**Project Orion**

**Report prototipo 1**

**Predizione delle intenzioni di voto da variabili demografiche e valoriali**

**Glossario**

*Target*: variabile predetta (e.g., intenzioni di voto)

*Feature*: variabile predittiva (e.g., età)

**Formalizzazione modelli predittivi**

*target* ~ *feature 1*, …., *feature N*

**Scopo**

Lo scopo dell’esperimento è costruire un modello mi machine learning in grado di utilizzare variabili demografiche e 5 variabili motivazionali per **predire le** **intenzioni di voto** con la massima accuratezza permessa **dai** **dati valoriali del Walden**.

Il report si articola nelle seguenti sezioni:

1. Descrizione e pulizia dati.
2. Costruzione modello di base da variabili demografiche.
3. Individuazione di 5 variabili valoriali da inserire nel modello.
4. Costruzione e valutazione modello.
5. **Descrizione e pulizia dati**

I microdati provengono dal database “**Walden integrated”**, comprendente i risultati delle indagini annuali dal 1997 al 2019, con 39781 risposte e 826 variabili.

Premesso che l’associazione tra le intenzioni di voto e le variabili demografiche e valoriali può cambiare nel tempo, sono stati utilizzati solo i dati più recenti, dal 2017 al 2019. Dopo aver rimosso il dato meno recente dei rispondenti che hanno risposto per più anni all’indagine, sono risultati **4504 rispondenti unici**.

Per migliorare la potenza predittiva del modello è stato scelto di ridurre il numero di categorie presenti nella variabile target **intenzioni di voto**, accorpando i partiti Lega, FdI, FdI-Alleanza Nazionale e FI come **Destra** e Sinistra italiana, Potere al Popolo, Rifondazione comunista e La Sinistra come **Sinistra**, risultando nella distribuzione sottostante:

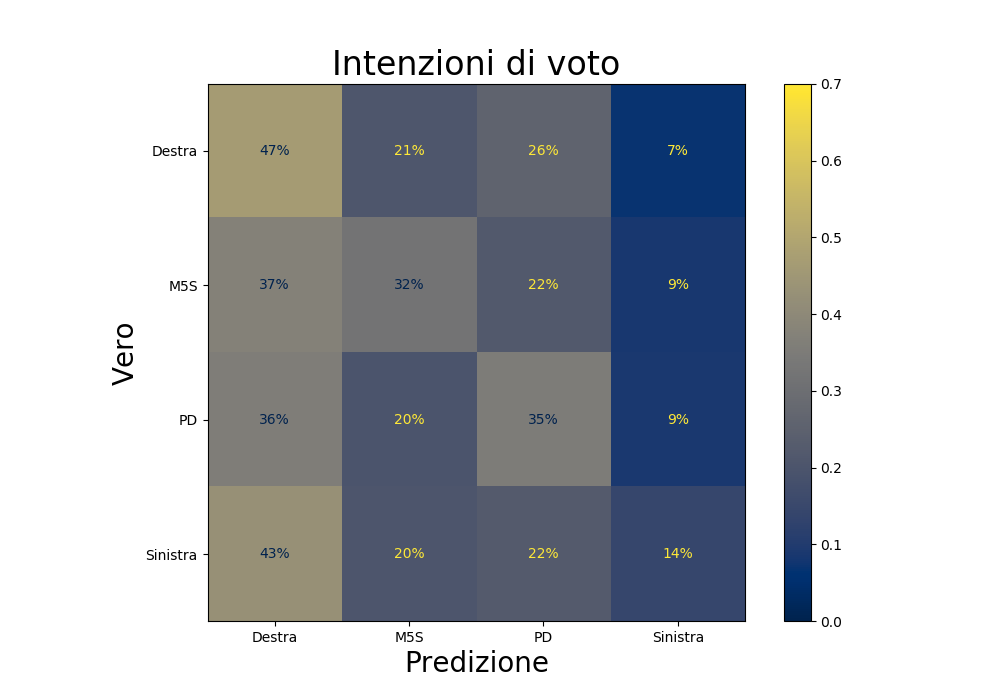
|  |  |
| --- | --- |
| **Partito** | **Intenzioni di voto** |
| Destra | 1086 |
| PD | 972 |
| M5S | 869 |
| Sono indeciso | 695 |
| Sinistra | 186 |
| Non andrei a votare | 157 |
| Preferisco non rispondere | 150 |
| Scheda bianca/nulla | 117 |
| +Europa | 86 |
| Altro partito | 62 |
| Liberi e Uguali | 56 |
| Verdi | 30 |
| Italia dei Valori | 9 |
| Noi con l'Italia UDC | 8 |
| Nuovo Centro Destra con UDC e PPI | 8 |
| Scelta Civica | 7 |
| altro partito di area di governo (SVP, Centro Democratico....) | 4 |
| Italia Unica di Corrado Passera | 2 |
| **TOTALE** | **4504** |

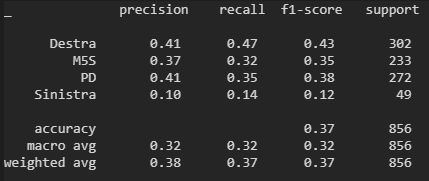
Per costruire il modello predittivo, sono stati utilizzati solo i partiti con più di 100 intenzioni di voto, escludendo i ‘sono indeciso’, ‘preferisco non rispondere’ e ‘scheda bianca/nulla’.

1. **Costruzione modello di base**

Per il modello di base abbiamo utilizzato l’algoritmo di classificazione **Random Forest Classifier** del pacchetto Python **Scikit-learn** (sklearn). In breve, il modello costruisce centinaia di alberi decisionali[[1]](#footnote-1), ognuno con una predizione. La predizione più votata è quella vincente.

**Modello base:** predizione delle intenzioni di voto con solo variabili demografiche [regione, età, sesso, professione, ampiezza centro]





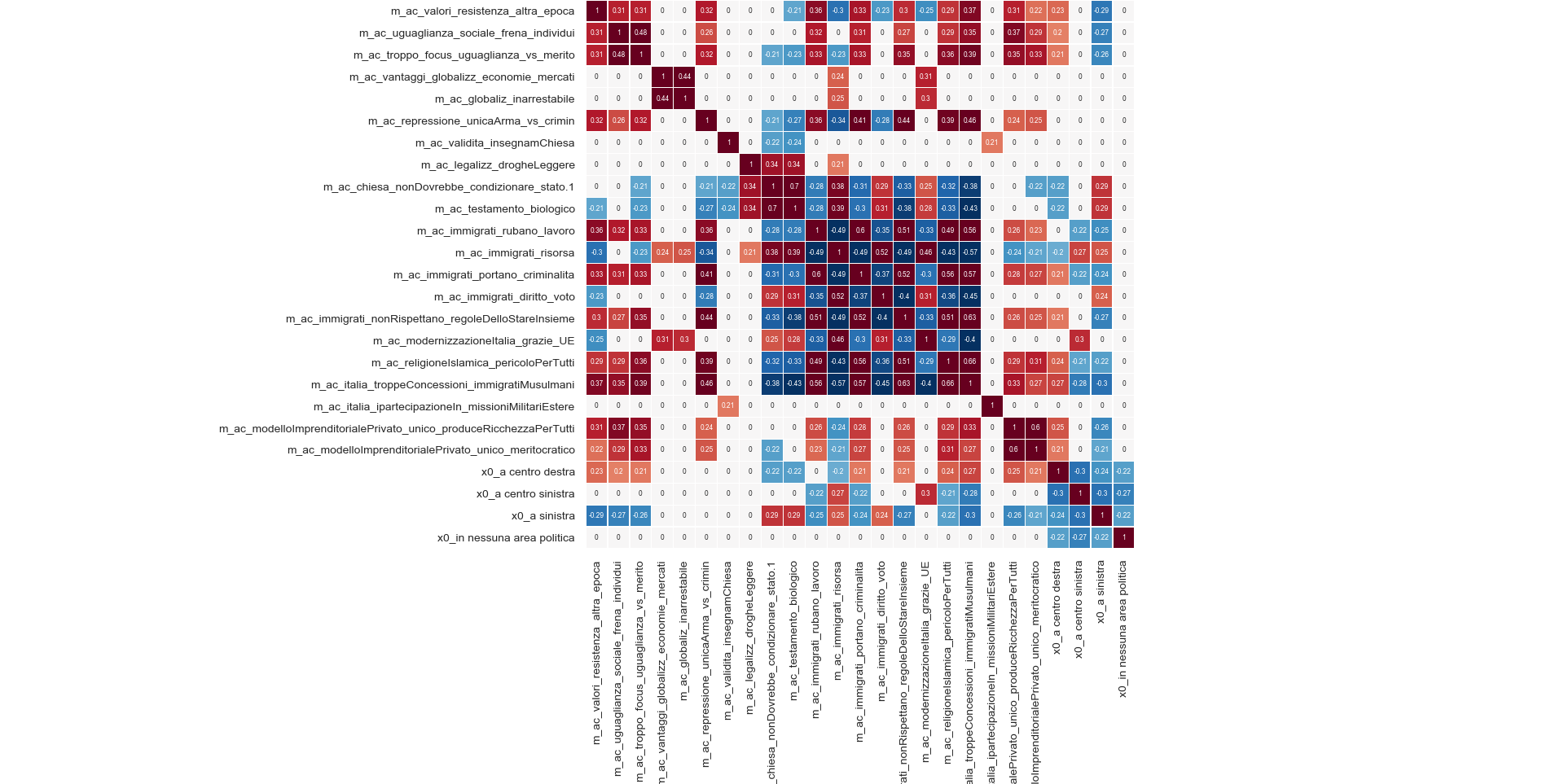


Selezione domande di accordo (m\_ac\_) che correlano con intenzioni di voto

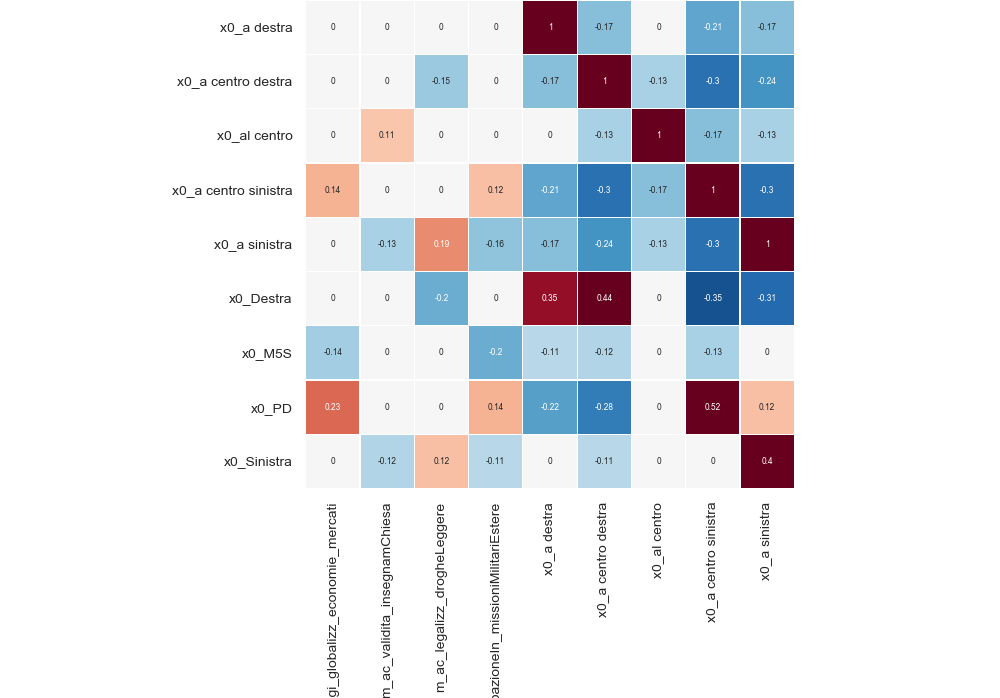
1. **Individuazione di 5 variabili valoriali da inserire nel modello**

Individuazione features più importanti con Random Forest.

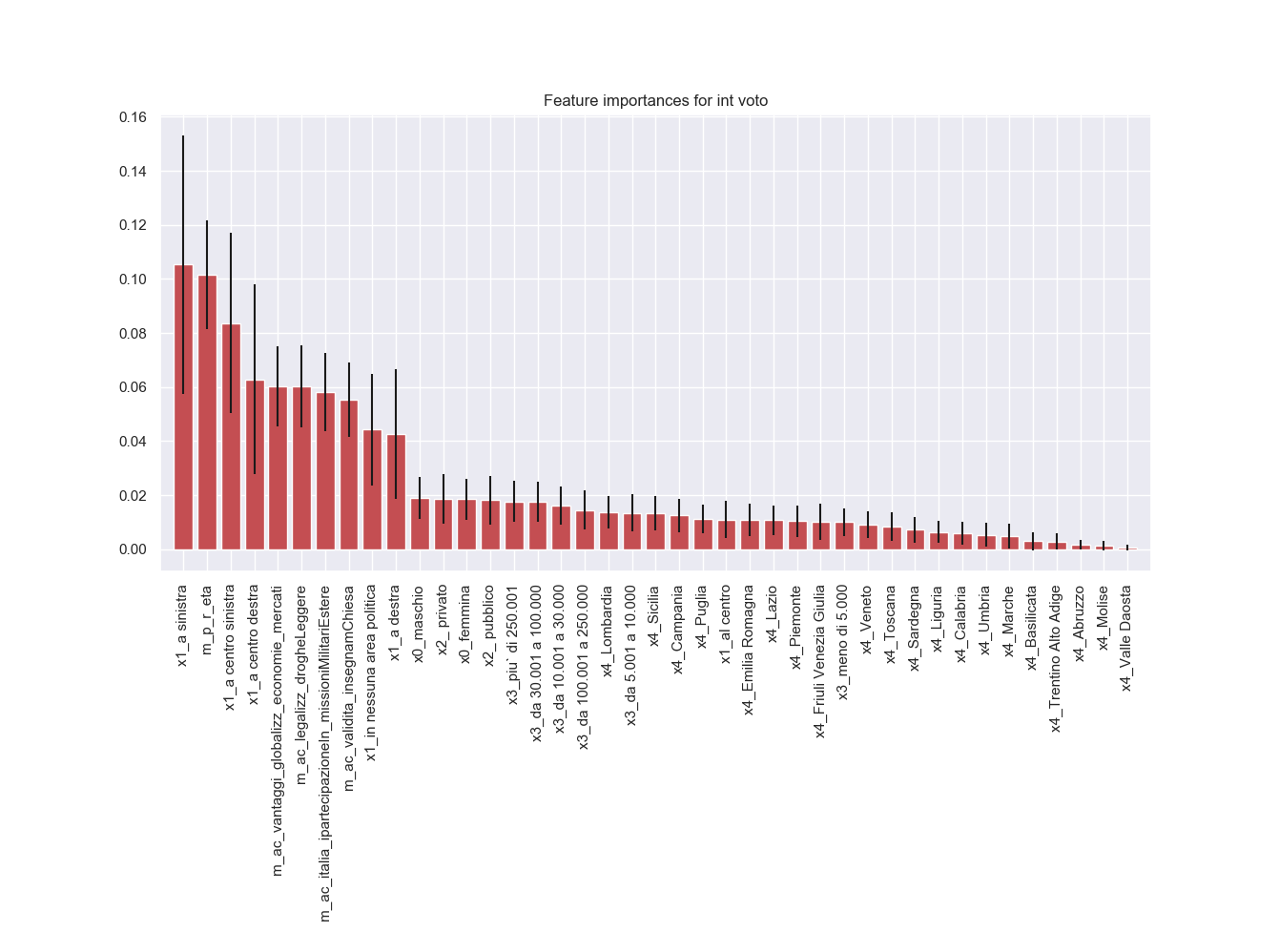
Check correlazione reciproca e selezioni variabili scorrelate con autocollocazione (ultime 4 colonne).  
*Correlazioni a 0 se coefficiente <.1.*



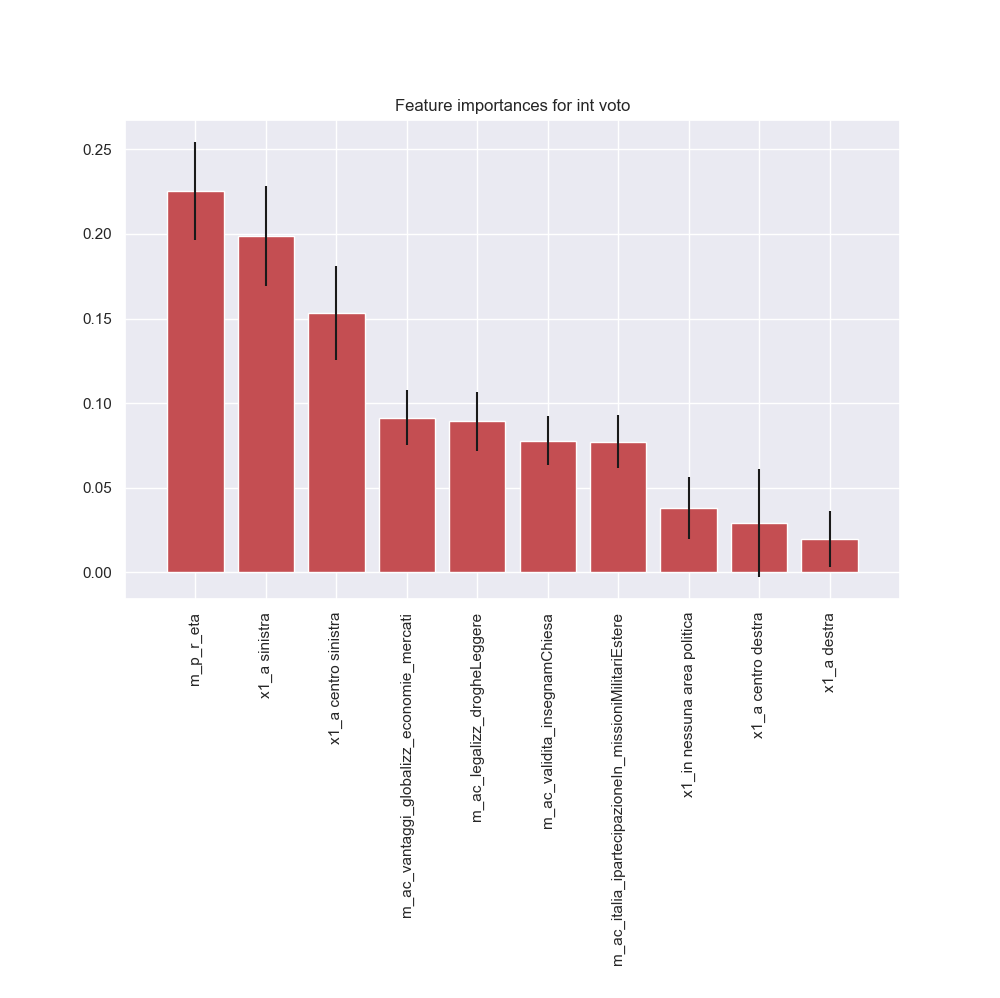
Check correlazioni nelle variabili più informative:



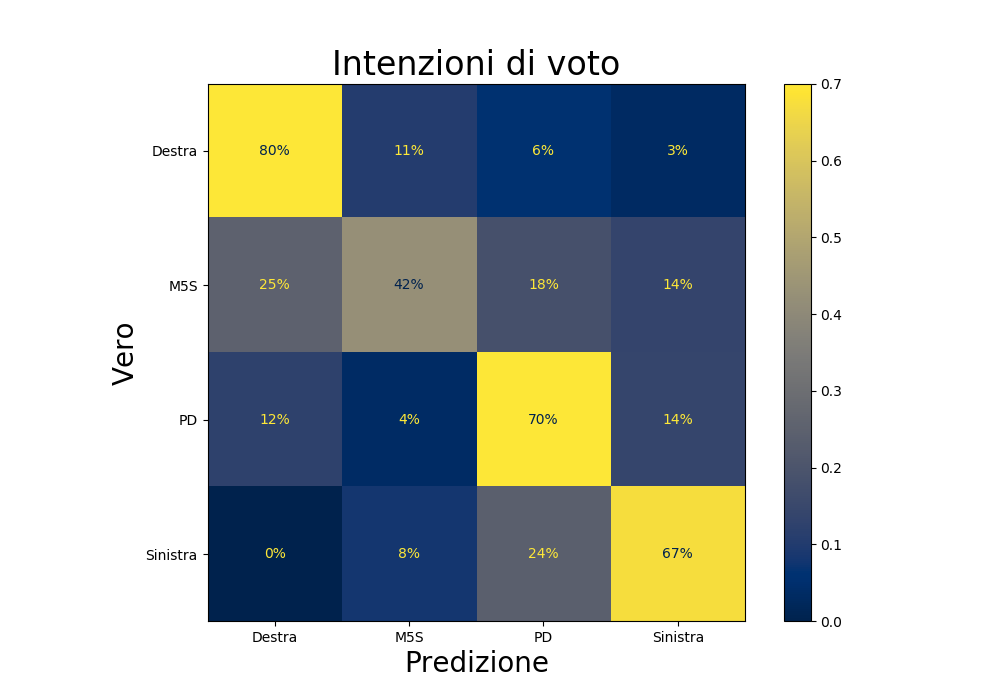
Features prima della selezione delle più rilevanti.

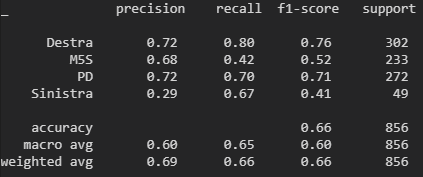


Importanza delle features più rilevanti:



Performance miglior modello

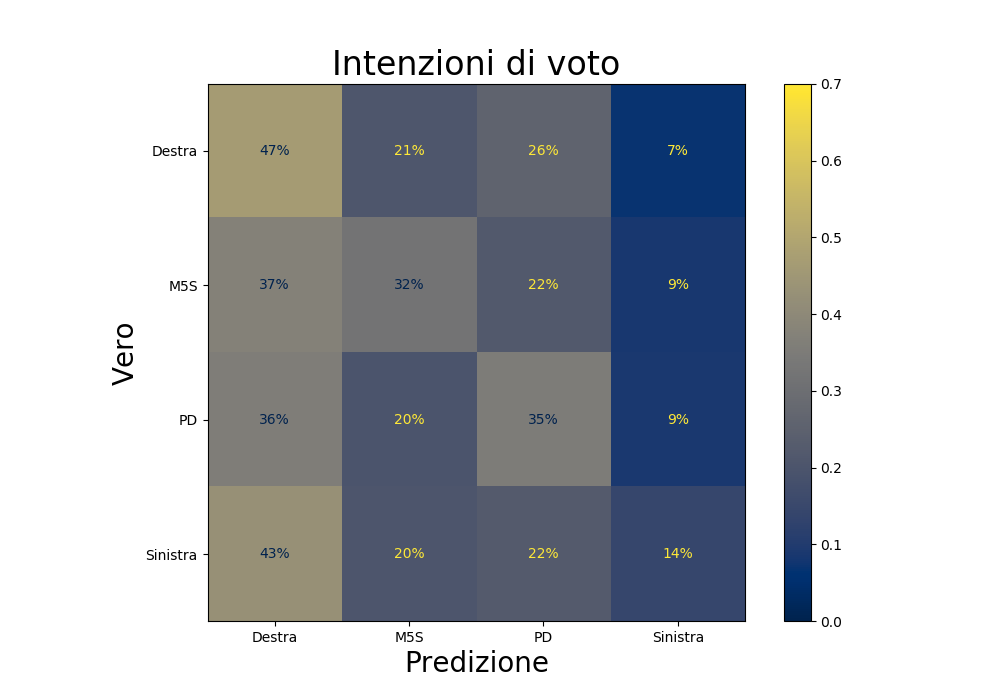
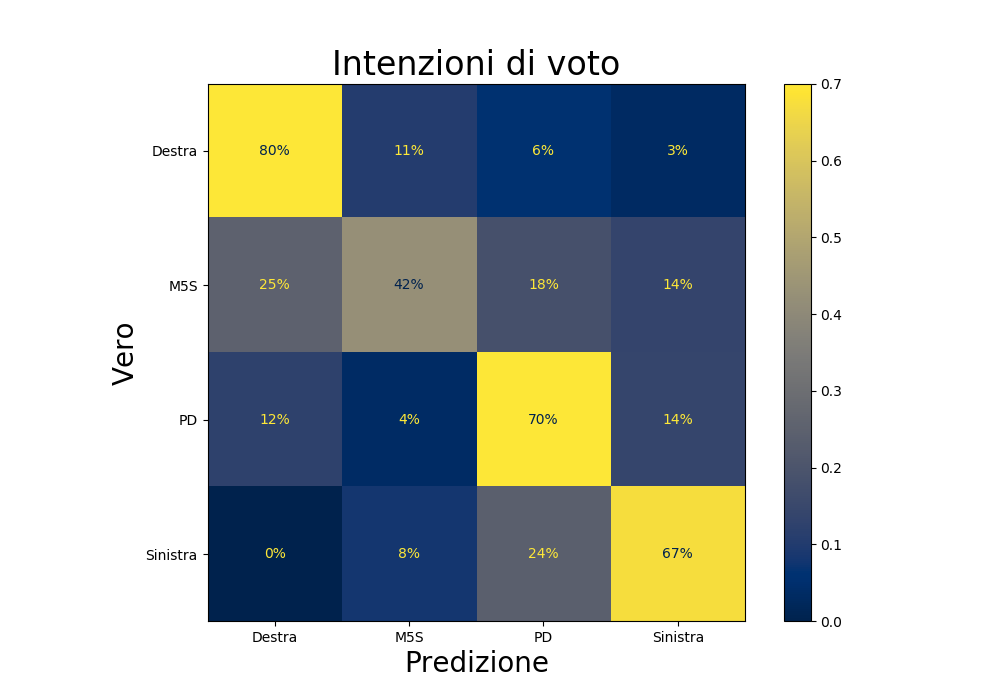


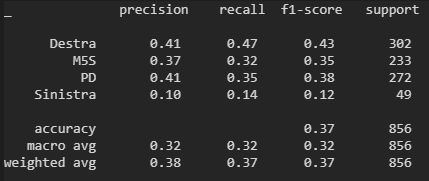
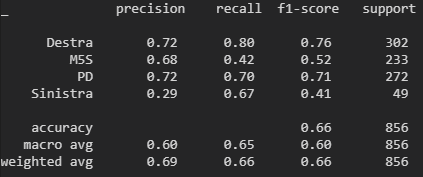




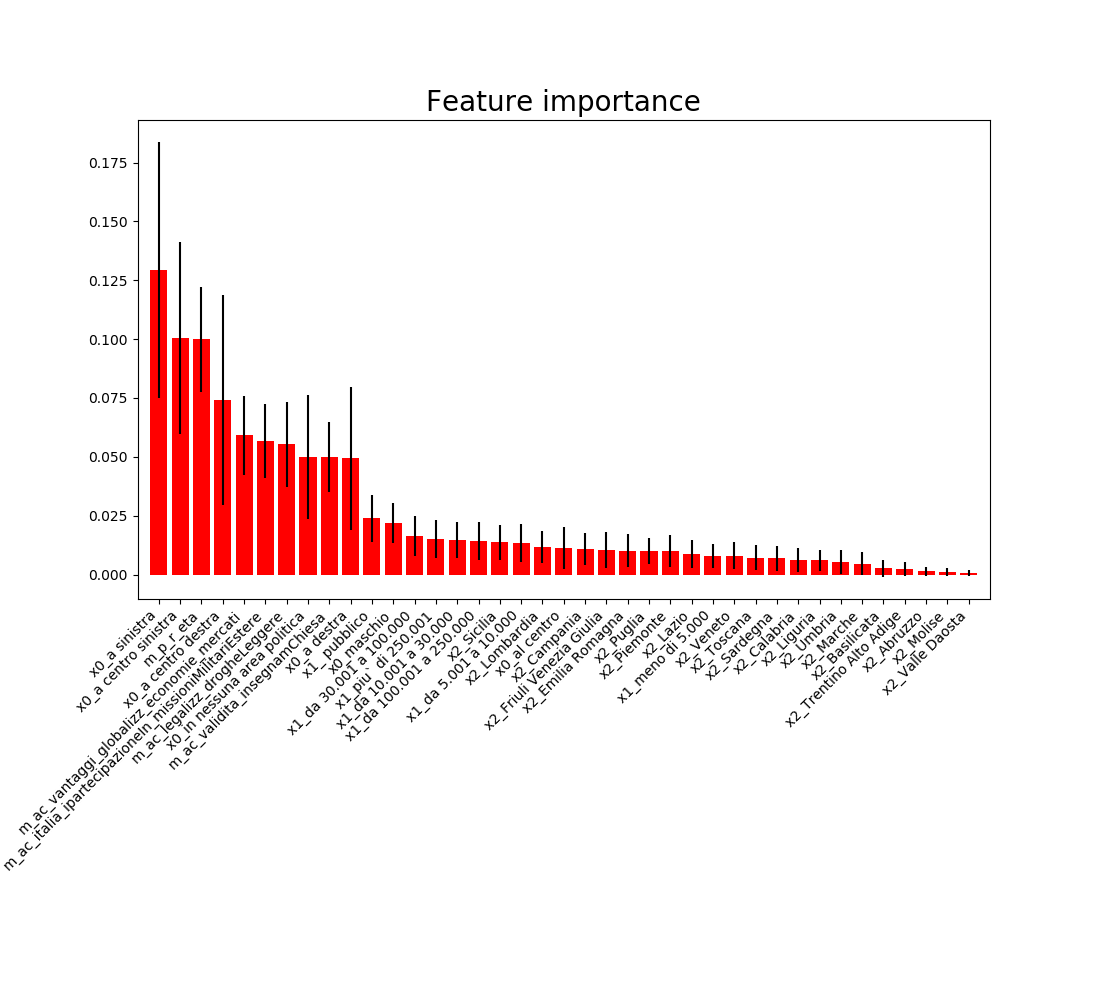
**Modelli a confronto**

**Base Con Valori**

Importanza features modello con Valori



1. Ogni albero decisionale viene costruito selezionando a caso e poi confrontando tra loro un gruppo di variabili. Tra queste viene poi individuata quella che ‘spacca’ di più sulla variabile target. Il processo viene ripetuto per ogni livello della spaccatura/ramificazione, fino al raggiungimento di una determinata profondità. [↑](#footnote-ref-1)