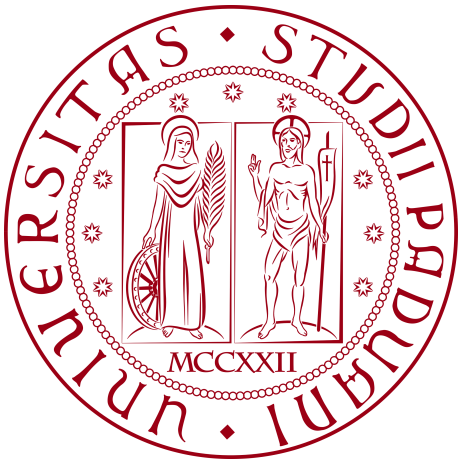
Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea in Informatica



Utilizzo dei Modelli di Machine Learning di AWS per la Classificazione e l'Estrapolazione di Informazioni contenute nelle Mail PEC

Tesi di Laurea Triennale

Relatore

Prof. Lamberto Ballan

Laureando

Riccardo Zaupa

Matricola 2034303

Anno Accademico 2023 - 2024

"I'm stronger, I'm smarter, I'm better"

— Homelander.

"Diventerò il re dei pirati"

— Monkey D. Luffy.

"Non devo fuggire"

— Shinji Ikari.

"Non dire gatto se non ce l'hai nel sacco"

— Nonna.

"GG\MM\AAAAAAA"

— Alex Scantamburlo.

"Ucciderò tutti i giganti"

— Eren Yeager.

"Con il superamento della revisione PB – a fronte di una prestazione estremamente deludente – avete concluso il vostro progetto didattico di IS."

— Tullio Vardanega.

"Non posso aiutarvi"

— Alessandro Staffolani.

"Mi avete fatto perdere mesi della mia vita"

— Riccardo Cardin.

Ringraziamenti

Desidero esprimere la mia gratitudine al professor Lamberto Ballan, mio relatore, per l'aiuto e il sostegno che mi ha dato durante la stesura dell'elaborato. Vorrei anche ringraziare, con affetto, i miei genitori per il loro sostegno, il grande aiuto e la loro presenza in ogni momento durante gli anni di studio. Desidero poi ringraziare i miei amici per i bellissimi anni trascorsi insieme e le mille avventure vissute.

Padova, Settembre 2024 Riccardo Zaupa

Sommario

Il presente documento descrive il lavoro svolto durante il periodo di stage svolta presso l'azienda Sanmarco Informatica S.p.A.. Lo stage, svolto alla conclusione del percorso di studi triennale in Informatica presso l'Università degli Studi di Padova, ha avuto una durata complessiva di trecentoore. L'obiettivo principale dello stage è stato quello di classificare e estrapolare informazioni contenute nelle mail PEC utilizzando i modelli di Machine Learning di AWS.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

In questo capitolo andremo ad enunciare la struttura del documento ed analizzeremo l'azienda ospitante il mio stage curricolare e l'offerta proposta per lo stage

1.1 Organizzazione del testo

Il secondo capitolo descrive ...

Il terzo capitolo approfondisce ...

Il quarto capitolo approfondisce ...

Il quinto capitolo approfondisce ...

Il sesto capitolo approfondisce ...

Nel settimo capitolo descrive ...

Riguardo la stesura del testo, relativamente al documento sono state adottate le seguenti convenzioni tipografiche:

- gli acronimi, le abbreviazioni e i termini ambigui o di uso non comune menzionati vengono definiti nel glossario, situato alla fine del presente documento;
- i termini in lingua straniera o facenti parti del gergo tecnico sono evidenziati con il carattere *corsivo*.

1.2 L'azienda

Sanmarco Informatica S.p.A.è un'azienda italiana di sviluppo software e consulenza informatica. Da oltre quarant'anni si dedica alla riorganizzazione dei processi aziendali in tutti i settori, progettando e implementando soluzioni digitali integrate.
L'azienda, che ad oggi conta più di 600 dipendenti e oltre 2500 aziende seguite ha come sede principale Villa Ramanelli a Grisignano di Zocco, in provincia di Vicenza, poco distante dai Centri di Ricerca e

Sviluppo (CRS) e dal Centro per la Formazione di Vicenza. Conta anche diverse filiati in Trentino-Alto Adige, Friuli-Venezia Giulia, Lombardia, Piemonte, Emilia-Romagna, Toscana, Campania e Puglia.
L'azienda è organizata in Business Unit, dei centri di competenza specifici e autonomi ma in relazione costante. Ognuna delle quali è specializzata in un settore specifico.
L'obiettivo principale è l'innovazione e il progresso tecnologico, con l'obiettivo di creare soluzioni software che siano in grado di rispondere alle esigenze dei clienti, garantendo la massima qualità e sicurezza.
La metodologia di lavoro, indipendentemente dalla Business Unit, è basata su un approccio [g]Agile implementata con il framework [g]Scrum. Agile è un approccio alla gestione dei progetti che si basa su principi di collaborazione, auto-organizzazione e flessibilità. Scrum è un framework Agile che permette di gestire progetti complessi, garantendo la massima trasparenza e la massima flessibilità.
Eventuali ulteriori informazioni sono disponibili sul sito web dell'azienda[1].

Figura 1.1: Logo di Sanmarco Informatica

1.3 L'offerta di stage

L'obiettivo dello stage consiste nella catalogazione delle Poste Elettroniche Certificate ([g]PEC), integrando tecnologie di Intelligenza Artificiale ([g]AI) per l'analisi e l'efficienza del processo.
Il modello di apprendimento automatico analizza il contenuto delle PEC e le classifica in base al contenuto.
Il progetto è stato proposto dall'azienda in occasione dell'evento Stage IT 2024, organizzato dall'Università degli Studi di Padova e promosso da Confindustria Veneto Est. Tale evento mira ad agevolare l'incontro tra studenti e aziende, offrendo la possibilità di svolgere uno stage formativo con specifico riferimento al settore ICT (Information and Communication Technology). Tale settore si riferisce all'insieme delle tecnologie utilizzate per la gestione e la comunicazione delle informazioni, incluse quelle legate all'informatica e alle telecomunicazioni.

[1] https://www.sanmarcoinformatica.com/

Capitolo 2

Descrizione dello stage

2.1 Introduzione al progetto

L'elaborazione intelligente dei documenti (IDP) è una tecnologia che automatizza il processo di immissione manuale dei dati da documenti cartacei o immagini digitali, integrandoli con altri processi aziendali digitali. Ad esempio, in un flusso di lavoro aziendale automatizzato, come l'invio di ordini ai fornitori al momento del calo delle scorte, l'IDP può sostituire l'immissione manuale dei dati da parte del team contabile. Invece di inserire manualmente i dati di una fattura ricevuta via e-mail, i sistemi di IDP estraggono automaticamente queste informazioni e le integrano direttamente nel sistema contabile, eliminando ostacoli e riducendo gli errori.

L'IDP offre numerosi vantaggi alle aziende. In termini di **scalabilità**, permette di elaborare documenti su larga scala con precisione, evitando errori umani e aumentando l'efficienza operativa. Promuove una **cultura dell'efficienza dei costi**, automatizzando attività ripetitive e riducendo i costi associati all'elaborazione manuale dei dati. Migliora anche la **soddisfazione dei clienti** grazie alla gestione più rapida e automatizzata dei documenti, come l'onboarding, le prenotazioni e i pagamenti, consentendo di fornire risposte personalizzate e veloci ai clienti.

Diversi settori traggono beneficio dall'IDP. Nel **settore sanitario**, facilita la gestione delle cartelle cliniche, migliorando l'estrazione e l'organizzazione dei dati dai documenti medici. Le **aziende finanziarie** lo utilizzano per automatizzare la gestione delle spese e l'elaborazione delle fatture, semplificando la gestione dei pagamenti. Nel **settore legale**, l'IDP analizza contratti e documenti legali, utilizzando tecnologie di elaborazione del linguaggio naturale (NLP) per estrarre informazioni chiave. Le aziende della **logistica** lo impiegano per tracciare spedizioni e documenti di transito, riducendo gli errori umani. Infine, nel settore delle **risorse umane**, l'IDP semplifica la selezione del personale, gestisce le buste paga e automatizza altre funzioni HR.

Le tecnologie alla base dell'IDP comprendono il **riconoscimento ottico dei caratteri (OCR)**, che converte immagini di testo in dati leggibili dalle macchine, e l'**elaborazione del linguaggio naturale (NLP)**, che analizza e comprende il linguaggio umano. L'**automazione robotica dei processi (RPA)** consente invece di automatizzare i flussi di lavoro aziendali ripetendo azioni umane predefinite.

Il processo di IDP si articola in diverse fasi: acquisizione e **classificazione dei documenti**, **estrazione dei dati** rilevanti tramite OCR e NLP, **convalida e successiva elaborazione dei dati** nei sistemi aziendali, e **apprendimento continuo** attraverso algoritmi di machine learning per migliorare l'accuratezza nel tempo. Inoltre, i sistemi di IDP offrono **report e analisi** per ottimizzare ulteriormente i flussi di lavoro aziendali.

Amazon Web Services (AWS) supporta l'implementazione dell'IDP attraverso servizi come **Amazon Textract**, che utilizza il machine learning per estrarre informazioni dai documenti senza interazioni manuali, e **Amazon Comprehend**, che sfrutta l'NLP per scoprire informazioni preziose nei testi. Entrambi i servizi consentono alle aziende di automatizzare la gestione dei documenti in modo efficiente e sicuro, integrandosi con altre piattaforme aziendali per un flusso di lavoro senza interruzioni.

2.2 Requisiti e obiettivi

Gli obiettivi sono stati definiti in accordo con il tutor aziendale e si identificano nel seguente modo:

[Priorità][Id]

- Priorità: indica la priorità dell'obiettivo, può essere obbligatorio o desiderabile;
- Id: composto da due cifre, identifica l'obiettivo in modo univoco rispetto alla priorità.

ID	**Categoria**	**Descrizione**
O01	Obbligatorio	Analisi dei servizi AWS per l'addestramento dei modelli AI
O02	Obbligatorio	Addestramento di un modello di apprendimento AI utilizzando i servizi AWS
O03	Obbligatorio	Analisi requisiti applicativi e tecnici per implementare la soluzione richiesta
O04	Obbligatorio	Implementare un modello di apprendimento automatico che analizzi il contenuto delle [g]PEC importate e assegni loro categorie appropriate in base al contenuto (mittente, destinatario, data e argomento)
D01	Desiderabile	Implementare algoritmi di IA in grado di adattarsi e apprendere continuamente dai dati per migliorare le prestazioni del sistema nel tempo. Ciò include l'ottimizzazione dei modelli di apprendimento automatico in base all'esperienza e ai feedback degli utenti
D02	Desiderabile	Integrazione con un sistema documentale per l'archiviazione delle PEC creando i metadati necessari con le informazioni estratte e collocandole nella corretta categoria di appartenenza

2.3 Pianificazione

2.3.1 Pianificazione settimanale

Il periodo di stage è stato suddiviso in 8 settimane, durante le quali sono previste le seguenti attività:

Settimana	**Dal**	**Al**	**Attività**
1	24-06-2024	28-06-2024	- Incontro con persone coinvolte nel progetto per discutere i requisiti e le richieste di implementazione - Ricerca, studio e documentazione per inquadramento progetto - Introduzione ai linguaggi di sviluppo - Introduzione agli ambienti di sviluppo - Introduzione dei servizi [g]awsg
2	01-07-2024	05-07-2024	- Analisi dei servizi AWS per l'addestramento di un modello di apprendimento - Addestramento di un modello di apprendimento utilizzando i servizi di AWS **Milestone:** Utilizzo dei servizi AWS per l'addestramento di un modello di apprendimento
3	08-07-2024	12-07-2024	- Studio della soluzione per definire i requisiti necessari per l'implementazione **Milestone:** Analisi dei requisiti applicativi e tecnici per implementare la soluzione
4	15-07-2024	19-07-2024	- Addestramento modello di apprendimento per catalogare le PEC in base al loro contenuto
5	22-07-2024	26-07-2024	- Implementazioni per interfacciarsi con il modello di apprendimento addestrato e per poter catalogare le PEC importate **Milestone:** Completamento obiettivi minimi
6	29-07-2024	02-08-2024	- Implementazione algoritmo di AI per l'autoapprendimento
7	05-08-2024	09-08-2024	- Studio e documentazione sulle API messe a disposizione dal documentale per poter catalogare le mail PEC - Implementazione dell'integrazione con il documentale producendo i metadati necessari per catalogare le PEC

Settimana	**Dal**	**Al**	**Attività**
8	12-08-2024	16-08-2024	- Verifica e test archiviazione PEC nel documentale **Milestone:** Completamento obiettivi massimi
9	19-08-2024	23-08-2024	- Recupero eventuali ritardi
10	26-08-2024	30-08-2024	- Recupero eventuali ritardi

Capitolo 3

Tecnologie

In questo capitolo verranno descritti i servizi e le tecnologie analizzate e pertinenti per il problema descritto, in quale modo possono essere impiegate e una panoramica finalizzata a chiarirne il constesto e il caso d'uso.

3.1 Amazon Web Services

Amazon Web Services (AWS) è una piattaforma di servizi cloud che offre potenza di calcolo, storage di database, distribuzione di contenuti e altre funzionalità per aiutare le aziende a scalare e crescere. AWS offre una vasta gamma di servizi che possono essere utilizzati per implementare soluzioni di Intelligenza Artificiale (AI) e Machine Learning ([g]ML). Per la realizzazione dell'applicazione sono stati individuati diversi servizi che hanno permesso di realizzare un'architettura scalabile e serverless.

3.1.1 Amazon Comprehend

Amazon Comprehend è un servizio di analisi del linguaggio naturale ([g]Natural Language Processing) che utilizza l'apprendimento automatico per identificare informazioni utili nei testi. Amazon Comprehend è stato utilizzato per analizzare i testi delle recensioni degli utenti dell'applicazione.

Figura 3.1: Logo di Amazon Comprehend

3.1.2 Amazon Textract

Amazon Textract è un servizio di [g]Optical Character Recognition che utilizza l'apprendimento automatico per riconoscere e analizzare il testo e i dati delle immagini. Amazon Textract è stato utilizzato

per estrarre il testo dalle immagini delle ricette.

Figura 3.2: Logo di Amazon Textract

3.1.3 Amazon S3

Amazon Simple Storage Service (Amazon S3) è un servizio di storage di oggetti che offre scalabilità, disponibilità dei dati, sicurezza e prestazioni. Amazon S3 è progettato per memorizzare grandi quantità di dati a un costo molto basso. Amazon S3 è stato utilizzato per memorizzare i file inerenti alle diversi fasi del progetto.

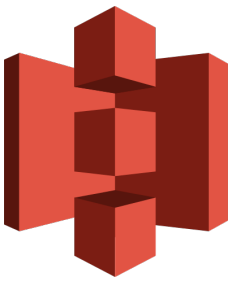


Figura 3.3: Logo di Amazon S3

3.1.4 AWS Lambda

AWS Lambda è un servizio di calcolo serverless che esegue il codice in risposta a eventi e gestisce automaticamente le risorse di calcolo richieste dal codice. AWS Lambda è stato utilizzato per implementare le funzioni di backend dell'applicazione.

Figura 3.4: Logo di AWS Lambda

3.1.5 Amazon DynamoDB

Amazon DynamoDB è un servizio di database NoSQL completamente gestito che offre prestazioni di singolo millisecondo a qualsiasi scala. Amazon DynamoDB è stato utilizzato per memorizzare i dati relativi ai vari utenti dell'applicazione.

Figura 3.5: Logo di Amazon DynamoDB

3.1.6 Amazon Step Functions

AWS Step Functions è un servizio di orchestrazione di serverless che consente di coordinare facilmente i componenti di applicazioni distribuite e microservizi utilizzando logica visuale.

Figura 3.6: Logo di Amazon Step Functions

3.2 Strumenti di sviluppo

3.2.1 Visual Studio Code

Visual Studio Code è un editor di codice sorgente sviluppato da Microsoft per Windows, Linux e macOS. Visual Studio Code è stato utilizzato per scrivere il codice dell'applicazione.

3.2.2 Git

[g]Git è un sistema di controllo di versione distribuito utilizzato per tenere traccia delle modifiche al codice sorgente durante lo sviluppo del software. Git è stato utilizzato per tenere traccia delle modifiche

Figura 3.7: Logo di Visual Studio Code

al codice sorgente dell'applicazione.

Figura 3.8: Logo di Git

3.2.3 Bitbucket

Bitbucket è un servizio di hosting di [g]Repository Git basato su cloud. Bitbucket è stato utilizzato per memorizzare il codice sorgente dell'applicazione.

Figura 3.9: Logo di Bitbucket

3.3 Linguaggi di programmazione

3.3.1 Python

Python è un linguaggio di programmazione ad alto livello, interpretato, adatto per lo sviluppo di applicazioni web, desktop e mobile. Python è stato utilizzato per la realizzazione del backend dell'applicazione.

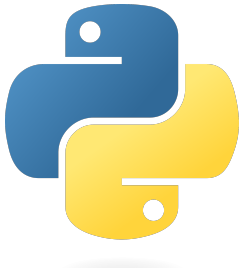


Figura 3.10: Logo di Python

Capitolo 4

Progettazione e codifica

Breve introduzione al capitolo

4.1 Tecnologie e strumenti

Di seguito viene data una panoramica delle tecnologie e strumenti utilizzati.

Tecnologia 1

Descrizione Tecnologia 1.

Tecnologia 2

Descrizione Tecnologia 2

4.2 Ciclo di vita del software

4.3 Progettazione

4.3.1 Namespace 1

Descrizione namespace 1.

Classe 1: Descrizione classe 1

Classe 2: Descrizione classe 2

4.4 Design Pattern utilizzati

4.5 Codifica

Capitolo 5

Conclusioni

Lorem [g]SDK

Lorem Application Program Interface

5.1 Consuntivo finale

Ipsum

5.2 Raggiungimento degli obiettivi

Sit amet

5.3 Conoscenze acquisite

5.4 Valutazione personale

Acronimi e abbreviazioni

AI Artificial Intelligence. 1, 15

API Application Programming Interface. 1, 15

AWS Amazon Web Services. 1, 15

ML Machine Learning. 1, 15

NLP Natural Language Processing. 1, 15

OCR Optical Character Recognition. 1, 15

PEC Posta Elettronica Certificata. 1, 15

SDK Software Development Kit. 1, 15

UML Unified Modeling Language. 1, 16

Glossario

. 1

Agile Nell'ambito dell'ingegneria del software con il termine Agile si intende 1, 2, 15

AI Per aritificial Intelligence (AI) si intende 1, 2, 7, 14

API In informatica con il termine *API* si indica ogni insieme di procedure disponibili al programmatore, di solito raggruppate a formare un set di strumenti specifici per l'espletamento di un determinato compito all'interno di un certo programma. La finalità è ottenere un'astrazione, di solito tra l'hardware e il programmatore o tra software a basso e quello ad alto livello semplificando così il lavoro di programmazione. 1, 13, 14

AWS Amazon Web Services (AWS) è una piattaforma di servizi cloud che offre potenza di calcolo, storage di database, distribuzione di contenuti e altre funzionalità per aiutare le imprese a scalare e crescere. 1, 14

Git Git è un sistema di controllo di versione distribuito gratuito e open source progettato per gestire tutto, dai piccoli ai grandi progetti, con velocità ed efficienza. 1, 9, 15

ML Per Machine Learning (ML) si intende 1, 7, 14

NLP Natural Language Processing (NLP) è ... 1, 7, 14

OCR Optical Character Recognition (OCR) è 1, 7, 14

PEC La *Posta Elettronica Certificata* (PEC) è un servizio di posta elettronica che garantisce l'invio e la ricezione di messaggi di posta elettronica con valore legale equivalente a quello della raccomandata con avviso di ricevimento.. 1, 2, 14

Repository Con il termine repository si intende .. . 1, 10, 15

Scrum In ingegneria del software, per Scrum si intende 1, 2, 15

SDK A software development kit (SDK) is a collection of software development tools in one installable package. They facilitate the creation of applications by having a compiler, debugger and sometimes a software framework. They are normally specific to a hardware platform and operating system combination. To create applications with advanced functionalities such as advertisements, push notifications, etc; most application software developers use specific software development kits. 1, 13, 14

UML **text** In ingegneria del software *Unified Modeling Language* (ing. linguaggio di modellazione unificato) è un linguaggio di modellazione e specifica basato sul paradigma object-oriented. L'*UML* svolge un'importantissima funzione di "lingua franca" nella comunità della progettazione e programmazione a oggetti. Gran parte della letteratura di settore usa tale linguaggio per descrivere soluzioni analitiche e progettuali in modo sintetico e comprensibile a un vasto pubblico. 1, 14

Bibliografia

Books

James P. Womack, Daniel T. Jones. *Lean Thinking, Second Editon*. Simon & Schuster, Inc., 2010.

Articles

Einstein, Albert, Boris Podolsky e Nathan Rosen. «Can Quantum-Mechanical Description of Physical Reality be Considered Complete?» In: *Physical Review* 47.10 (1935), pp. 777–780. DOI: 10.1103/PhysRev.47.777.

Siti web consultati

AWS. URL: https://aws.amazon.com/.

Manifesto Agile. URL: http://agilemanifesto.org/iso/it/.