si proteggono vicendevolmente dagli errori dei singoli»*[59][Yiu T., 2021].

Potrebbero essere usati anche *artificial neural networks* (ANN) o altri metodi di deep learning, ma questi soffrono nella backpropagation: hanno il problema del vanishing gradient, che però è stato superato con gli *LSTM* sebbene per quest'ultimo sarebbero necessarie elevate risorse computazionali.

Negli ultimi anni sono stati sviluppati BERT e GPT-3, che utilizzano modelli pre allenati a cui si possono aggiungere features o contesti tramite

fine-tuning.
BERT, Bidirectional Encored Representrations from Transformers, è stato sviluppato e pre-allenato da Google ed è stato messo in produzione nel 2019; GPT-3, Generative Pre-trained Transformer 3, è stato creato da OpenAI ed è stato poi acquistato da Microsoft nel 2020.
«È stato dimostrato come i modelli di linguaggio pre allenati migliorassero molti task di NLP [...] a livello frasale»[60][Devlin et al., 2018]. BERT, in particolare, supera i vincoli imposti dalla monodirezionalità del flusso di dati tramite un *masked language model* che *«maschera in modo casuale alcuni token dall'imput per predire l'id della parola nascosta basandosi sul contesto [...] permettendo di allenare un trasformatore bidirezionele [...] Bert migliora lo stato dell'arte per undici task di NLP»*[60][Devlin et al., 2018].
GPT-3, un language model autoregressivo con 175 miliardi di parametri che *ottiene buone prestazioni su molti dataset NLP, come la traduzione a la risposta alle domande.*[61][Brown et al.,2020]. Anche se sono state sviluppate architetture agnostiche rispetto al compito da eseguire, rimane comunque la necessità di avere dei dataset specifici e una fase di finetuning per avere buone prestazioni su un determinato compito e *«un potenziale approccio per risolvere questa problematica è costituito dal meta-learning, in altre parole il modello sviluppa un insieme di capacità nell'abilità di riconoscere i pattern in fase di allenamento e utilizza queste abilità »*[61][Brown et al.,2020] per la classificazione. Per GPT-3 si parla di *few shot learning* in quanto la classificazione si basa su un numero limitato di campioni di training.

GPT-3 e BERT rappresentano lo stato dell'arte nella classificazione di testo in NLP, ma come sempre in questi casi, la scelta dell'approccio adottato è un compromesso tra le risorse computazionali richieste, la velocità di esecuzione, le performance e la manutenibilità del codice.
Scegliere un approccio facendo riferimento soltanto all'accuratezza della classificazione, non consentirebbe di prende in considerazione una moltitudine di fattori che potrebbero essere impattanti a lungo termine.

6.3 Il classificatore attuale

La scelta di Synapta è ricaduta sul classificatore di tipo Random Forest. «Il classificatore CPV usato in Contrattipubblici.org consiste in una batteria di classificatori binari, uno per ogni classe di primo livello (45 in totale)» . Ogni classificatore esprime la probabilità, confidence, che un dato bando o un contratto appartengano a una classe CPV.

Il valore di soglia stabilisce la minima probabilità accettata per classificare un contratto. La sua scelta è un punto critico nel processo di classificazione: una soglia troppo bassa porterebbe a classificare più contratti, ma, di contro, farebbe aumentare il tasso di errore nella classificazione. D'altro canto, un valore di soglia alto permetterebbe di ridurre l'errore, però consentirebbe la classificazione di un numero inferiore di contratti.

Ad ogni modo, questa scelta operativa è giustificata dal fatto che questo il classificatore random forest *«gestisce facilmente un gran numero di features, fornisce una misura dell'importanza delle singole features per la classificazione [mentre le] reti neurali vanno trattate come black-box. [...] [D'altro canto] i modelli prodotti occupano molto spazio su disco [ed] è più difficile interpretare le decisioni di un classificatore RF rispetto a quelle di un singolo albero decisionale, ma le maggiori performance giustificano la scelta»*[1]. Inoltre *«per ottenere dei risultati con accuratezze significativamente superiori utilizzando delle reti neurali bisognerebbe fornire un elevato numero di esempi e maggiori risorse computazionali, specialmente in fase di training. Non si può scegliere un algoritmo da utilizzare in produzione guardando solo l'accuratezza ottenuta, bisogna anche considerare la manutenibilità del codice e delle librerie usate, le risorse computazionali richieste ed il costo dello sviluppo»*[2].

Con l'ultima versione del classificatore, l'accuratezza ottenuta è di circa

[1]Carducci, G. - Risposta a motivazioni dietro la scelta di random forest.

[2]Monti, D. - Risposta a motivazioni dietro la scelta di random forest.

l'80%. Risulta fondamentale ai fini della classificazione la lunghezza del contratto: più questo è definito e articolato, maggiore è la possibilità che venga classificato; il classificatore elimina le stopwords e le parole che hanno meno di quattro caratteri.

	Processing testo	Doc totali	Parole totali	Parole uniche	Media caratteri oggetto	Media parole oggetto
Contratti	no	100k	875k	80.1k	65.6	8.7
	si	87.6k	706k	59.4k	46.6	7.1
Bandi	no	100k	1.89M	95k	140.8	18.9
	si	98.4k	1.35M	53.2k	91.9	13.5

Figura 6.1: Appalti processati divisi tra contratti e bandi. Per gentile concessione di Giulio Carducci.

Le prestazioni del classificatore sono migliori sui bandi rispetto ai contratti. I primi sono generalmente verbosi e ben descritti, situazione che ne rende più agevole la classificazione.I contratti, invece, hanno meno parole al loro interno. Di conseguenza, non è possibile classificarne circa il 13%: non soddisfano i requisiti per poter essere classificati.

Vengono classificati i contratti al primo livello dell'albero dei CPV, per un totale di 45 classi.
Lo scopo dell'ulteriore lavoro di classificazione è quello di poter discernere in modo accurato le classi sottostanti.

Il classificatore implementato è inserito in cascata al classificatore attuale per poter scendere verticalmente, di almeno due livelli, nella gerarchia ad albero dei CPV. In breve, se il classificatore di primo livello dà come output la classe 33, il nuovo classificatore riesce a disambiguare a quali delle sottoclassi appartiene il contratto, affinandone ulteriormente la classificazione.

In questo modo ContrattiPubblici.org offre ai clienti un prodotto maggiormente accurato e, oltre a questo proposito squisitamente commerciale, il portale riesce a fornire alle stazioni appaltanti informazioni più accurate riguardanti le maggiori aree di interesse della spesa pubblica.

Sul punto, Paolo Coppola, presidente della Commissione parlamentare di inchiesta sul livello di digitalizzazione nella PA tra il 2013 e 2018, così si esprime: *«l'utilizzo di strumenti come questo potrebbe stimolare i decisori politici a cambiare il proprio approccio all'analisi dei dati, basando le proprie decisioni sui risultati di queste ultime anziché cercando [...] di trovare conferma della bontà delle proprie scelte selezionando solo i dati più utili allo scopo. [...] [Inoltre] Gli strumenti di analisi, se utilizzati correttamente, possono servire ad aumentare la trasparenza negli acquisti, ridurre il margine di errore nelle scelte da intraprendere, portare alla luce potenziali conflitti di interesse»*[62][Synapta, 2021].

6.4 Le classi di secondo e terzo livello

Prima della classificazione vera e propria è stato necessario creare le classi per cui è possibile la classificazione.

È stato supposto che le classi maggiormente rappresentate venissero classificate in modo più accurato e si sono verificate le accuratezze per ciascuna classe di primo livello.
L'obiettivo della classificazione è quello di assegnare con elevata confidenza una classe di appartenenza ai contratti, scegliendo di privilegiare l'accuracy rispetto alla recall e minimizzando i falsi positivi.

Quindi sono state analizzate le prestazioni del classificatore di primo livello e si è notato come le classi poco rappresentate avessero una recall elevata. Quindi si è dovuta inserire una soglia stringente per scegliere quali classi è possibile classificare mantenendo accuracy e precision alte.

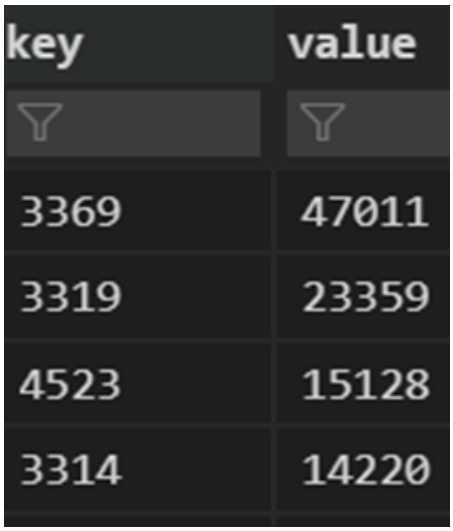
Per la creazione delle classi di livelli sottostanti sono stati esaminati tutti i contratti con CPV, ovvero 7.2 M contratti sui circa 60 M presenti sul sito, ovvero il 12% del totale.

Analizzando il CPV sunburst, grafico che mostra la distribuzione dei contratti, si rileva che 1,25 milioni di contratti appartengono alla divisione 33 e 430 mila alla divisione 45. Molte classi sono largamente rappresentate, altre lo sono poco, se non per nulla.
Il numero minimo di contratti ritenuto rappresentativo della classe per cui si vuole creare il modello è pari a 16k contratti: questo valore è la soglia minima di cardinalità.

L'export che è stato creato comprende 497 mila tra contratti e bandi della PA aventi la stessa distribuzione del mercato reale e un codice CPV associato. Di conseguenza la soglia è stata proporzionalmente modificata, assegnandole un valore di 1060. La soglia scelta viene rilassata nel momento in cui si scende verticalmente nella gerarchia dei CPV di un coefficiente di 0,75 per il secondo livello e di 0,60 per il terzo livello.

Dall'export creato vengono troncati i CPV alla quarta cifra, quindi vengono contate le occorrenze per ogni classe. Già dalla struttura che si forma si evince come alcune classi siano largamente rappresentate, mentre altre per nulla.

Si è dovuto pertanto pensare a una struttura facile da scorrere allo scopo di verificare per quali classi fosse possibile la classificazione (Figura 6.3).
Questa struttura è ricorsiva: ha come chiave la divisione, ovvero la classe di primo livello e al suo interno contiene la classe di secondo livello e il numero di contratti appartenenti alla divisione. Scendendo ancora, si trovano le



key	value
3369	47011
3319	23359
4523	15128
3314	14220

Figura 6.2: Numero di contratti per classe di terzo livello.

classi di terzo livello con il numero di contratti della classe di secondo livello. Infine, il numero di contratti della classe di terzo livello.

Se le classi di secondo livello non superano la soglia richiesta, allora queste non vengono ulteriormente classificate, perché si mantiene la classificazione di primo livello. Se, invece, la classe di secondo livello supera la soglia ed esiste almeno una classe di terzo livello sopra soglia, allora gli altri fratelli con un numero di contratti inferiore alla soglia vengono aggregati nella classe <cpv classe di secondo livello>_ altro.

6.5 Preprocessing

Quindi, le classi vengono appiattite, in modo da poter estrarre i dati positivi e negativi per effettuare la fase di training (Figura 6.4). I dati positivi sono

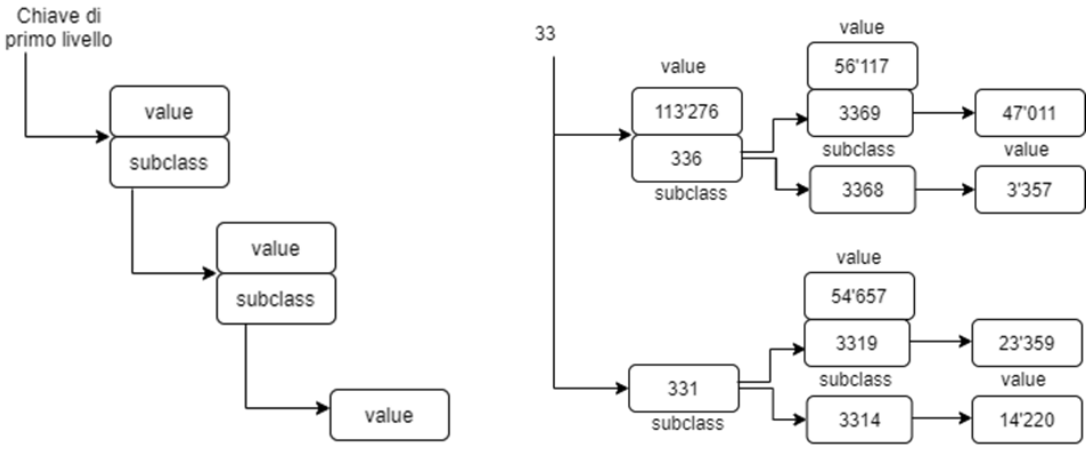


Figura 6.3: Struttura dati usata per livelli inferiori.

quelli appartenenti alla classe che si vuole analizzare, quelli negativi sono tutti gli altri contratti o bandi appartenenti alla stessa divisione.

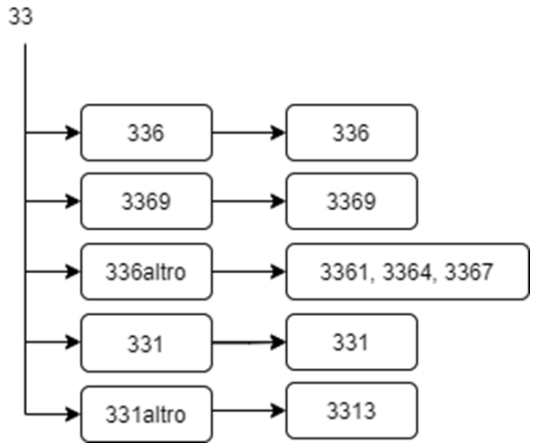


Figura 6.4: Classi di livelli inferiori.

Dopo aver creato gli esempi positivi e negativi, è possibile allenare i modelli.

Dato che non tutte le parole aggiungono al contratto semantica e significato, è necessario rimuoverle, pertanto vengono eliminate le stopwords.

Occorre tenere presente che nella lingua italiana sono presenti variazioni morfologiche: è necessario pertanto riportare la parola esaminata ad una forma nota. Le possibili soluzioni in tal senso sono lo stemming e la lemmatizzazione.

Il primo è un processo di riduzione della parola al suo stemma, ovvero la sua radice, ma è possibile incorrere in un errore comune: diverse parole con significati diversi possono avere lo stesso stemma, si pensi agli omofoni. La lemmatizzazione, computazionalmente più costosa, riconduce la parola al suo lemma, *«la forma di citazione di una parola in un dizionario»*[63][Beccaria, 2004].

Per il training sono stati creati alcuni vocabolari accessori basati su Bag of Words (BoW), uno dei metodi più diffusi per la categorizzazione di immagini o di testo, e su N-grammi, ovvero i gruppi ricorrenti di n parole.

«Il modello BoW è usato come strumento per generare e classificare le features»[64][Qader et al., 2019] del testo e in prima approssimazione, è basato sulla frequenza delle parole all'interno di un sottoinsieme. Questo modello non tiene traccia dell'ordine con cui le parole sono concatenate, si limita a tracciare la frequenza delle singole parole.

L'approccio basato su Bag of Words *«presenta un grande inconveniente. Omette le informazioni sulla semantica e sull'ordine delle parole. Quindi disperde il significato cruciale delle frasi. [...] N-gram soffre la sparsità dei dati anche se supera i problemi tipici delle BoW»*[65][Indhraom & Umarani, 2019].

6.6 Risultati sintetici

È stato effettuato uno split tra i dati assegnando il 70% dei dati al dataset di training, 20% al test set e 10% al validation set. Dopo aver processato il testo, vengono addestrati i modelli. Quindi vengono caricati i contratti con il codice CPV alla seconda cifra, che indica la divisione e vengono ordinati per classe di primo livello. Per la classificazione vengono caricati tutti i modelli che fanno riferimento alla divisione. Viene preso per buono l'output dal classificatore di primo livello: in questa ulteriore fase di classificazione avviene la disambiguazione tra le classi di livello inferiore. Sono stati creati 235 modelli di livelli inferiori su un totale di 1231 tra classi di secondo e terzo livello, si tratta del 19% circa dell'ammontare di classi possibili.

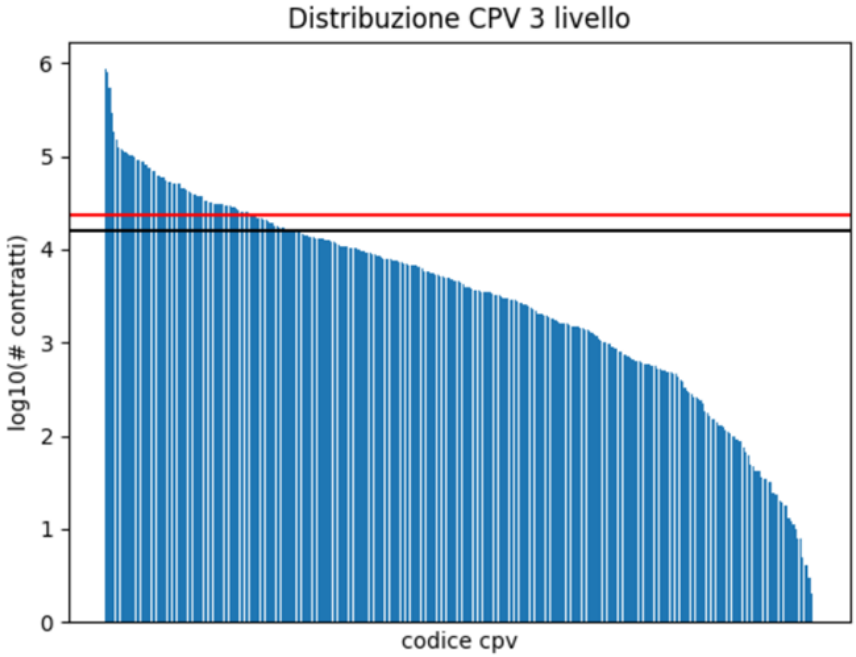


Figura 6.5: Distribuzione del mercato per codice CPV di terzo livello.

In Figura 6.5 stato plottato il numero di contratti collegati ai CPV di terzo livello. In rosso è presente la media dei contratti; la linea nera rappresenta la soglia. Le classi considerate rappresentano più il 93% del contratti totali con codice CPV.

Sono stati classificati i contratti con diverse soglie minime. Dato che gli alberi che compongono RF esprimono in uscita la probabilità che un dato elemento appartenga o meno a una classe, bisogna accertarsi che questa probabilità sia abbastanza elevata. In altre parole, occorre che ci sia una confidenza minima di classificazione, altrimenti si potrebbe incorrere in errori di classificazione ricorrenti.

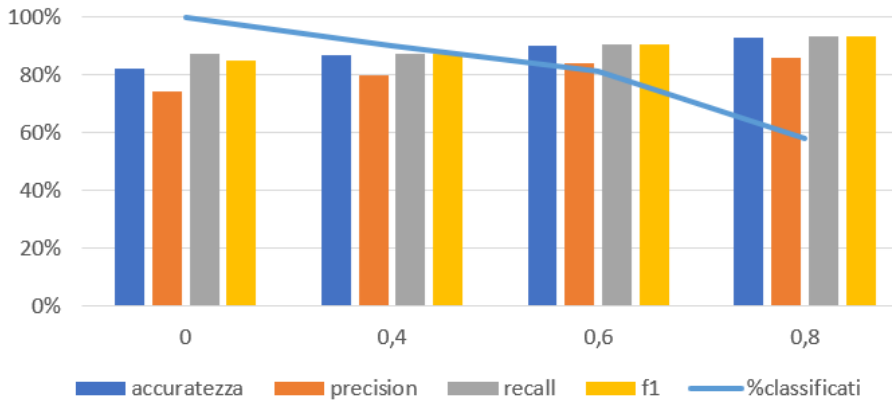


Figura 6.6: Statistiche sul test set al variare del valore di soglia.

Come è possibile notare dalla figura 6.6, senza alcuna soglia vengono classificati tutti i contratti aventi almeno quattro parole a seguito della rimozione delle stopword, seppure con accuratezza minore rispetto alle classificazioni con una soglia alta.
La precision massima, l'86%, viene raggiunta con il valore di soglia di 0,8. Questo risultato è in linea con quanto riportato dalla letteratura: avere un

valore di soglia alto implica che i contratti previsti come appartenenti a una classe siano effettivamente tali. In generale, quindi, con il valore di soglia crescente, anche la precision dovrebbe incrementarsi.

L'aumento della precision implica un minor numero di falsi positivi. Impostando la soglia a 0,8 vengono classificati il 58% dei contratti con un'accuratezza del 92%. L'accuratezza scende al 90% con la soglia a 0,6 ma vengono classificati l'82% dei contratti.

Capitolo 7

Conclusioni

L'evoluzione tecnologica coinvolge in maniera sempre più crescente la pubblica amministrazione, che, a sua volta, genera quotidianamente una ingente mole di dati in riferimento a contratti e appalti.

La gestione di questi enormi volumi di dati - se effettuata in maniera puntuale e adeguata - concretamente dovrebbe tradursi, *ipso facto*, in strumento operativo per la *business intelligence*, ma soprattutto, in garanzia di una effettiva trasparenza dell'azione amministrativa.

L'esperienza svolta in Synapta, che gestisce ContrattiPubblici.org, motore di ricerca dei contratti pubblici italiani, ha sollecitato le mie riflessioni tanto sui punti di forza – di cui si è già detto avanti - del sistema di classificazione dei contratti, quanto sui limiti e sulle criticità dell'attuale sistema di classificazione, al fine di prospettarne opportunità di possibili implementazioni e ipotesi di alternative future

Seppure le attuali tecniche di machine learning riescano a semplificare la complessità della gestione di grandi volumi di dati, non sempre essi hanno gli attributi per poter essere gestiti dalla macchina: l'alberatura di categorie merceologiche cui oggi si fa riferimento è il CPV, *Common Procurement* Vocabulary, aggiornato per l'ultima volta nel 2008, nonostante il proliferare

di forniture e servizi richiesti.

È di tutta evidenza che una categoria merceologica molto ampia - perché non dettagliatamente definita e circoscritta - crea un setaccio a maglie abbastanza larghe da provocare la dispersione di tanti dati, appunto quelli riferiti ai contratti non classificabili.
In quest'ottica il dizionario di regole riesce a superare i limiti del classificatore, creando delle classificazioni verticali e rendendo possibili progetti di annotazione ad-hoc sotto richiesta dei clienti che hanno l'interesse ad approfondire la conoscenza di un'area di mercato. L'approccio secondo regole costituisce una soluzione efficace con risultati estraibili in tempi ragionevoli.

I modelli di machine learning migliorano le performance in base ai dati che utilizzano ma – e non sembra superfluo ribadirlo – la classificazione non può funzionare, o non funzionerà come si vorrebbe, nel caso in cui siano stati omessi dati in fase di compilazione del contratto. Tuttavia BERT e GPT-3 possono attutire il rumore, dato che sono pre-allenati, e si potrebbero prestare molto bene al compito di classificazione, giacché ne costituiscono lo stato dell'arte.

Aldilà di ogni valida considerazione su avveniristici metodi di classificazione dei contratti, si può affermare - senza tema di smentita - che il sistema ibrido di classificazione, che coinvolge intervento umano e machine learning, attualmente adottato da Synapta, sia il più efficace.
In conclusione lo stato dell'arte leggittima dunque a pensare che sarebbe auspicabile migliorare il sistema delle regex in riferimento sia alla loro creazione che alla loro manutenzione.

Bibliografia

[1] *Regio Decreto 18 Novembre 1923, n. 2440*. URL: https://www.bosettiegatti.eu/info/norme/statali/1923_2440.htm (cit. a p. 3).

[2] *Regio Decreto 18 novembre 1923, n. 2440*. 1923. URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:regio.decreto:1923-11-18;2440!vig= (cit. a p. 4).

[3] Tulino Domenico. *L'evoluzione normativa dei contratti Pubblici in Italia*. Apr. 2013. URL: http://www.salvisjuribus.it/levoluzione-normativa-dei-contratti-pubblici-in-italia/ (cit. a p. 4).

[4] *Legge 11 Dicembre 1984, n. 839*. 1984. URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1984-12-11;839!vig= (cit. a p. 4).

[5] *Legge 7 Agosto 1990, n. 241*. 1990. URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1990-08-07;241 (cit. alle pp. 4, 5).

[6] *Legge 6 Novembre 2012, n. 190; Legge Anticorruzione*. 2012. URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2012-11-06;190 (cit. a p. 5).

[7] *Decreto Legislativo 14 Marzo 2013, n. 33; Decreto Trasparenza*. 2013. URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2013-03-14;33!vig (cit. a p. 5).

[8] Dipartimento della Funzione Pubblica. *FOIA - Ministro per la Pubblica Amministrazione*. 2019. URL: https://www.funzionepubblica.gov.it/foia-7 (cit. a p. 6).

[9] *Decreto Legislativo 18 Aprile 2016, n. 50*. Apr. 2016. URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2016-04-18;50 (cit. a p. 6).

[10] Francesco Caringella. *Manuale di diritto amministrativo*. 12ª ed. Roma, RO: Dike giuridica, 2018 (cit. alle pp. 6, 25).

[11] Pisanu. *Procurement, Le proposte dell'antitrust per le riforma del settore*. Mag. 2021. URL: https://www.agendadigitale.eu/procurement/procurement-le-proposte-dellantitrust-per-le-riforma-del-settore/ (cit. alle pp. 7, 8).

[12] Sanna Silvia | ContrattiPubblici.org. *Specifiche Tecniche Anac: L'interpretazione delle Pa. Specifiche tecniche ANAC: l'interpretazione delle PA*. Mar. 2019. URL: https://contrattipubblici.org/blog/2019/03/25/specifiche-tecniche-anac-interpretazione-delle-pubbliche-amministrazioni/ (cit. alle pp. 7, 20, 21).

[13] *Decreto Legislativo 2005, n. 82; Codice dell'Amministrazione Digitale*. 2005. URL: https://www.agid.gov.it/sites/default/files/repository_files/leggi_decreti_direttive/dl-7-marzo-2005-82_0.pdf (cit. a p. 8).

[14] AGID - Agenzia per l'Italia Digitale. *Formati Aperti*. URL: https://www.agid.gov.it/it/dati/formati-aperti (cit. a p. 8).

[15] Morando Federico | ContrattiPubblici.org. *L'importanza degli Open Data nel mondo della PA*. Ott. 2020. URL: https://contrattipubblici.org/blog/2020/10/06/importanza-open-data-nel-mondo-della-pa/ (cit. alle pp. 8, 9, 22, 36, 37).

[16] Open Data Handbook. *Il manuale degli Open Data*. URL: http://opendatahandbook.org/guide/it/ (cit. a p. 9).

[17] Open Data Institute - ODI. *Knowledge & opinion.* URL: https://theodi.org/article/what-is-open-data-and-why-should-we-care/ (cit. a p. 9).

[18] C. Bizer, T. Heath e T. Berners-Lee. «Linked data - the story so far». In: *International Journal on Semantic Web and Information Systems* 5.3 (2009), pp. 1–22 (cit. a p. 10).

[19] Antonio Vetrò, Lorenzo Canova, Marco Torchiano, Camilo Orozco Minotas, Raimondo Iemma e Federico Morando. «Open data quality measurement framework: Definition and application to Open Government Data». In: *Government Information Quarterly* 33.2 (2016), pp. 325–337. ISSN: 0740-624X. DOI: https://doi.org/10.1016/j.giq.2016.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0740624X16300132 (cit. alle pp. 11, 14, 16).

[20] Hugo Moreno | Forbes. *The importance of data quality - good, bad or ugly.* Giu. 2017. URL: https://www.forbes.com/sites/forbesinsights/2017/06/05/the-importance-of-data-quality-good-bad-or-ugly/ (cit. alle pp. 11, 12).

[21] ISO/IEC. *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Data quality model.* ISO/IEC 25012. Geneva, Switzerland: International Organization for Standardization, 2008 (cit. alle pp. 12, 14).

[22] Sanna Silvia | ContrattiPubblici.org. *Big data e Pubblica Amministrazione: Perché sono importanti?* Ago. 2019. URL: https://contrattipubblici.org/blog/2019/08/22/big-data-pubblica-amministrazione-perche-importanti/ (cit. a p. 14).

[23] Agenzia per l'Italia Digitale DOCS Italia. *Piano Triennale per l'informatica nella Pubblica Amministrazione 2020 - 2022, Executive Summary.* 2020. URL: https://docs.italia.it/italia/piano-triennale-ict/pianotriennale-ict-doc/it/2020-2022/executive_summary.html (cit. a p. 15).

[24] Agenzia per l'Italia Digitale DOCS Italia. *Piano Triennale per l'informatica nella Pubblica Amministrazione 2020 - 2022, Interoperabilità: Obiettivi e risultati attesi*. 2020. URL: https://docs.italia.it/italia/piano-triennale-ict/pianotriennale-ict-doc/it/2020-2022/capitolo-5-interoperabilit%C3%A0/obiettivi-e-risultati-attesi-4.html (cit. a p. 15).

[25] Agenzia per l'Italia Digitale DOCS Italia. *Piano Triennale per l'informatica nella Pubblica Amministrazione 2020 - 2022, Obiettivi e risultati attesi*. 2020. URL: https://docs.italia.it/italia/piano-triennale-ict/pianotriennale-ict-doc/it/2020-2022/capitolo-2-dati/obiettivi-e-risultati-attesi-1.html (cit. a p. 15).

[26] Marco Torchiano, Antonio Vetrò e Francesca Iuliano. «Preserving the Benefits of Open Government Data by Measuring and Improving Their Quality: An Empirical Study». In: *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 1. 2017, pp. 144–153. DOI: 10.1109/COMPSAC.2017.192 (cit. alle pp. 17, 18).

[27] Autorità Nazionale AntiCorruzione ANAC. *Specifiche tecniche per la pubblicazione dei dati ai sensi dell'art. 1 comma 32 Legge n. 190/2012*. Giu. 2016. URL: https://www.anticorruzione.it/portal/rest/jcr/repository/collaboration/Digital%5C%20Assets/anacdocs/Servizi/ServiziOnline/AdempimentoLegge190/Specifiche%5C%20Tecniche%5C%20Legge%5C%20190%5C%20v1.2_finale.pdf (cit. alle pp. 19–21).

[28] Artusio Claudio | ContrattiPubblici.org. *Dati dei Contratti pubblici: è obbligatorio aggiornarli?* Dic. 2019. URL: https://contrattipubblici.org/blog/2019/12/10/dati-dei-contratti-publici-bisogna-aggiornarli/ (cit. a p. 21).

[29] Open Knowledge Group. *Open Definition*. URL: https://opendefinition.org/ (cit. a p. 22).

[30] Morando Federico | ContrattiPubblici.org. *Open data: Tra La fossa della Disillusione e le Fondamenta della Piramide*. Nov. 2019. URL: https://contrattipubblici.org/blog/2019/11/15/open-data-tra-la-fossa-della-disillusione-e-le-fondamenta-della-piramide/ (cit. a p. 23).

[31] Zangarini Emanuele | ContrattiPubblici.org. *Fact checking con i dati di ContrattiPubblici.org*. Nov. 2020. URL: https://contrattipubblici.org/blog/2020/11/05/fact-checking-con-i-dati-di-contrattipubblici-org/ (cit. a p. 23).

[32] Synapta | ContrattiPubblici.org. *Anna Cavallo, CSI Piemonte: "con ContrattiPubblici.org possiamo avere una visione d'insieme sul mercato della PA"*. Ott. 2021. URL: https://contrattipubblici.org/blog/2021/10/21/anna-cavallo-csi-piemonte-contrattipubblici-org/ (cit. a p. 24).

[33] Synapta | ContrattiPubblici.org. *Daniela Galletti, Rekeep: "ContrattiPubblici.org è uno strumento essenziale per la pianificazione commerciale"*. Nov. 2021. URL: https://contrattipubblici.org/blog/2021/11/02/daniela-galletti-rekeep/ (cit. a p. 24).

[34] Synapta | ContrattiPubblici.org. *Stefano Arnoldi, Arkytec | 360° – SPM: "da ContrattiPubblici.org i dati che servono per progettare e realizzare la smart city"*. Feb. 2022. URL: https://contrattipubblici.org/blog/2022/02/28/stefano-arnoldi-arkytec-360-spm-smart-city/ (cit. a p. 24).

[35] Brocardi.it. *Oggetto del contratto, Dizionario Giuridico*. URL: https://www.brocardi.it/dizionario/1888.html (cit. a p. 25).

[36] Synapta | ContrattiPubblici.org. *Glossario, Stazione Appaltante*. URL: https://contrattipubblici.org/glossario/stazione-appaltante (cit. a p. 25).

[37] Synapta | ContrattiPubblici.org. *Glossario, Procedura Scelta Contraente.* URL: https://contrattipubblici.org/glossario/scelta-contraente (cit. a p. 26).

[38] Sanna Silvia | ContrattiPubblici.org. *Gare d'appalto: Quante Tipologie Esistono? Gare d'appalto: quante tipologie esistono?* Giu. 2020. URL: https://contrattipubblici.org/blog/2020/06/22/gare-dappalto-quante-tipologie-esistono/ (cit. a p. 26).

[39] Synapta | ContrattiPubblici.org. *Glossario, Procedura Scelta Contraente.* URL: https://contrattipubblici.org/glossario/procedura-aperta (cit. a p. 26).

[40] Synapta | ContrattiPubblici.org. *Glossario, Procedura Scelta Contraente.* URL: https://contrattipubblici.org/glossario/procedura-negoziata (cit. a p. 26).

[41] Synapta | ContrattiPubblici.org. *Glossario, Procedura Scelta Contraente.* URL: https://contrattipubblici.org/glossario/procedura-ristretta (cit. a p. 26).

[42] Synapta | ContrattiPubblici.org. *Glossario, Procedura Scelta Contraente.* URL: https://contrattipubblici.org/glossario/sistema-dinamico-di-acquisizione (cit. a p. 26).

[43] Synapta | ContrattiPubblici.org. *Glossario, Procedura Scelta Contraente.* URL: https://contrattipubblici.org/glossario/codice-identificativo-gara (cit. a p. 27).

[44] Synapta | ContrattiPubblici.org. *Glossario, Procedura Scelta Contraente.* URL: https://contrattipubblici.org/glossario/common-procurement-vocabulary (cit. a p. 28).

[45] European Commition. *Public procurement in the European Union Guide to the Common Procurement Vocabulary.* Giu. 2008. URL: https://simap.ted.europa.eu/documents/10184/36234/cpv_2008_guide_en.pdf (cit. alle pp. 28–30).

[46] Martiello et al. *In quali Casi non serve IL CIG*. Mag. 2019. URL: https://www.funzionarioamministrativo.it/2019/05/04/no-codice-cig/ (cit. a p. 30).

[47] Autorità Nazionale AntiCorruzione ANAC. *GUIDA PRATICA ALL'USO DI SIMOG*. Nov. 2020. URL: https://www.anticorruzione.it/portal/rest/jcr/repository/collaboration/Digital%5C%20Assets/anacdocs/Attivita/Atti/Delibere/2020/del.1120.2020.pdf (cit. a p. 30).

[48] Ruggiero G. *Come ottenere Lo Smartcig*. Lug. 2021. URL: https://www.funzionarioamministrativo.it/2019/05/04/come-chiedere-smartcig/ (cit. a p. 31).

[49] Oracle. *Che cosa significa business intelligence? Che cos'è la business intelligence?* URL: https://www.oracle.com/it/what-is-business-intelligence/ (cit. alle pp. 36, 38).

[50] SeedScientific. *How much data is created every day? [27 powerful stats]*. Ott. 2021. URL: https://seedscientific.com/how-much-data-is-created-every-day/ (cit. a p. 38).

[51] AGID. *La spesa ICT nella PA italiana 2020. Principali trend e percorsi in atto*. Gen. 2021. URL: https://www.agid.gov.it/sites/default/files/repository_files/report_sulla_spesa_ict_nelle_pa_2020_0.pdf (cit. a p. 45).

[52] Confindustria. *IL DIGITALE IN ITALIA 2021. Previsioni 2021-2024 e Policy*. Nov. 2021. URL: https://preparatialfuturo.confindustria.it/wp-content/uploads/2021/11/il_digitale_in_italia_-_previsioni_2021_24_e_policy_Anitec-Assinform.pdf (cit. a p. 46).

[53] IBM Cloud Education. *What is natural language processing?* URL: https://www.ibm.com/cloud/learn/natural-language-processing (cit. a p. 60).

[54] S. Gupta. *Sentiment analysis: Concept, analysis and applications*. Gen. 2019. URL: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17 (cit. a p. 61).

[55] K. D. Chaudhuri. *Building language models in NLP. Analytics Vidhya*. Gen. 2022. URL: https://www.analyticsvidhya.com/blog/2022/01/building-language-models-in-nlp/ (cit. a p. 62).

[56] Amanpreet Singh, Narina Thakur e Aakanksha Sharma. «A review of supervised machine learning algorithms». In: *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. 2016, pp. 1310–1315 (cit. alle pp. 62, 63).

[57] A. R. Rout. *Advantages and disadvantages of logistic regression*. Set. 2020. URL: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/ (cit. a p. 62).

[58] David Opitz e Richard Maclin. «Popular Ensemble Methods: An Empirical Study». In: *J. Artif. Int. Res.* 11.1 (lug. 1999), pp. 169–198. ISSN: 1076-9757 (cit. a p. 63).

[59] T. Yiu. *Understanding random forest*. Set. 2021. URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (cit. a p. 63).

[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee e Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805 (cit. a p. 64).

[61] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. DOI: 10.48550/ARXIV.2005.14165. URL: https://arxiv.org/abs/2005.14165 (cit. a p. 64).

[62] Synapta. *Paolo Coppola e il Ruolo di contrattipubblici.org nella commissione sulla digitalizzazione della pa.* Ott. 2021. URL: https://contrattipubblici.org/blog/2021/10/27/paolo-coppola-commissione-pa/ (cit. a p. 67).

[63] G.L. Beccaria. *Dizionario di linguistica e di filologia, metrica, retorica.* Nuova serie. Einaudi, 2004. ISBN: 9788806169428. URL: https://books.google.it/books?id=GJPlAAAAMAAJ (cit. a p. 71).

[64] Wisam Abdulaziz, Musa M. Ameen e Bilal Ahmed. «An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges». In: giu. 2019, pp. 200–204. DOI: 10.1109/IEC47844.2019.8950616 (cit. a p. 71).

[65] Indhraom M e Umarani Srikanth. «Survey of Sentiment Analysis Using Deep Learning Techniques». In: apr. 2019, pp. 1–9. DOI: 10.1109/ICIICT1.2019.8741438 (cit. a p. 71).