

Il Machine Learning al lavoro sui bandi di gara della Pubblica Amministrazione

Rosa Meo con Mirko Lai, Paolo Pasteris, Reni Hoxhaj, Joana Cfarku, Alessandro Marrazzo, Alessia Ambu





Di cosa parleremo...



- La collaborazione tra UNITO e ANAC: obiettivi
- Il percorso di analisi dati
- Risultati
- Spiegazione
- Lavori in corso e futuri



La collaborazione tra UNITO e ANAC



La collaborazione con ANAC

ANAC e l'Università di Torino hanno stipulato nel 2019 una convenzione per la ricerca e analisi dati sui contratti pubblici (referente per Unito la Prof. Gabriella Racca)

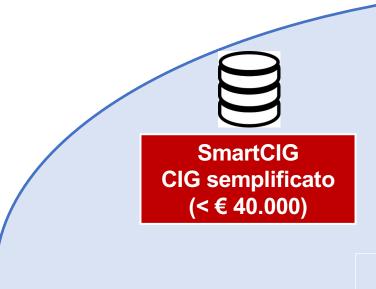
La *BDNCP* è un *data lake* che comprende:

- Sistema Informativo Monitoraggio Gare (SIMOG) con dati sulle procedure di gara
- SMART CIG con dati sugli appalti di importo sotto soglia (<40 mila euro)
- Anagrafe Unica delle Stazioni Appaltanti (AUSA)
- CEL, banca dati dei Certificati di Esecuzione Lavori



Banca Dati Nazionale dei Contratti Pubblici (BDNCP)







S.I.MO.G.
Sistema informativo monitoraggio gare (CIG)



C.E.L.
Certificati Esecuzione
Lavori

DATALAKE della B.D.N.C.P.



Altre banche

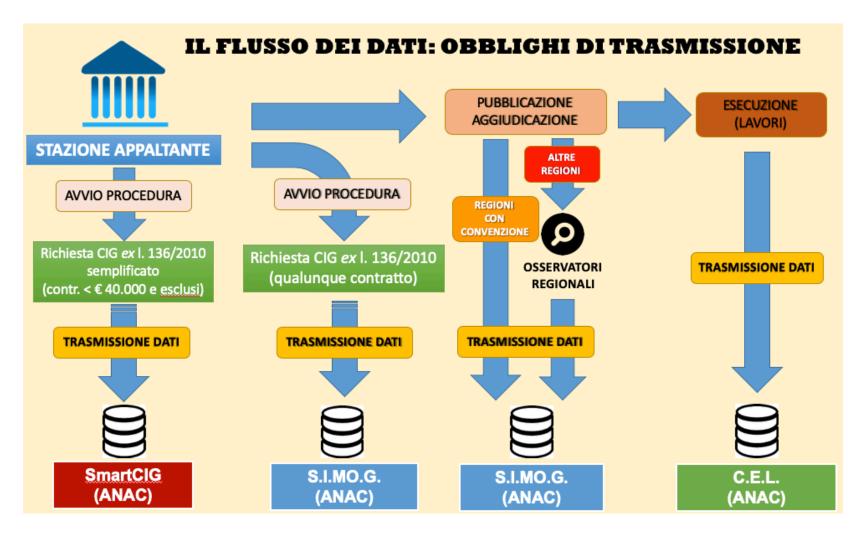
dati



A.U.S.A.
Anagrafe Unica
Stazioni Appaltanti



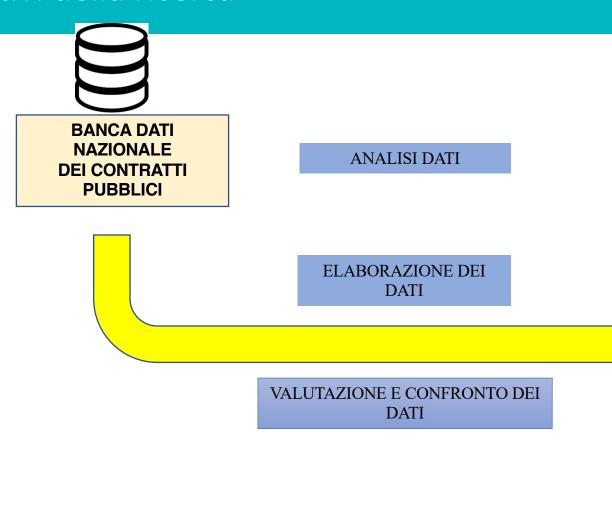
Il flusso dei dati





Obiettivi della ricerca





DEFINIZIONE E UTILIZZO

DI INDICI

POTENZIAMENTO DEI SUPPORTI CONOSCITIVI PER LE **DECISIONI PUBBLICHE**

PROMOZIONE E DIFFUSIONE DELLE BEST PRACTICES

EFFICIENZA E INTEGRITA' **DELLA PUBBLICA AMMINISTRAZIONE**

MIGLIORAMENTO DELLA QUALITA' DEI SERVIZI E DELLA TRASPARENZA **AMMINISTRATIVA**

PREVENZIONE DEI FENOMENI DISTORSIVI



Il percorso di analisi dati

Analisi Prescrittiva

Analisi Predittiva

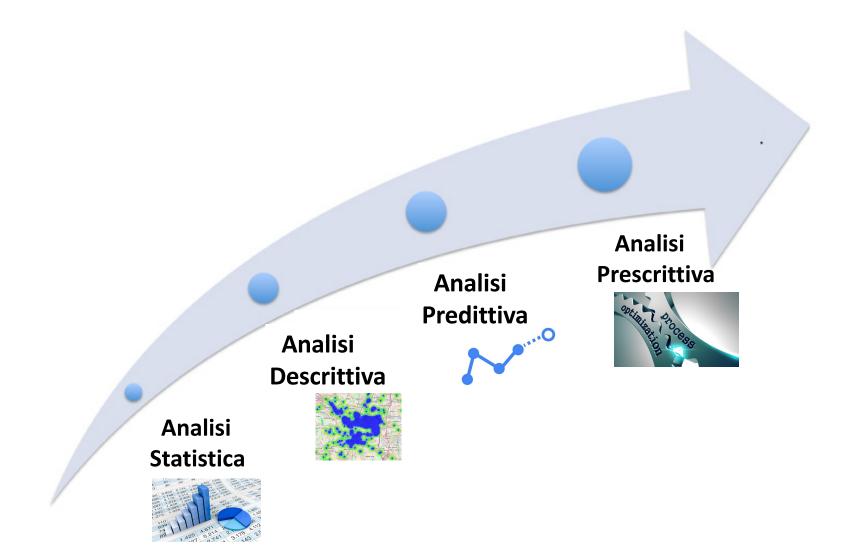
Analisi Descrittiva

Analisi Statistica



Il percorso di analisi dati







Librerie Software usate nel progetto



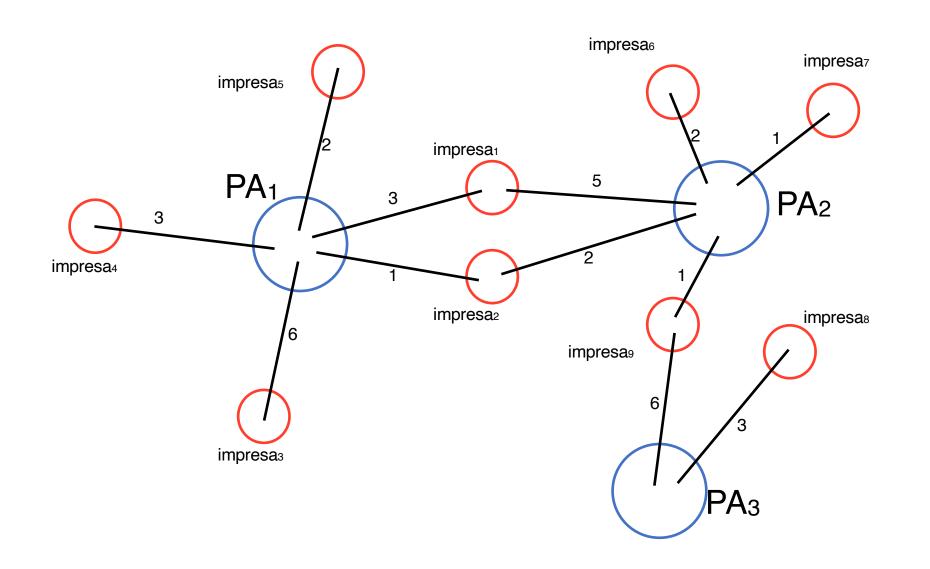


Il grafo delle collaborazioni tra PA e imprese



Analisi descrittiva: il grafo degli appalti pubblici







Analisi grafo per macro-area



In questa nuvola che rappresenta il grafo degli appalti pubblici, rappresentiamo per leggibilità solo le stazioni appaltanti



Stazioni appaltanti di grosse dimensioni bandiscono contratti vinti da tutta Italia (e sono centrali nel grafo)

Esempi:

CNR, segnaletica stradale, Tecnositaf (sicurezza stradale)



Altri risultati dell'analisi sui grafi



Abbiamo analizzato i dati dei bandi di gara pubblicati dalle pubbliche amministrazioni per formulare indicatori di anomalia

Esempio:

numero di bandi ripetutamente aggiudicati alla stessa impresa dalla stazione appaltante

Risultati:

Esiste una forte disomogeneità nelle consuetudini di stipula di appalti nei vari enti (comuni, provincie, enti pubblici a partecipazione privata, dipartimenti dello stato)



Il percorso di analisi dati

Analisi Prescrittiva

Analisi Predittiva

Analisi Descrittiva

Analisi Statistica





Raccomandazione degli importi alle stazioni appaltanti



Abbiamo analizzato la distribuzione degli importi nei lotti dei bandi di gara sugli ordinativi di un determinato bene (risme carta) in una stessa tipologia di stazione appaltante (le Università)

Il diagramma mostra sull'asse delle ordinate (y) la frazione dei bandi di gara che hanno un importo fino a un valore dell'ascissa (x) considerato come valore estremo:



Escludiamo gli importi troppo piccoli o troppo grossi (per cui vi è solo il 10% delle gare che richiedono il bene per un importo oltre quel valore) \Rightarrow troviamo la fascia di importi usati nel 80% dei bandi

Suggerimento nelle linea guida



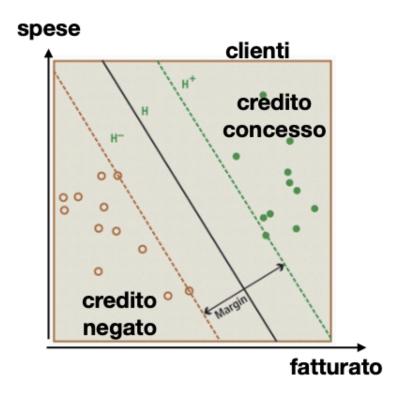
Predizione dei bandi di gara con varianti in corso d'opera

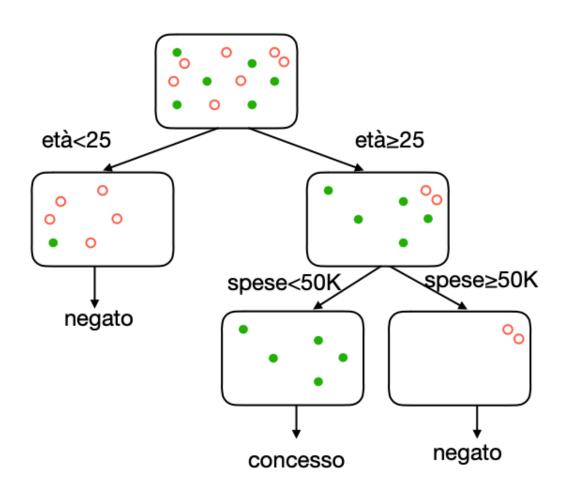


Modelli di apprendimento automatico



Concessione del credito





Support Vector Machines

Alberi di decisione



Analisi predittiva per riconoscere se un appalto di un lavoro pubblico darà luogo ad una variante in corso d'opera



Analisi di tipo **supervisionato**:

un esperto associa un *quid di informazione* (etichetta: normale o con variante) ad un insieme di casi di esempio (gare)

ANAC ci ha fornito l'elenco dei contratti che hanno dato luogo a varianti, con:

- etichettatura delle gare passate

Noi abbiamo:

- costruito un modello di riconoscimento delle gare con varianti
- applicazione del modello alle gare in corso

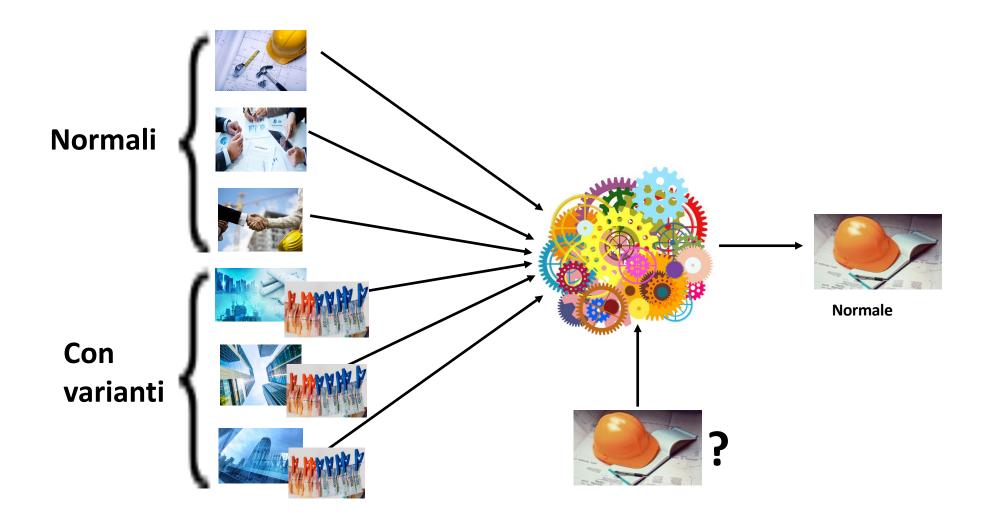
Riconoscimento di pattern comuni tra le gare con varianti

Predizione nel futuro



Predizione gare con varianti







Descrizione della Raccolta Dati



- Abbiamo raccolto i dati per costruire un modello di classificazione che sia capace di predire se i bandi pubblicati dalla pubblica amministrazione daranno luogo a varianti in corso d'opera
- Per ottenere il modello, abbiamo preso il database di ANAC (dal 2010 ad oggi) delle gare (oltre 5 milioni di casi)
- Abbiamo selezionato solo i bandi relativi a lavori, che sono stati aggiudicati (escludendo i bandi andati deserti, o i contratti annullati)



Costruzione del data set



- Abbiamo trovato la presenza di varianti in: 51.569 appalti
- Li abbiamo etichettati come esempi positivi
- Abbiamo estratto casualmente altrettanti esempi di lavori:
 - aggiudicati
 - senza varianti in corso d'opera,
- Etichettiamo questi appalti come negativi

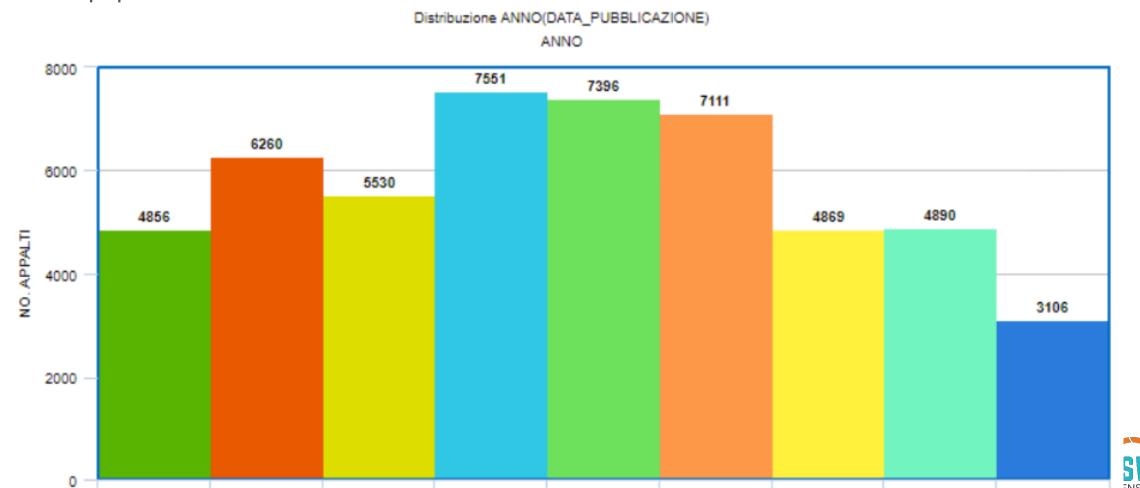
In modo che il classificatore impari anche dagli esempi negativi. Il dataset è **bilanciato** (stesso numero di positivi e negativi)



Distribuzione degli appalti nel tempo



Abbiamo usato l'accortezza di selezionare gli appalti negativi nel tempo in modo equivalente agli esempi positivi



Un bando nella base dati è descritto da...



- CIG (Codice Identificativo Gara)
- CIG Accordo Quadro
- CF_amministrazione appaltante
- Denominazione stazione appaltante
- Centro di costo
- Data pubblicazione bando
- Data scadenza bando
- CPV (oggetto prevalente)
- CVP descrizione
- Importo complessivo gara
- Numero lotti
- Scelta contraente
- Tipo scelta contraente
- Modalità di realizzazione
- Oggetto principale contratto
- Luogo
- Escluso (in deroga alla norma)
- Motivo esclusione

- Codice esito
- Esito
- Data aggiudicazione
- Criterio aggiudicazione
- Importo aggiudicazione
- Numero imprese offerenti
- Ribasso aggiudicazione
- Quadro economico base (importo lavori)
- Quadro economico (importo servizi)
- Quadro economico (importo forniture)
- Quadro economico (importo sicurezza)
- ...
- Data stipula contratto
- Data inizio effettiva
- Data termine contrattuale
- Data effettiva ultimazione
- Quadro economico fine (importo lavori)

Esempi



Tipo scelta contraente:

Affidamento in economia

Procedura ristretta

Procedura aperta

Procedura negoziata previa pubblicazione

Procedura negoziata senza previa indizione di gara (ex art 221 DLgs 163)

Modalità di realizzazione:

Acquisizione in economia

Contratto di concessione di lavori

Contratto d'appalto

Contratto di concessione di servizi e/o forniture

Oggetto principale contratto:

Servizi, lavori, forniture

• Escluso:

- (in deroga alla norma Codice dei contratti pubblici):
- Art. 20, 21, 25 D.Lgs. 163/2006

• Criterio aggiudicazione:

- Prezzo più basso
- offerta economicamente più vantaggiosa

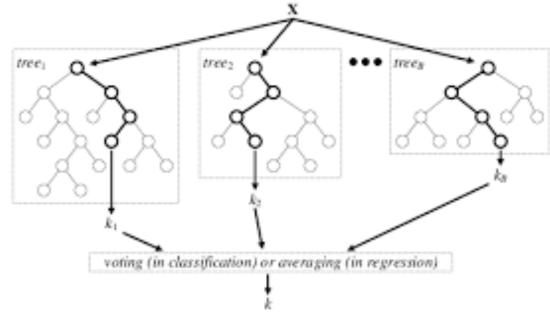


Il Modello usato per la Classificazione



Abbiamo scelto **Random Forest** * come Modello di Classificazione: è molto usato e tende a dare risultati migliori e robusti.





(*)

- Ho, Tin Kam (1995). «Random Decision Forests».

Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

- Breiman L (2001). «Random Forests». Machine Learning. 45 (1): 5–32.



Altre caratteristiche fornite in input:



- Tra le caratteristiche descrittive dei bandi, date di input al nostro modello, abbiamo calcolato anche le differenze tra le date:
 - Differenza tra la Data di Scadenza Offerta e la Data di Pubblicazione
 - Differenza tra Data Aggiudicazione Definitiva e Data Pubblicazione;
 - Differenza tra Data Stipula Contratto e Data Aggiudicazione Definitiva.

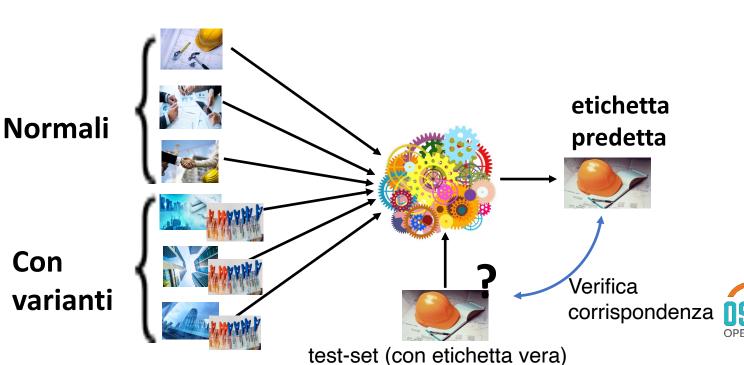


Test accuratezza modello predittivo



- Per verificare l'accuratezza delle predizioni del modello ci si prepara il test set, scelto in maniera casuale (ad es, in proporzione 70%-30%) dall'insieme di esempi disponibili e con etichetta nota
- Gli esempi di test non vengono forniti all'algoritmo di apprendimento in fase di addestramento (training) ma vengono usati per verificare la precisione del modello.

training-set (con etichetta vera)



Risultati dei modelli



Falsi positivi

C

Ε

E Z

- Il risultato ottenuto dopo l'addestramento, usando 80 stimatori diversi nel classificatore Random Forest, dà i seguenti valori di prestazione in classificazione:
 - Accuratezza: 79%
 - Media (armonica) 80% tra:
 - precisione (correttezza) 75%
 - recall (completezza) 87%
 - Matrice di confusione:
 - Risultati simili si hanno con un numero di stimatori minori (20-40-60)

	Risposte Negative	Risposte Positive	
Veri Negativi	10984	4486	
Veri Positivi	2011	13495	87%
		75%	

Falsi regativi



Le caratteristiche che sono risultate più importanti sono:



- DATE DIFF STIP AGGIUDICAZIONE
- QE FINE IMPORTO LAVORI
- TIPO SCELTA_CONTRAENTE = "Procedura selettiva ex art. 238 c.7, D.Lgs. 163/2006"
- QE_BASE_IMPORTO_LAVORI
- NUM IMPRESE OFFERENTI
- MODALITA_REALIZZAZIONE = "Contratto d'appalto"
- MODALITA_REALIZZAZIONE = "Accordo quadro/Convenzione"
- QE_BASE_SOMME_A_DISPOSIZIONE
- CPV=45454100-5 e 45454000-4("Lavori di restauro" e "Lavori di ristrutturazione")
- MODALITA_REALIZZAZIONE = "Acquisizione in economia"
- MODALITA_REALIZZAZIONE = "Contratto d'appalto discendente da Accordo quadro/Convenzione senza successivo confronto competitivo"
- QE_FINE_SOMME_A_DISPOSIZIONE





Riconoscimento appalti con contenzioso presso la Giustizia Amministrativa



- Raccolta dei dati per costruire un modello di classificazione che sia capace di predire per tutti i bandi pubblicati dalla pubblica amministrazione se daranno luogo a contenzioso tra le parti ricorsi per irregolarità, contestazioni e richiesta di annullamento gare, ecc) presso la Giustizia Amministrativa.
- Per ottenere il modello, abbiamo preso il database di ANAC (dal 2010 ad oggi) delle gare (oltre 5 milioni di casi)



Processamento dei dati



Consultazione del portale della Giustizia Amministrativa (Sezione Decisioni e Pareri)

Ricerca tramite un programma (Spyder o BOT) da noi creato appositamente tramite l'uso della libreria software *Mechanize*, dei bandi di gara presenti sul portale della

Giustizia Amministrativa

Giustizia Amministrativa Consiglio di Stato Tribunali Amministrativi Regionali							Ita~
Portale del cittadino		Portale dell'avvocato			Portale del magistrato		
	Giustizia Amministrativa	Consiglio di Presidenza	Consiglio di Stato	CGA Sicilia	TAR	Studi e approfondimenti	Decisioni e pareri

Decisioni e Pareri

Ricerca libera:	7100568613	F	Ricerca Avanzata	
Risultati per pagina:	20 \$			
Tipo Provvedimento:	\$			
Sede:	\$			
Anno e numero provvedimento:	*			
		Annulla Cerca		
Trovati 1 risultati				
Risultati da 1 a 1 di 1 totali - Pagine	: 1		Filtra per:	



Risultato del BOT



Abbiamo trovato sul Portale della Giustizia Amministrativa:

5024 appalti (0.1%)

- 3159 di tipo "FORNITURE" (63%)
- 1130 di tipo "LAVORI" (22.5%)
- 735 di "SERVIZI" (14.5%).
- Etichettiamo questi appalti come positivi



Processamento dei dati



- Abbiamo preso altri 5024 bandi di gara che abbiamo etichettato come negativi in modo che il data set sia bilanciato
- Affinchè l'algoritmo di apprendimento del modello impari non solo dagli esempi positivi ma anche per confronto con i negativi
- Abbiamo scelto i bandi di gara negativi in modo casuale, ma siamo stati attenti a mantenere sui negativi lo stesso rapporto relativo all'Oggetto Principale del Contratto:
 - 63% sono "FORNITURE",
 - 22.5% sono "LAVORI",
 - 14.5% sono "SERVIZI".

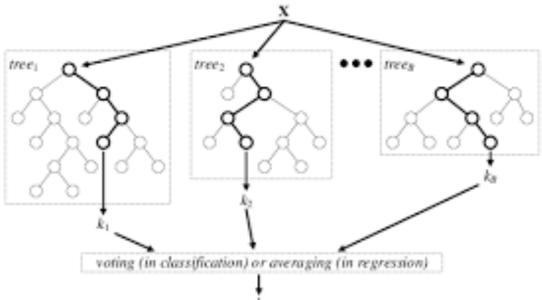


Il Modello usato per la classificazione



- Completamento del dataset con la descrizione dei bandi etichettati come positivi e negativi, con ulteriori caratteristiche delle imprese aggiudicatarie.
 I dati serviranno come input per costruire ed allenare il modello predittivo
- Addestramento di Random Forest come Modello di Classificazione: è molto usato e tende a dare risultati migliori e robusti.







Risultati dei modelli



- Il risultato ottenuto dopo l'addestramento, usando 50 stimatori (alberi) diversi nel classificatore Random Forest, dà i seguenti valori di prestazione in classificazione:
 - Accuratezza: 91.8%
 - Media tra la precisione (correttezza) e la recall (completezza): 91%
 - Matrice di confusione:

	Risposte Negative	Risposte Positive	
Veri Negativi	1367	163	
Veri Positivi	82	1403	



Risultati dei modelli



- Il risultato ottenuto dopo l'addestramento, usando 50 stimatori (alberi) diversi nel classificatore Random Forest, dà i seguenti valori di prestazione in classificazione:
 - Accuratezza: 91.8%
 - Media tra la precisione (correttezza) e la recall (completezza): 91%

Matrice di confusione:

Falsi positivi

	Risposte Negative	Risposte Positive	
Veri Negativi	1367	163	
Veri Positivi	82	1403	

Falsi negativi



Risultati dei modelli



- Il risultato ottenuto dopo l'addestramento, usando 50 stimatori (alberi) diversi nel classificatore Random Forest, dà i seguenti valori di prestazione in classificazione:
 - Accuratezza: 91.8%
 - Media tra la precisione (correttezza) e la recall (completezza): 91%
 - Matrice di confusione:

	Risposte Negative	Risposte Positive	
Veri Negativi	1367	163	
Veri Positivi	82	1403	94.4%
		89.6%	

OMPLETEZZA



Quali caratteristiche sono importanti per la predizione?



IMPORTO_LOTTO	1.496607e-01
IMPORTO_COMPLESSIVO_GARA	1.286814e-01
DATE_DIFF_SCAD_PUB	1.094781e-01
TIPO_SCELTA_CONTRAENTE_Procedura aperta	8.627650e-02
TIPO_SCELTA_CONTRAENTE_Affidamento in economia - affidamento diretto	4.419985e-02
MODALITA_REALIZZAZIONE_Acquisizione in economia	4.349765e-02
DATE_DIFF_AGG_PUB	3.843708e-02
IMPORTO_AGGIUDICAZIONE	2.934897e-02
MODALITA_REALIZZAZIONE_Contratto d'appalto	2.866990e-02
N_LOTTI_COMPONENTI	2.548224e-02
NUM_IMPRESE_OFFERENTI	2.243243e-02
ESITO_0	1.284519e-02
ESITO_Aggiudicata	1.282269e-02
RIBASSO_AGGIUDICAZIONE	1.139650e-02
TIPO_SCELTA_CONTRAENTE_Procedura negoziata senza previa pubblicazione	9.712806e-03
TIPO_SCELTA_CONTRAENTE_Affidamento in economia - cottimo fiduciario	9.496342e-03
DATE_DIFF_STIP_AGG	8.948370e-03
OGGETTO_PRINCIPALE_CONTRATTO_LAVORI	7.501162e-03
TIPO_SCELTA_CONTRAENTE_Affidamento diretto in adesione ad accordo quadro/convenzione	6.777853e-03
MODALITA_REALIZZAZIONE_Accordo quadro/Convenzione	6.444676e-03



Quali caratteristiche non sono importanti?



- Le caratteristiche dei bandi di gara meno importanti sono:
 - CPV
 - la tipologia del prodotto
 - MOTIVO ESCLUSIONE
 - luogo della PA



Ulteriori esperimenti



Extreme Gradient Boosting è un altro modello ora molto in voga (evoluzione delle Random Forest)

Ha ottenuto risultati molto simili di accuratezza (90.94%)

- Friedman J, Hastie T, Tibshirani R, et al. (2000). "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." The annals of statistics, 28(2), 337–407.
- Friedman JH (2001). "Greedy function approximation: a gradient boosting machine." Annals of Statistics, pp. 1189–1232.
- xgboost: eXtreme Gradient Boosting a software package implemented in R and in Python



Variante classificatore con il pieno uso dei campi descrittivi



Aggiunta della descrizione dell'oggetto del contratto

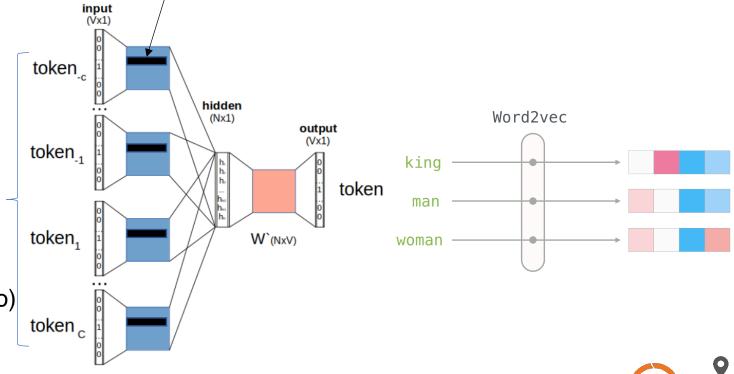


- Abbiamo ulteriormente arricchito il classificatore fornendo caratteristiche aggiuntive in input:
- la descrizione testuale dell'oggetto del contratto

• non è assegnata tale e quale ma trasformata in dato numerico (vettore o sequenza di componenti numeriche) detta word embedding*, tramite il metodo Doc2vec

Doc2vec è "simile" a Word2vec
e fornisce una rappresentazione
del testo da parte di una rete
neurale artificiale (a due strati),
addestrata a partire dalle
descrizioni testuali dei bandi
di gara, date in input.

parole della frase (contesto)



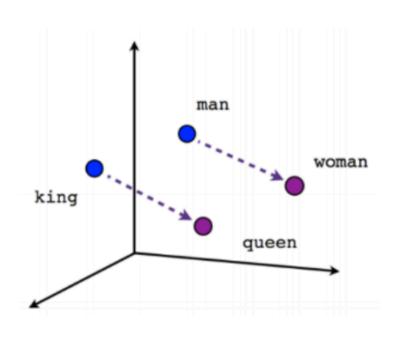
(*) Thomas Mikolov, *Efficient estimation of word representations in vector space*, in *Proceedings of NIPS*, 2013.

Software libraries for computing word embeddings are Word2vec, Gensim, Glove, Deeplearning4j, fastText

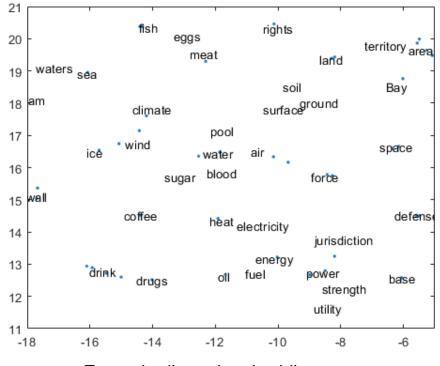
Word embedding



La rappresentazione a *word embedding* permette di calcolare la **similarità semantica** tra due testi tramite un'operazione numerica di calcolo della distanza **Esempio**:



king = queen - woman + man



Esempio di word embedding (fastText) in lingua Inglese



Risultati dei modelli di classificazione



- Il risultato del classificatore ottenuto dopo l'addestramento sul dataset contenente anche i word embeddings degli **oggetti dei bandi di gara**, usando **50** stimatori (alberi) nel classificatore Random Forest, dà i seguenti valori di prestazione in classificazione:
 - Accuratezza: 97% (era 91.8% senza oggetto del bando)
 - Recall (completezza): **97**% (era 94.4%)
 - Precisione (correttezza): 96% (era 89.6%)
 - Media tra la precisione e recall: 97% (era 91%)

Conclusione:

i word embeddings sulla descrizione di un bando/gara di appalto hanno un contenuto informativo notevole e migliorano in modo significativo la predizione della presenza di ricorso su un bando di gara





Analisi esplicativa del contenuto dei bandi di gara e dei contratti

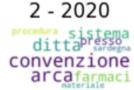




Words clouds degli oggetti degli appalti, nel tempo (mese per mese):

word clouds

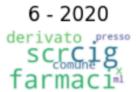
















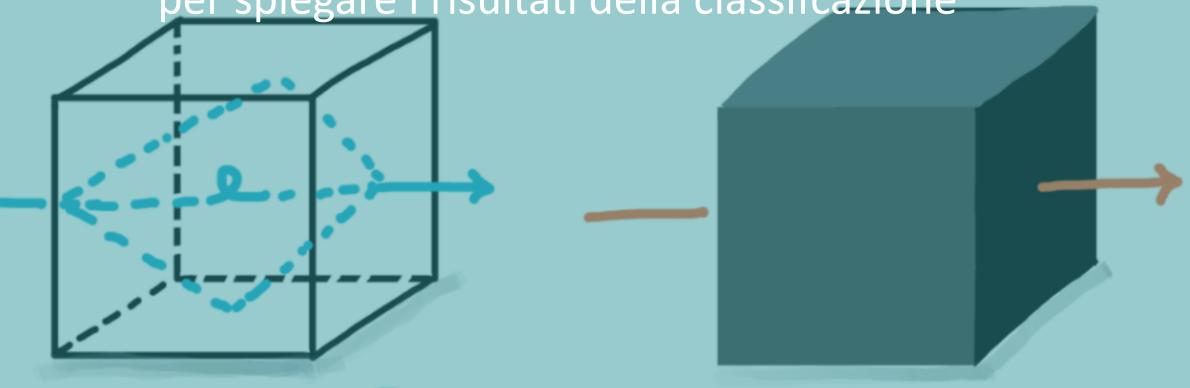








Utilizzo di strumenti di Explainable Al per spiegare i risultati della classifcazione



EXPLAINABLE (AI)
Interpretable Machine Learning

A Guide for Making Black Box Models Explainable. By Christoph Molnar







Spiegazione della presenza di ricorso con il metodo di regressione logistica



Caratteristiche che impattano positivamente sulla probabilità di ricorso

Durata gara:

per ogni aumento di **1 mese** della durata della gara ⇒ la probabilità che vi sia ricorso aumenta di 1.2 volte, per ogni **anno** aggiuntivo ⇒ probabilità di ricorso aumenta di 8.84 volte

Importo lotto: per ogni aumento di 1 mln ⇒ probabilità di ricorso aumenta di 1.05 volte per un aumento di 10 mln ⇒ probabilità aumenta di 1.61 volte per un aumento di 100mln ⇒ probabilità aumenta di 120 volte

Procedure di selezione:

per una procedura **aperta** ⇒ la probabilità di ricorso aumenta di 5.64 volte rispetto a un affidamento diretto per una procedura **ristretta** ⇒ la probabilità di ricorso aumenta di 1.81 volte « « « « per una procedura **negoziata** ⇒ la probabilità di ricorso aumenta di 1.55 volte « « « «

Per i **Servizi** ⇒ la probabilità di ricorso aumenta di 2.8 volte rispetto alle forniture Per le procedure con l'**offerta econom. più vantaggiosa** ⇒ la probabilità di ricorso aumenta di 2.22 volte rispetto alla scelta con il prezzo più basso

Ribasso: ogni 10% di ribasso ⇒ la probabilità di ricorso aumenta di 1.15 volte **A.T.I.**: ⇒ la probabilità di ricorso aumenta di 1.92 volte rispetto alle imprese singole **Classe di rating dell'impresa**:

1° classe ⇒ la probabilità di ricorso aumenta di 1.08, 10° classe ⇒ la probabilità di ricorso aumenta di 2.1 rispetto alla 1° classe

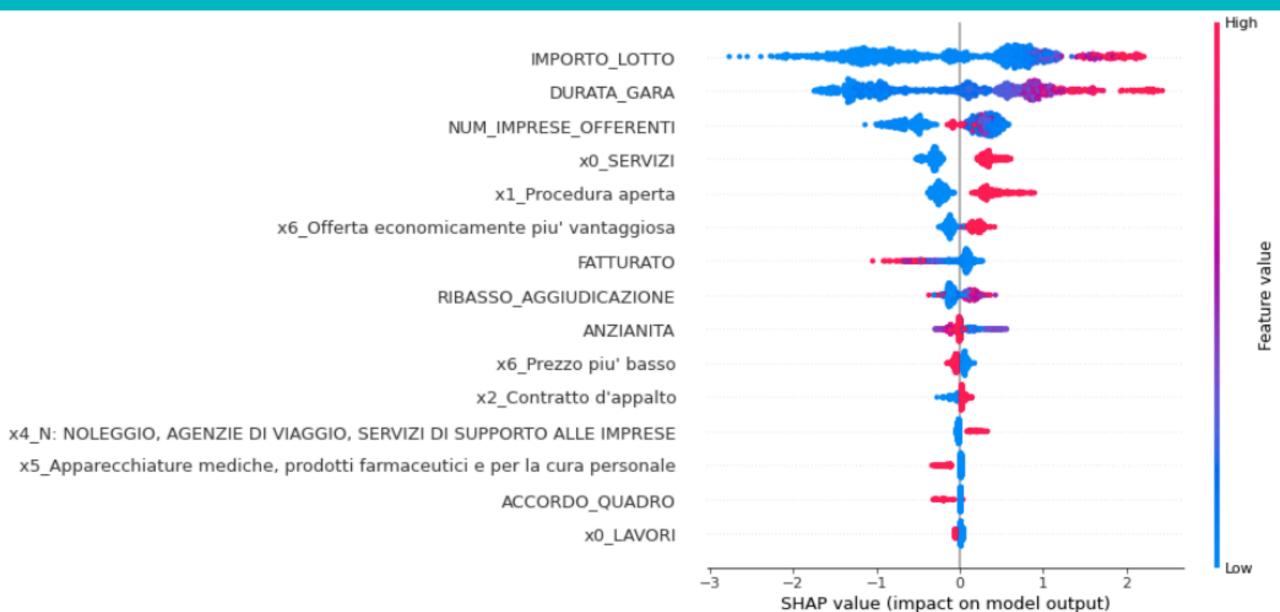
Caratteristiche che impattano negativamente sulla probabilità di ricorso

Accordo quadro: ⇒ la probabilità di ricorso scende di 0.36



Spiegazioni globali del modello a black box (XGBoost)



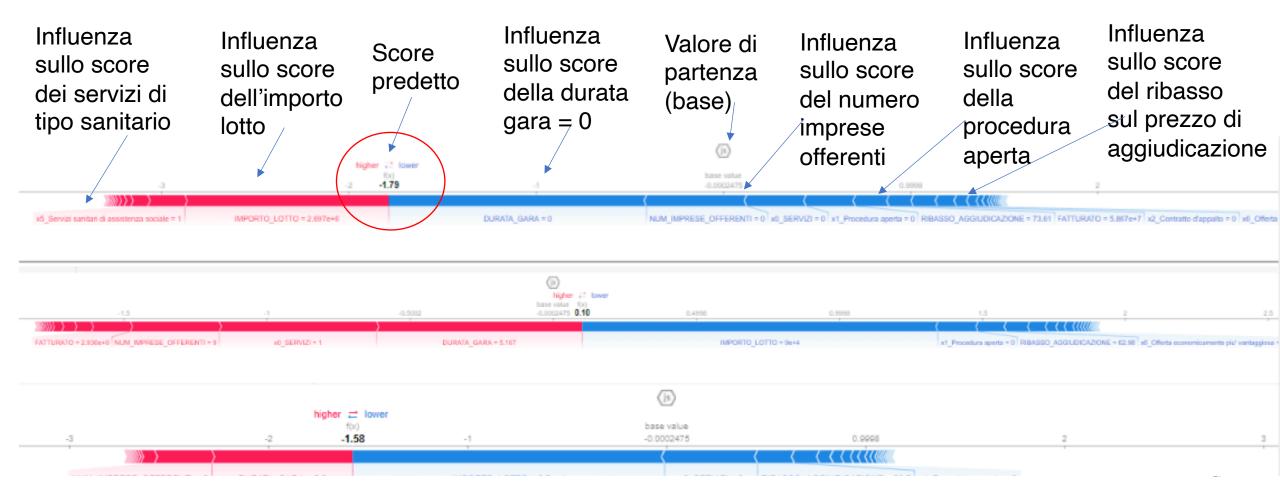


SHAPLEY value: spiegazioni locali del modello a black box (XGBoost)



Spiegazione di come si modificherebbe lo score predittivo per un singolo esempio se le variabili avessero valore diverso

Il colore, e l'ampiezza delle barre evidenziano la direzione del cambiamento



SHAPLEY values: spiegano l'importanza delle caratteristiche degli esempi



- Shapley values *:
 Basati su osservazione dell'output del sistema di Al se sollecitato in input con esempi simili (perturbazioni) di quelli precedenti
- Si ispirano alla teoria dei giochi

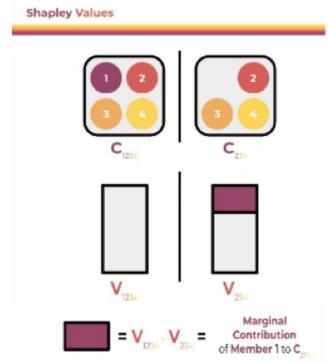
 Rappresentano un esempio come se fosse una squadra e cancellano una caratteristica di input dall'esempio per verificare l'effetto che fa la sua

mancanza sull'output

(*)

- Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2.28 (1953): 307-317

- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017
- We used the *shap* software package in scikit-learn Python library







CONCLUSIONI



Conclusioni



- Abbiamo analizzato i dati dei bandi di gara pubblicati dalle pubbliche amministrazioni
- Analisi dell'indicatore numero di bandi ripetutamente aggiudicati alla stessa impresa dalla stazione appaltante
- Analisi predittiva per riconoscere in anticipo, al tempo di aggiudicazione, se un bando di gara potrebbe dar luogo a:
- varianti in corso d'opera
- contenziosi presso la Giustizia Amministrativa
- I risultati sono buoni, i metodi sono promettenti e possono essere spiegati con i metodi di XAI

Lavori futuri

- Costruzione di altri BOT per confrontare i dati con altre fonti (TED) e analizzare la qualità dei dati
- Analisi dei bandi andati deserti
- Analisi delle opera incompiute
- Analisi dei ritardi nei pagamenti tramite incrocio con SIOPE (pagamenti elettronici)































