Documento di *onboarding*

*Versione marzo 2022*

Indice dei contenuti

[Indice dei contenuti 1](#_Toc88138202)

[Introduzione 1](#_Toc88138203)

[Rete aziendale, email, repository e la VPN 1](#_Toc88138204)

[I dati SWG, la Community, le rilevazioni e le principali elaborazioni 2](#_Toc88138205)

[Ponderazione e ricodifiche 2](#_Toc88138206)

[Data visualisation 3](#_Toc88138207)

[Arricchimento dati 3](#_Toc88138208)

[Tecnologie utilizzate 4](#_Toc88138209)

[APPENDICE 5](#_Toc88138210)

# Introduzione

Il presente documento serve ad aiutare l’onboarding dei nuovi membri del team Rachael, illustrando le risorse a nostra disposizione e i principali processi di lavoro.

L’obiettivo principale di Rachael è aiutare i clienti a comprendere in maniera veloce ed efficace i propri dati, arricchendoli ed integrandoli con le basi dati SWG e con dati pubblici. All’interno di questo obiettivo generale, è nostra intenzione creare un gruppo e ambiente di lavoro dove, nonostante gli inevitabili alti e bassi, potremo imparare cose nuove, divertirci e perseguire i nostri obiettivi con il massimo della professionalità!

Per poter iniziare questo percorso, sarà necessario innanzitutto connettersi ai nostri sistemi…

# Rete aziendale, email, repository e la VPN

Per accedere alla **rete aziendale** (WIFIMAIN) è necessario richiedere il rilascio delle credenziali al reparto IT. Queste permettono anche l’accesso al nostro **Repository** – un’unità di rete da 3 TB utilizzata come spazio di archiviazione remoto e scambio file, simile a Google Drive[[1]](#footnote-1). Tutti i file legati ai nostri progetti (dati, codici, file e altre risorse), indipendentemente da dove vengono sviluppati, vanno SEMPRE salvati al suo interno[[2]](#footnote-2). Per poter accedere al repository è necessario chiedere l’abilitazione al reparto IT e seguire le indicazioni presenti nell’Appendice. Per lo scambio dei file è preferibile sempre utilizzare questo mezzo rispetto alla mail. È importante sapere anche che l’accesso al repository è possibile solo se connessi alla rete aziendale, pertanto, per potervi accedere da casa è necessario farlo tramite **VPN**. Il suo utilizzo prevede l’installazione di un software e di apposite credenziali di accesso, rilasciate su richiesta dall’IT assieme alle istruzioni di installazione e configurazione.

Un altro repository che contiene dati di interesse per il lavoro si trova al percorso di rete [**\\file.swg.it\repository**](file:///\\file.swg.it\repository)**.** Questo volume contiene la stragrande maggioranza dei dati aziendali di SWG (protocolli, moduli, infografiche, pubblicazioni ecc). Di utilità per noi c’è sotto 14 la firma aziendale e 11 postazioni telefoniche aziendali.

Oltre all’accesso al WiFi e al repository, verrà fornita una **mail aziendale** gestita tramite il pacchetto GSuite di Google. È utile sapere che dei servizi offerti da Google utilizziamo spesso anche la chat di Hangouts e Meet per le riunioni.

Ad ogni lavoratore di Rachael verrà richiesto di compilare un Time Sheet, un foglio excel in cui il dipendente spiega in maniera sintetica su cosa ha lavorato di settimana in settimana. Questo è importante per poter tenere traccia dei progressi sulla gestione dei progetti e per poter provare di fare Research and Development.

Il teama di Rachael possiede un gruppo su Slack, un software che aiuta a gestire progetti in smart working. E’ consigliato installarlo ed imparare ad usarlo.

Infine, è necessario ottenere e configurare le credenziali per AWS (Amazon Web Services), un calderone di servizi online (che vanno dalla raccolta e gestione di script e database fino ad una loro interazione “automatica”) con il quale viene gestita la pipeline di lavoro per il Sistemone (il progetto cardine di Rachael che prevede di pubblicare sotto forma di abbonamento tutti i dati raccolti da SWG negli ultimi 25 anni). Chiedere a Gabriele.

# I dati SWG, la Community, le rilevazioni e le principali elaborazioni

Uno dei principali asset di Rachael sono i dati provenienti dalle rilevazioni di SWG. Queste possono essere fatte ad hoc per specifiche commesse da parte di clienti, oppure possono essere autoprodotte, ovvero sponsorizzate dalla stessa SWG. Nel secondo caso rientra l’**Osservatorio** (anche chiamato internamente Walden), che consiste in un’indagine annuale (iniziata nel 1997) condotta a livello nazionale su un campione di ~1500 rispondenti (in maggior parte con persone sempre diverse di anno in anno)[[3]](#footnote-3). Settimanalmente viene inoltre fatto l’**Omnibus**, un questionario contenente un mix di domande ad uso di SWG e commesse.

Alla base dei dati di SWG sta la **Community**. Costituita da quasi 60.000 persone rappresenta un bacino di rispondenti che occasionalmente o regolarmente rispondono ai sondaggi. Permette in tal modo di pianificare un campione rappresentativo in base alle variabili desiderate.

Per quanto riguarda la metodologia usata per effettuare le **rilevazioni**/**interviste**, queste possono essere di 4 tipi: **CATI, CAWI, CAPI e CAMI**. Essi si differenziano in base alla modalità utilizzata per contattare ed intervistare il rispondente: chiamata sul fisso, tramite web, di persona/face-to-face, chiamata sul mobile. Gli acronimi stanno rispettivamente per Computer Assisted Telephone/Web/Person/Mobile Interview., dove “computer assisted” indica il fatto che l’intervistatore (assente però nel CAWI) utilizza un software per leggere le domande ed inserire le risposte date direttamente nel computer.

## Ponderazione e ricodifiche[[4]](#footnote-4)

Acciocché un campione sia rappresentativo di una popolazione, è necessario che ne rifletta le caratteristiche demografiche. Ad esempio, se la popolazione italiana è rappresentata al 50% da donne, tale deve essere la proporzione nel campione dell’indagine. Succede di regola che non tutti i rispondenti contattati rispondono al questionario e quindi la proporzione potrebbe sbilanciarsi. In questo caso la Ricerca effettua una **ponderazione** (o pesatura) dei rispondenti. Continuando l’esempio, se nel campione intervistato le donne sono solo il 40%, sarà necessario aumentare il peso delle loro risposte e diminuire quello degli uomini, in modo tale che la voce di entrambi i gruppi corrisponda alla proporzione della popolazione (in questo esempio, 50/50). La pesatura viene svolta solo su indagini campionarie, e, dove presente, è essenziale tenerne conto.

Un’ulteriore elaborazione effettuata dai ricercatori è la **ricodifica** dei dati dei questionari. Per ricodifica si intende l’inserimento di nuove colonne che permettono di identificare i rispondenti in base a raggruppamenti standard di altre variabili o gruppi di variabili. Ad esempio, la variabile contenente il titolo di studio più alto conseguito può venire ricodificata nella variabile “scolarità” a 3 livelli, associando ciascun titolo di studio ad una scolarità bassa, media o alta. Lo stesso può venire fatto con il reddito familiare, ricodificandolo nella variabile “ceto” (alto, medio o popolare).

# Data visualisation

La **data visualisation** è volta alla creazione di dashboard narrative ed interattive, ed è attualmente implementata con il software *Tableau Desktop*[[5]](#footnote-5). Le dashboard vengono sia salvate nella nostra Repository, sia caricate sul servizio cloud di Tableu (*Tableau Online*) e la loro consegna al cliente viene effettuata inviando un link di accesso (ristretto e protetto da credenziali ad-hoc) al cloud.

# Arricchimento dati

Allo scopo di arricchire i database di clienti applichiamo tecniche di machine learning (classificazione e regressione). Nello specifico, per **arricchimento** si intende l’applicazione su un database di un algoritmo predittivo già allenato, il quale predice (e aggiunge) una nuova variabile nel database del cliente a livello individuale o di gruppo/cluster[[6]](#footnote-6). A rendere possibile l’arricchimento sono principalmente i dati rilevati da SWG, sui quali alleniamo e testiamo gli algoritmi predittivi, applicandoli poi a database che presentano le stesse variabili di input al modello (i predittori). I database dei clienti su cui applicare i modelli vengono spesso chiamati CRM (Customer Relationship Management) e di regola contengono tutte le informazioni che le aziende raccolgono sui propri clienti. La presente procedura è soggetta a costanti modifiche e aggiornamenti derivanti dalla ricerca e sviluppo condotta internamente.

# Tecnologie utilizzate

Il gruppo di lavoro al momento utilizza specifiche tecnologie per svolgere le diverse mansioni. Anche se non vincolanti, consigliamo di adottarle o almeno di familiarizzarvi per facilitare la collaborazione.

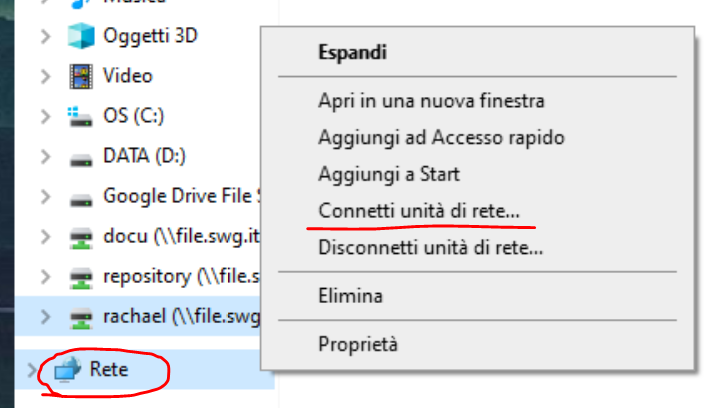
|  |  |
| --- | --- |
| Task | Linguaggi di programmazione (pacchetti) e software |
| Data preprocessing | Python (numpy, pandas), R, MS Excel, Tableau Prep |
| Feature engineering | Python (scikit-learn, factor\_analyzer) |
| Machine Learning | Python (scikit-learn, UMAP) |
| Deep Learning | Python (Keras/TensorFlow, PyTorch) |
| Statistica inferenziale | Python (pacchetti: Scipy), R, SPSS, Jamovi, Jasp |
| Version control | CodeCommit (AWS) |
| Data visualisation | Tableau, Python (matplotlib, seaborn) |
| Code editing | Python: VSCode (<https://code.visualstudio.com/>), Jupyter Lab  R: RStudio (<https://rstudio.com/>) |

# APPENDICE

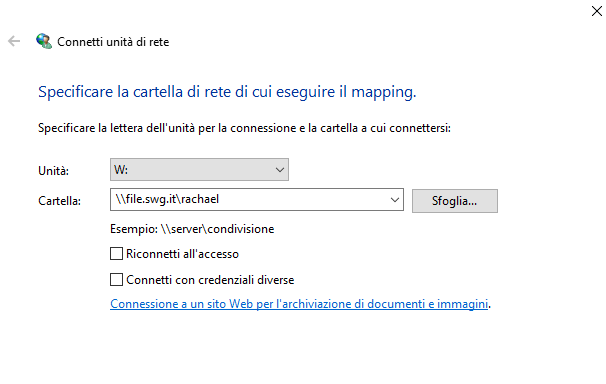
**Connessione al Repository Rachael**

Dopo aver ottenuto dall’IT l’abilitazione ad accedere:

1. Esplora file > click destro su Rete > Connetti all’unità di rete…



1. Specificare un’unità non occupata e inserire l’indirizzo: **\\file.swg.it\rachael-lavoro**

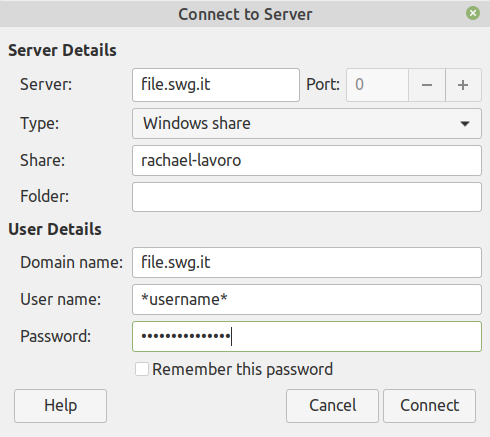


1. Alla comparsa del prompt, immettere le credenziali di accesso alla rete aziendale rilasciate dall’IT

**Connessione al Repository Aziendale (da sistemi Linux)**

Dopo aver ottenuto dall’IT l’abilitazione ad accedere:

* Esplora file > click su File > Connect to Server…
* Compila i campi come riportato e sostituisci le tue credenziali



1. L’unità non permette il version control dei file o di codice, ma rappresenta un comodo e semplice mezzo di scambio e archiviazione. [↑](#footnote-ref-1)
2. Anche SWG ha un suoi repository: il “repository”, che contiene documenti e moduli aziendali, e il “docu” (detto anche “file docu”), dove sono contenuti tutti i lavori/progetti. Nell’ultimo caso, è importante sapere che ogni progetto della Ricerca è associato ad un identificativo a 5 cifre chiamato “numero job” o solo “job”. [↑](#footnote-ref-2)
3. Dai dati dell’Osservatorio è stato anche scritto [un libro](https://www.amazon.it/modo-diverso-1997-2017-cambiata-lopinione/dp/8862506759), ed ogni anno viene prodotto un nuovo report [disponibile nel repository rachael > Risorse > Osservatori SWG (Walden)]. [↑](#footnote-ref-3)
4. Per implementare queste ed altre elaborazioni (e per creare i questionari) i Ricercatori utilizzano i software proprietari Macp3 e Macp4 che utilizzano un proprio formato dati. Solitamente i dati del Macp3/4 ci vengono inviati in formato .sav (utilizzato da SPSS) che il software permette estrarre. Per poter convertire il SAV in CSV e salvarne i metadati abbiamo creato un apposito script in Python. [↑](#footnote-ref-4)
5. L’utilizzo di Tableau necessita di una licenza. Per chi risulta ancora iscritto all’università, è possibile richiedere a Tableau una licenza gratuita di 1 anno, con la possibilità di installare l’applicazione su 2 computer. [↑](#footnote-ref-5)
6. Come in questo caso, spesso la parola “cluster” verrà usata in generale per indicare un gruppo, ovvero il risultato di un raggruppamento o aggregazione in base ad una caratteristica già nota (ad esempio, il cluster dei ventenni). È importante quindi tra di noi non fare confusione con il concetto di cluster utilizzato nel data science per indicare il risultato dell’applicazione di un algoritmo non supervisionato. [↑](#footnote-ref-6)