Università degli Studi di Milano Bicocca

Dipartimento di Informatica

Corso di Laurea Magistrale

Progetto di ML - Analisi della Qualità dell'Aria nell'est Asia

ANNO ACCADEMICO 2024/2025

Isceri Alessandro - 879309 &

Farioli Alessio - 879217 &

Ghilotti Riccardo - 879259

Indice

1. Introduzione

1.1 Descrizione del Dominio

Il dominio applicativo di questo studio si concentra sull'analisi dell'inquinamento atmosferico nelle nazioni del sud-est asiatico. I dati utilizzati comprendono una vasta gamma di misurazioni relative alla qualità dell'aria, integrate da informazioni demografiche che possono influenzare significativamente i livelli di inquinamento nelle diverse aree analizzate.

Questo insieme di variabili ambientali e sociali permette di esplorare non solo l'entità dell'inquinamento, ma anche i fattori che lo determinano, offrendo una visione più completa del fenomeno.

1.2 Obiettivi dell'Elaborato

Un primo obiettivo di questo elaborato è analizzare i dati relativi alla qualità dell'aria e ai fattori demografici, al fine di identificare relazioni significative tra le variabili e la qualità dell'aria.

Un ulteriore scopo consiste nell'indurre dei modelli di Machine Learning in grado di prevedere la qualità dell'aria in diverse aree geografiche in base alle misurazioni effettuate. Questi modelli consentiranno di valutare la qualità dell'aria in una qualsiasi zona solamente attraverso le misurazioni dei parametri specificati, in questo modo è possibile valutare in breve tempo l'abitabilità di una zona per quanto riguarda l'inquinamento atmosferico.

2. Analisi ed esplorazione del dataset

2.1 Descrizione del Dataset

Il dataset analizzato riguarda i livelli di inquinamento atmosferico nelle regioni del sud-est asiatico, come descritto nel capitolo introduttivo. Esso comprende un totale di 5000 righe e 10 colonne, ciascuna rappresentante una variabile significativa per l'analisi.

La variabile target, **Air Quality**, è di tipo categorico e descrive la qualità dell'aria attraverso quattro classi:

1. **Good**: Aria pulita con livelli di inquinamento bassi, che non presentano rischi per la salute.
2. **Moderate**: Qualità dell'aria accettabile, ma con concentrazioni di alcuni agenti inquinanti che potrebbero influire leggermente sui gruppi più sensibili della popolazione.
3. **Poor**: Livelli di inquinamento che potrebbero causare problemi di salute per individui sensibili, come bambini, anziani e persone con malattie respiratorie o cardiache.
4. **Hazardous**: Livelli estremamente alti di inquinamento che rappresentano un serio rischio per la salute pubblica.

Le colonne del dataset contengono dati relativi a misurazioni ambientali e fattori demografici che potrebbero influenzare la qualità dell'aria. Di seguito, vengono riportate tali feature:

1. **Temperature (°C)**: La temperatura media registrata nella regione, espressa in gradi Celsius.
2. **Humidity**: La percentuale di umidità relativa nell'aria, un fattore che può influenzare la dispersione degli agenti inquinanti.
3. **PM2.5 Concentration (µg/m³)**: La concentrazione di particolato fine, noto per i suoi effetti nocivi sulla salute.

4. **PM10 Concentration (µg/m³)**: La concentrazione di particolato più grossolano, che contribuisce al deterioramento della qualità dell'aria.

5. **NO_2 Concentration (ppb)**: Il livello di biossido di azoto, un gas prodotto principalmente da attività industriali e traffico veicolare.

6. **SO_2 Concentration (ppb)**: La concentrazione di biossido di zolfo, spesso legata alla combustione di combustibili fossili.

7. **CO Concentration (ppm)**: Il livello di monossido di carbonio, un agente inquinante generato da processi di combustione incompleta.

8. **Proximity to Industrial Areas (km)**: La distanza dalla zona industriale più vicina, un parametro che può influire significativamente sui livelli di inquinamento locale.

9. **Population Density (people/km²)**: La densità di popolazione, espressa in persone per chilometro quadrato.

I dati utilizzati per questa analisi sono stati raccolti da diverse organizzazioni internazionali che si occupano del monitoraggio della qualità dell'aria e di altri fattori ambientali:

- **World Health Organization (WHO)**: Organizzazione Mondiale della Sanità, che fornisce risorse e dati dettagliati sull'inquinamento atmosferico e i suoi effetti sulla salute (`https://www.who.int/health-topics/air-pollution`).

- **World Bank**: mette a disposizione dati globali relativi a fattori demografici e ambientali, tra cui la densità di popolazione (`https://data.worldbank.org/indicator/EN.POP.DNST`).

2.2 Analisi Esplorativa

In questa sezione viene presentata un'analisi esplorativa del dataset, con l'obiettivo di comprendere meglio l'influenza delle variabili in esame sulla qualità dell'aria. Di seguito è riportata una matrice che mostra alcune statistiche descrittive per ciascuna feature del dataset, ovvero la media, la deviazione standard, il valore minimo, il massimo e i quartili.

	Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	PIA	PD
mean	30.03	70.06	20.14	30.22	26.41	10.01	1.5	8.43	497.42
std	6.72	15.86	24.55	27.35	8.9	6.75	0.55	3.61	152.75
min	13.4	36	0	-0.2	7.4	-6.2	0.65	2.5	188
25%	25.1	58.3	4.6	12.3	20.1	5.1	1.03	5.4	381
50%	29	69.8	12	21.7	25.3	8	1.41	7.9	494
75%	34	80.3	26.1	38.1	31.9	13.73	1.84	11.1	600
max	58.6	128.1	295	315.8	64.9	44.9	3.72	25.8	957

In questa matrice le feature **Proximity_to_Industrial_Areas** e **Population_Density** sono chiamate rispettivamente PIA e PD.

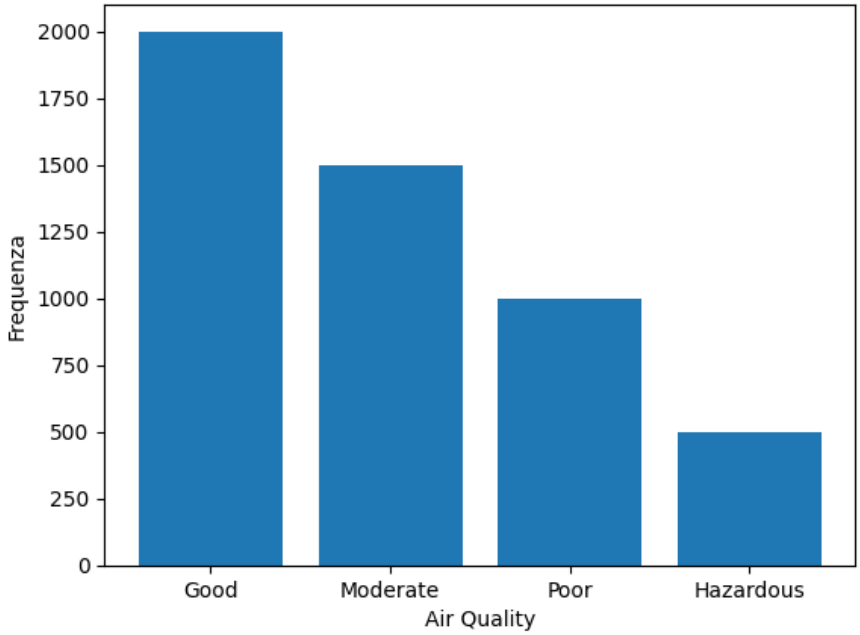

Figura 2.1: Distribuzione dei valori della colonna target.

Dal grafico sopra riportato emerge che il dataset presenta uno **sbilanciamento** tra le classi, poiché non tutte sono rappresentate con la stessa frequenza. In particolare, la qualità dell'aria nelle regioni analizzate tende prevalentemente ad essere classificata come **Good** o **Moderate**, mentre le classi **Poor** e **Hazardous** risultano meno comuni.

Questo squilibrio potrebbe influire sulle prestazioni dei modelli di classificazione, che potrebbero essere portati a privilegiare le classi più rappresentate (**Moderate** e **Good**) a scapito di quelle meno frequenti (**Poor** e **Hazardous**). Sarà quindi importante valutare i risultati dei modelli tenendo a mente che è presente uno sblianciamento.

L'approccio adottato per esplorare quanto una feature influenza le classi target consiste nel suddividere i suoi valori in quattro categorie, al fine di analizzare la distribuzione delle classi target per ciascuna di esse. Tuttavia, è importante sottolineare che questa analisi non garantisce una relazione diretta tra la feature considerata e la qualità dell'aria.

Ad esempio, analizzando un grafico a barre che rappresenta la distribuzione di **Air Quality** rispetto a gruppi di temperatura (**Temperature Group**), si osserva che la qualità dell'aria tende a essere migliore nelle regioni con temperature più basse. Questo potrebbe suggerire che una temperatura bassa favorisca una migliore qualità dell'aria, anche se ulteriori approfondimenti sarebbero necessari per confermare questa ipotesi.

La suddivisione dei valori della feature in categorie è stata effettuata utilizzando i percentili, con i seguenti intervalli:

- **Low**: (0, 25° percentile)
- **Med-low**: (25° percentile, 50° percentile)
- **Med-high**: (50° percentile, 75° percentile)
- **High**: (75° percentile, ∞)

Questa categorizzazione consente di osservare in modo più chiaro eventuali correlazioni tra i valori della feature analizzata e la distribuzione delle classi di qualità dell'aria.

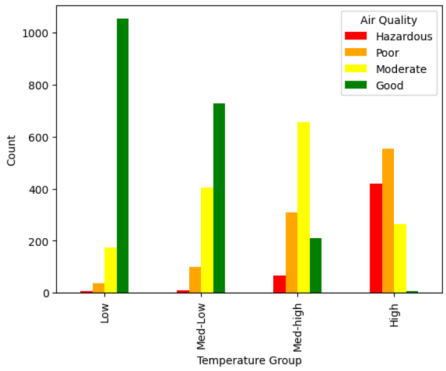

Figura 2.2: Distribuzione label in base alla feature temperatura.

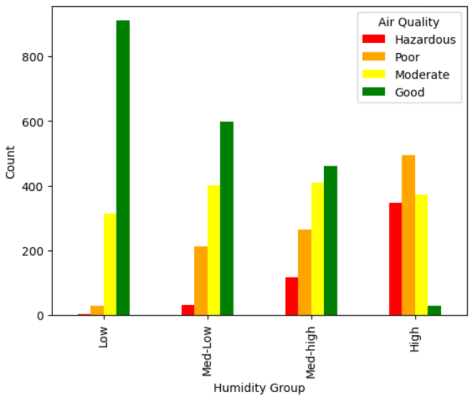

Figura 2.3: Distribuzione label in base alla feature umidità.

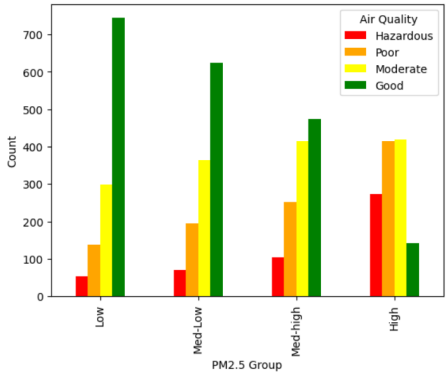

Figura 2.4: Distribuzione label in base alla feature PM2.5.

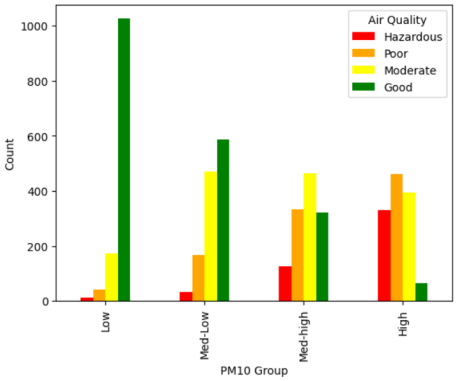

Figura 2.5: Distribuzione label in base alla feature PM10.

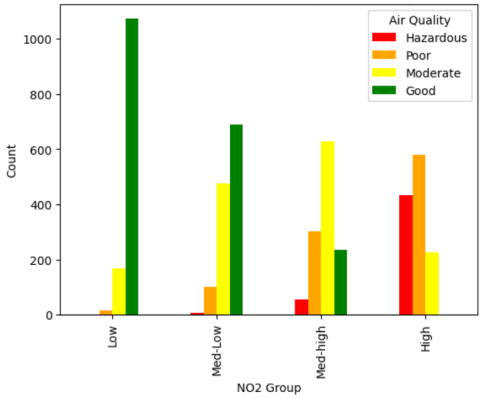

Figura 2.6: Distribuzione label in base alla feature NO_2.

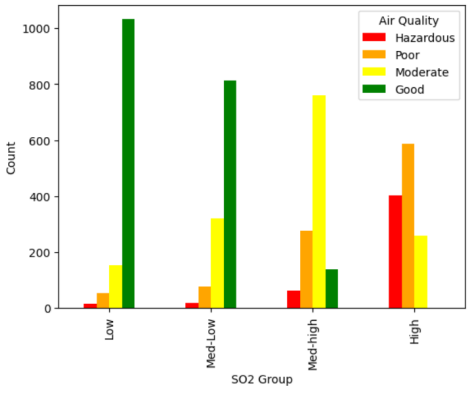

Figura 2.7: Distribuzione label in base alla feature SO_2.

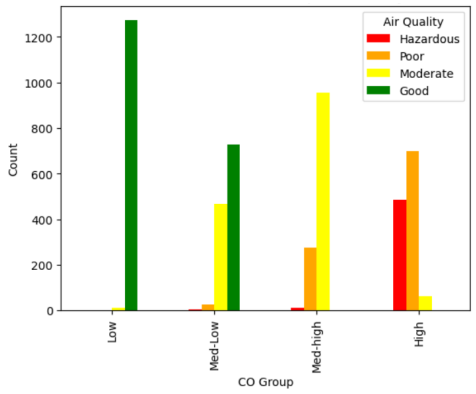

Figura 2.8: Distribuzione label in base alla feature CO.

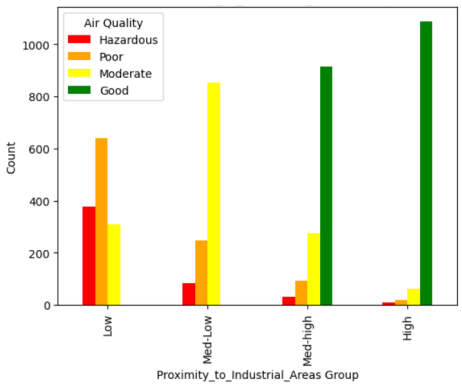

Figura 2.9: Distribuzione label in base alla feature Proximity_to_Industrial_Areas.

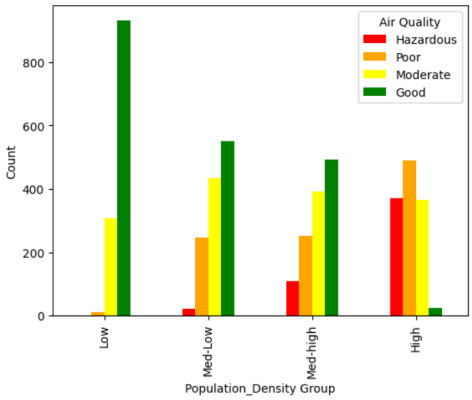

Figura 2.10: Distribuzione label in base alla feature Population_Density.

Poiché le classi **Good** e **Moderate** sono le più rappresentate nel dataset, è naturale che queste classi siano prevalenti nella maggior parte dei grafici analizzati. Tuttavia, è comunque possibile fare alcune osservazioni significative sulla distribuzione delle classi target in relazione alle feature.

Dai grafici, emerge che le feature che sembrano influenzare maggiormente la qualità dell'aria (in termini di distribuzione delle classi) sono:

- **Proximity to Industrial Areas**
- **Temperature**

- **Humidity**
- **CO Concentration**

In particolare, la temperatura, l'umidità e la concentrazione di CO risultano avere una relazione inversa con la qualità dell'aria: all'aumentare di questi valori, la qualità dell'aria tende a peggiorare. Al contrario, la prossimità a aree industriali sembra essere direttamente proporzionale alla qualità dell'aria, con la qualità che peggiora all'avvicinarsi a queste aree.

Per queste feature in particolare, si può notare che c'è una netta differenza tra la frequenza delle classi nelle categorie low/med-low e med-high/high.

Al contrario, la feature PM2.5 sembra meno influente. Infatti, la distribuzione delle classi tra le diverse categorie di PM2.5 non mostra una varianza significativa, in particolare nelle categorie med-high e high, indicando che questa variabile sembra non avere un impatto così forte sulla qualità dell'aria nelle regioni analizzate.

2.2.1 PCA

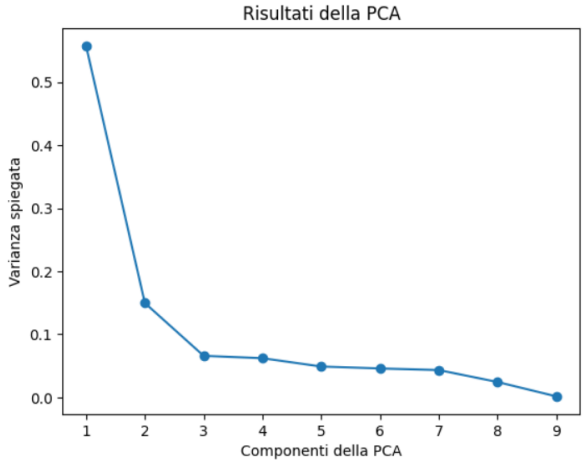

Figura 2.11: Varianza spiegata dalle componenti PCA.

Da questo grafico si può notare che è possibile giustificare una buona varianza del dataset utilizzando un numero ridotto di componenti. In particolare, con sole 2 componenti si riesce a spiegare circa

il 70% della varianza complessiva. Con 3 componenti, invece, la varianza spiegata sale a quasi il 77%.

Di conseguenza è possibile ridurre significativamente il numero di feature, e quindi la complessità dei modelli, utilizzando solo un piccolo numero di componenti senza compromettere le prestazioni. Questo risultato è probabilmente dovuto al fatto che, come osservato in precedenza, diverse feature sono già in grado di spiegare efficacemente la variabilità della classe target.

Aumentando il numero di componenti oltre le prime tre, il guadagno in termini di informazione non è significativo. Pertanto, è stato deciso di utilizzare le prime 3 componenti per semplificare alcuni modelli.

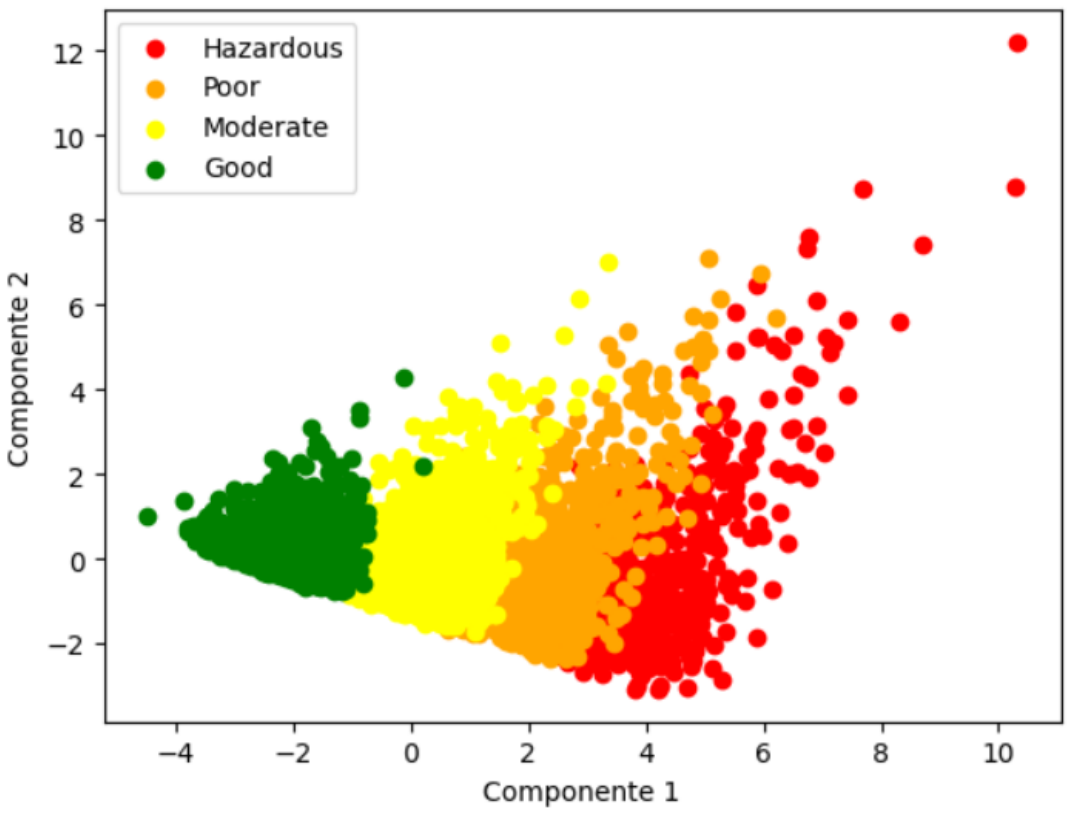

Figura 2.12: Distribuzione dei vettori a due dimensioni calcolati tramite PCA nel piano.

Questo grafico conferma che, utilizzando solo due componenti, è possibile ottenere una buona separazione delle classi nel piano. Le classi sono chiaramente distinte, ad eccezione di alcune

sovrapposizioni tra classi simili, come ad esempio quelle tra **Hazardous** e **Poor**.

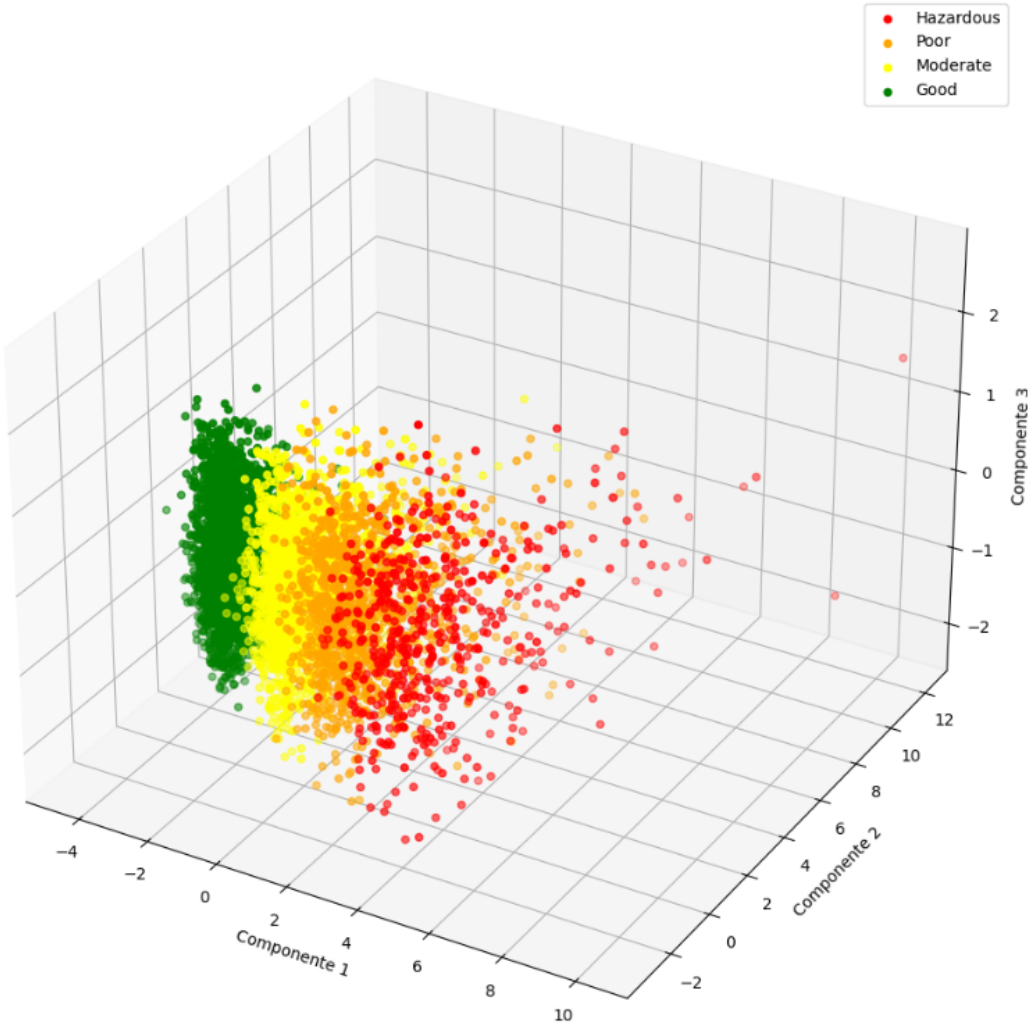

Figura 2.13: Distribuzione dei vettori a tre dimensioni calcolati tramite PCA a nello spazio.

Inoltre, per ottenere una visione più completa della distribuzione delle istanze, è stato tracciato un grafico che mostra la loro distribuzione nello spazio, utilizzando le prime 3 componenti.

Anche questo grafico conferma una buona separazione dei dati, il che fa supporre che i modelli di machine learning addestrati su questi dati siano in grado di raggiungere ottime performance, nonostante il ridotto numero di feature.

3. Modelli di Clustering

3.1 K-Means

Il **clustering** è stato utilizzato anche per proseguire l'analisi esplorativa, fornendo una misura tangibile della separazione dei dati nello spazio. Una delle ragioni di questa scelta è che il clustering non richiede una conoscenza approfondita a priori dei dati, rendendolo un metodo utile per esplorare la distribuzione dei dati.

Per l'analisi è stato impiegato il modello **K-Means** con un numero di cluster k pari a 4, sia sui dati originali che sulle istanze indotte mediante PCA.

3.1.1 PCA

Dopo aver analizzato i grafici 2.12 e 2.13, ci si aspetta che il clustering non fornisca risultati eccezionali, ma nemmeno scarsi. Questo perché le classi sono parzialmente sovrapposte, in particolare le classi **Poor** e **Hazardous**, e, non avendo a disposizione le etichette durante la fase di addestramento, è difficile ricostruire accuratamente le classi originali. Infatti, come si nota dai grafici, punti molto vicini tra loro potrebbero appartenere a classi differenti, complicando la distinzione tra le categorie, soprattutto dato è stato utilizzato l'algoritmo K-Means.

4. Modelli di Classificazione

4.1 SVM

È stato scelto il modello **SVM** (Support Vector Machine) perché si pensava che potesse ottenere buone performance sui dati trasformati tramite PCA, come suggerito dai risultati grafici e dal clustering. L'obiettivo era anche quello di confrontare le performance del modello SVM ottenuto utilizzando le istanze fornite dalla PCA con il modello allenato sui dati originali, che non erano necessariamente linearmente separabili, eventualmente utilizzando **metodi kernel** più complessi. Tuttavia, non è stato necessario ricorrere a kernel complessi, poiché il modello ha ottenuto risultati migliori utilizzando il kernel lineare.

Per allenare il modello, l'**iperparametro C**, che indica il fattore di tolleranza dei vettori nel margine, è stato impostato inizialmente a 1. Questo perchè un valore di C basso favorisce la ricerca di un margine ampio, riducendo il costo dei vettori presenti nel margine. Poiché, tracciando i vettori ottenuti tramite PCA, si notava già una buona separazione delle classi, si è deciso di incrementare il valore di C a 5, riducendo così la tolleranza per i vettori presenti nel margine.

Dato che si tratta di un problema multiclasse, gli iperpiani separatori sono stati calcolati seguendo lo schema **one-vs-one**.

4.2 Reti Neurali

Poiché il dataset possiede un gran numero di feature e sono presenti più di due classi target, sono state scelte le **reti neurali** per la loro potenza nel trattare dati complessi e multi-classe.

Inoltre, l'obiettivo era anche quello di confrontare due modelli simili, in quanto entrambi cercano un iperpiano separatore, anche se in modo diverso.

La rete neurale allenata sui dati originali presenta 5 layers **fully connected**. Nel primo layer (strato di input) sono presenti 9 neuroni (uno per ogni feature). Nel secondo e terzo strato sono presenti 18 neuroni, in questo modo la rete può apprendere funzioni di separazione più complesse in minor tempo. Il penultimo strato presenta 8 neuroni, inseriti per evitare una compressione troppo improvvisa dello spazio (da 18 a 4), mentre l'ultimo strato ha 4 neuroni (uno per ciascuna classe target).

Il modello di reti neurali è stato indotto sia sui dati originali che sui dati trasformati tramite l'operazione PCA.

Per i vettori ottenuti dalla PCA, è stata scelta una rete **fully connected** a 4 strati, con un numero diverso di neuroni per ciascun layer. In particolare, il primo strato utilizza 3 neuroni (uno per ogni componente), il secondo 9, il terzo 6 e il quarto 4. Il numero di neuroni è stato scelto in maniera analoga alla rete precedente.

Come **funzioni di attivazione**, sono state utilizzate **leaky ReLU** in tutti i layer, eccetto l'ultimo, dove è stata utilizzata **softmax** per ottenere la probabilità che un'istanza appartenga a una delle 4 classi.

Per entrambi i modelli è stato utilizzato come ottimizzatore **Adam**, come loss function la **categorical crossentropy**, poiché si tratta di un problema multiclasse, e come metrica da misurare durante il training l'**accuracy**.

Inoltre, per entrambi i modelli è stata utilizzata la tecnica di **early stopping** per fermare la fase di addestramento nel caso in cui la loss non dovesse diminuire oltre una soglia predefinita per tre iterazioni consecutive, evitando l'apprendimento si bloccasse in **minimi locali**.

5. Risultati Ottenuti

5.1 Clustering

5.1.1 Matrici di Confusione

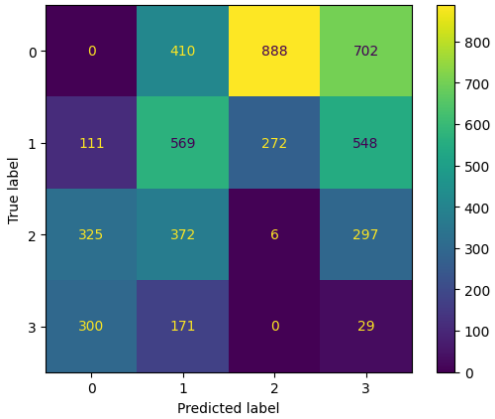

Figura 5.1: Matrice di Confusione K-Means sulle istanze originali.

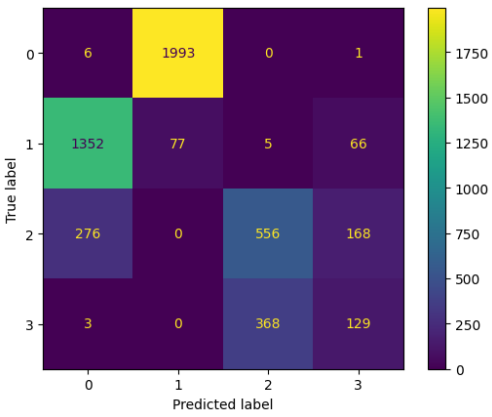

Figura 5.2: Matrice di Confusione K-Means sulle istanze PCA.

5.1.2 Metriche di Valutazione

Silhouette

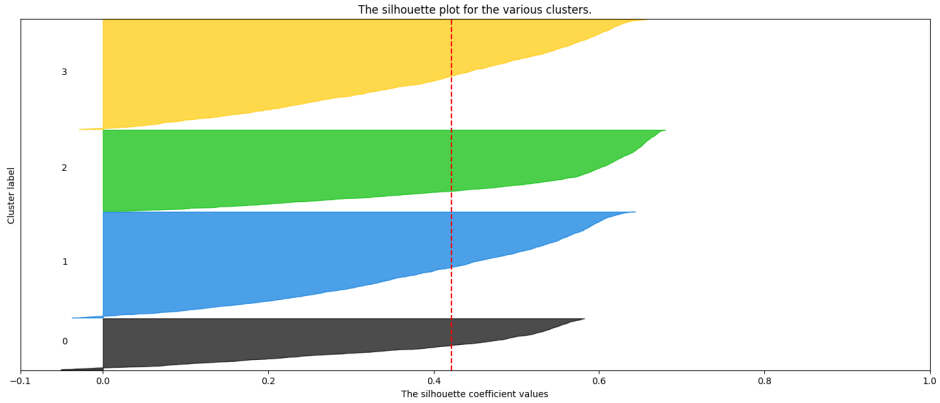

Figura 5.3: Silhouette del clustering sulle istanze originali.

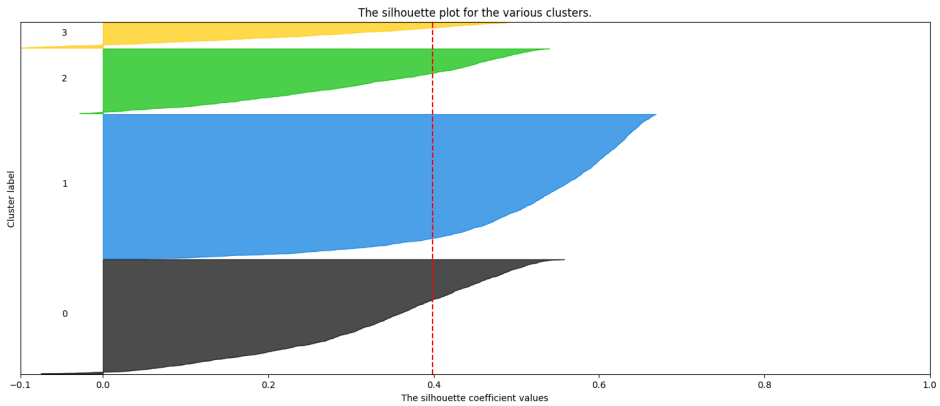

Figura 5.4: Silhouette del clustering sulle istanze PCA.

Indice di Rand aggiustato

L'**indice di rand** è una misura utilizzata in statistica in particolare nel clustering. L'indice di Rand è una misura della similarità tra due risultati di clustering, nel nostro caso il primo "clustering" corrispondeva alle labels originali, mentre il secondo al risultato di k-means. Inoltre, questa metrica è utile quando si vuole misurare la similarità di due modi di assegnare delle label a dei dati, ignorando la permutazione delle label.

	K-means dati originali	**K-means dati PCA**
Indice di Rand	0.12	0.75
Avg Silhouette	0.42	0.40

5.2 Metodi di Validazione

Come metodo di validazione, è stata effettuata una divisione tra **train** e **test**, in particolare, il test set è composto dal 20% delle istanze e il train set dall'80%.

Ovviamente, il training set è stato utilizzato per allenare i 4 modelli di classificazione, mentre il test set è stato utilizzato per calcolare le **matrici di confusione**, le **metriche** e le **curve ROC**.

5.3 SVM

5.3.1 Matrici di Confusione

Figura 5.5: Matrice di Confusione sulle istanze originali.

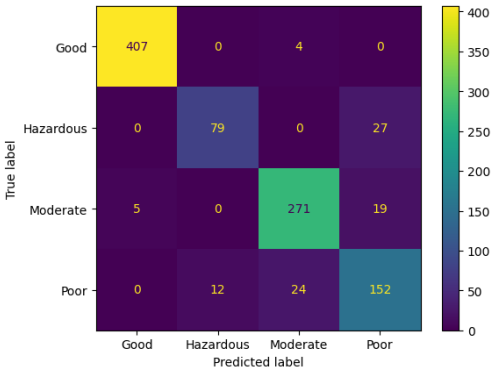

Figura 5.6: Matrice di Confusione sulle istanze PCA.

5.3.2 Metriche di Valutazione

SVM: Numero di support vectors: 628

	Precision	**Recall**	**F1-score**	**Support**
Good	1.00	1.00	1.00	411
Moderate	0.95	0.97	0.96	295
Poor	0.85	0.86	0.85	188
Hazardous	0.87	0.79	0.83	106
Accuracy			0.94	1000
Macro avg	0.92	0.90	0.91	1000
Weighted avg	0.94	0.94	0.94	1000

Tabella 5.2: Metriche di SVM sulle istanze originali.

SVM-PCA: Numero di support vectors: 854

	Precision	Recall	F1-Score	Support
Good	0.99	0.99	0.99	411
Moderate	0.91	0.92	0.91	295
Poor	0.77	0.81	0.79	188
Hazardous	0.87	0.75	0.80	106
Accuracy			0.91	1000
Macro Avg	0.88	0.87	0.87	1000
Weighted Avg	0.91	0.91	0.91	1000

Tabella 5.3: Metriche di SVM sulle istanze PCA.

5.3.3 Curve ROC

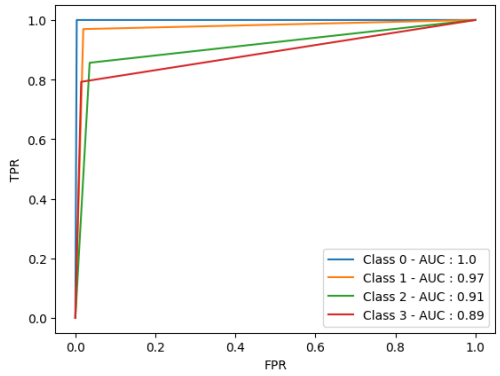

Figura 5.7: Curva ROC di SVM applicato sulle istanze originali.

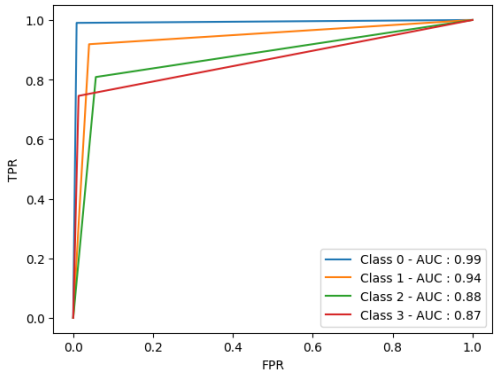

Figura 5.8: Curva ROC di SVM applicato sulle istanze PCA.

5.4 Reti Neurali

5.4.1 Matrici di confusione

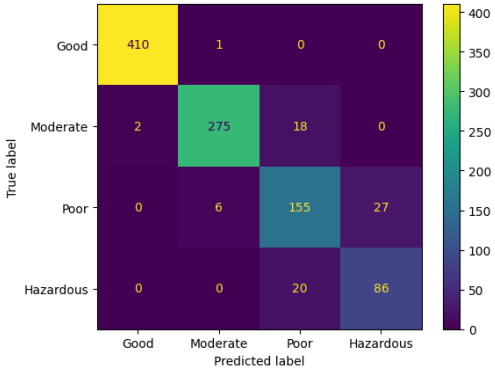

Figura 5.9: Matrice di Confusione sulle istanze originali.

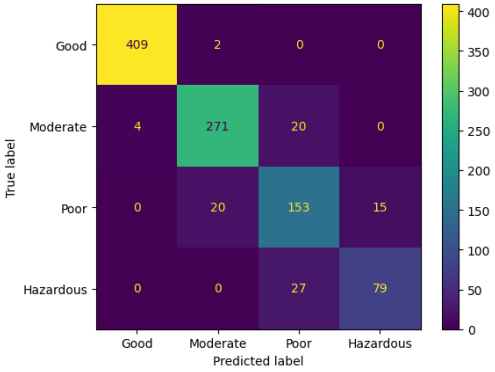

Figura 5.10: Matrice di Confusione sulle istanze PCA.

5.4.2 Metriche di Valutazione

	Precision	**Recall**	**F1-Score**	**Support**
Good	0.99	1.00	0.99	411
Moderate	0.92	0.97	0.95	295
Poor	0.86	0.77	0.81	188
Hazardous	0.80	0.78	0.79	106
Accuracy			0.93	1000
Macro Avg	0.89	0.88	0.89	1000
Weighted Avg	0.92	0.93	0.92	1000

Tabella 5.4: Metriche di reti neurali sulle istanze originali.

	Precision	Recall	F1-Score	Support
Good	0.99	0.99	0.99	411
Moderate	0.92	0.92	0.92	295
Poor	0.77	0.82	0.80	188
Hazardous	0.85	0.76	0.81	106
Accuracy			0.91	1000
Macro Avg	0.88	0.87	0.88	1000
Weighted Avg	0.91	0.91	0.91	1000

Tabella 5.5: Metriche di reti neurali sulle istanze PCA.

5.4.3 Curve ROC

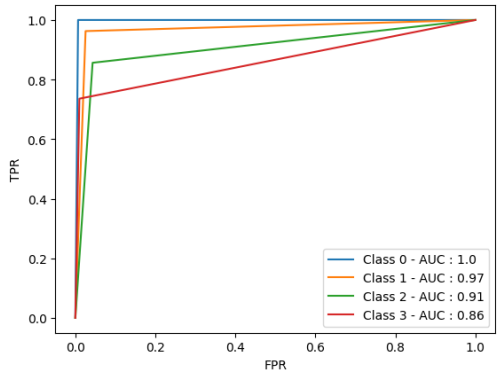

Figura 5.11: Curva ROC di reti neurali applicate sulle istanze originali.

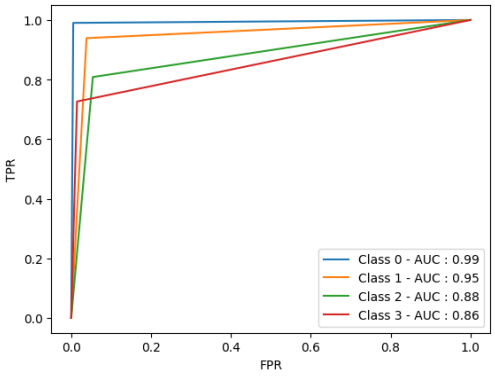

Figura 5.12: Curva ROC di reti neurali applicate sulle istanze PCA.

5.4.4 Accuracy e Loss dei modelli

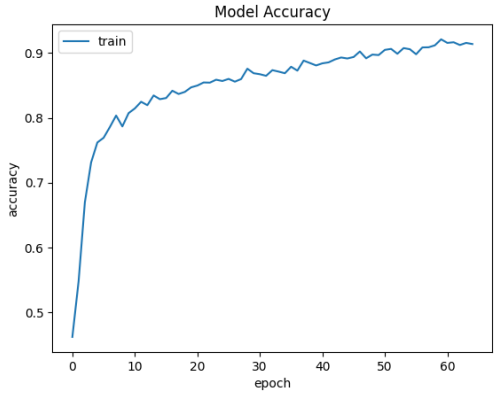

Figura 5.13: Accuracy di reti neurali applicate sulle istanze originali.

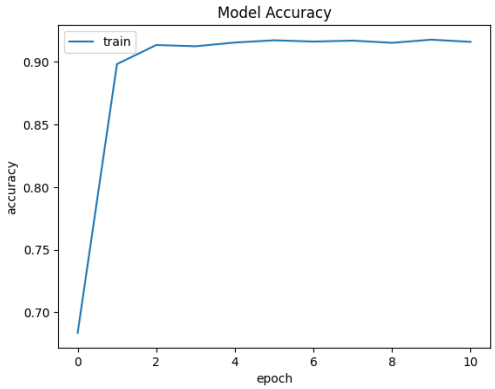

Figura 5.14: Accuracy di reti neurali applicate sulle istanze PCA.

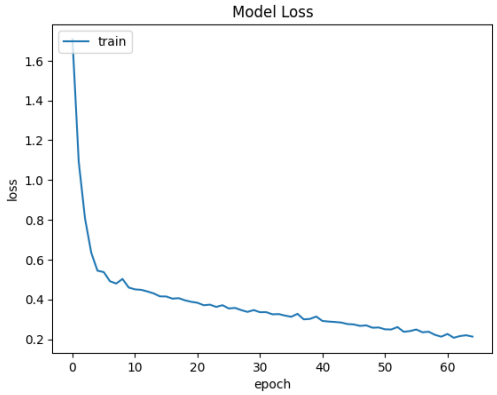

Figura 5.15: Loss di reti neurali applicate sulle istanze originali.

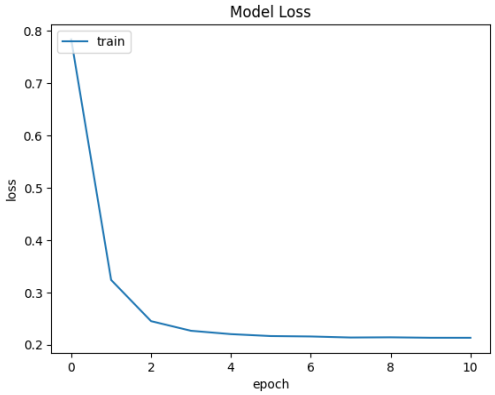

Figura 5.16: Loss di reti neurali applicate sulle istanze PCA.

6. Analisi dei Risultati

6.1 Clustering

6.1.1 Matrici di confusione

Le **matrici di confusione** 5.1.1 mostrano che il **clustering** sui dati grezzi non è ottimale. In particolare, i dati che appartengono a classi diverse tendono ad essere posizionati vicino tra di loro nello spazio originale, causando una bassa separabilità tra i cluster.

Al contrario, quando i dati vengono trasformati tramite **PCA**, la matrice di confusione migliora significativamente. In questo caso, le prime due classi vengono correttamente separate, mentre nelle ultime due classi si osserva una maggiore sovrapposizione. Questo fenomeno è confermato dal grafico in 2.13, che evidenzia una forte sovrapposizione tra le classi **Poor** e **Hazardous**.

6.1.2 Silhouette

A differenza delle matrici di confusione, il grafico delle **silhouette** 5.1.2 mostra performance leggermente migliori quando i dati non sono stati trattati tramite **PCA**. Questo è confermato dal grafico 2.13, dove si osserva che i dati risultano essere molto vicini tra loro. Di conseguenza, la **distanza inter-cluster** deve essere bassa, il che porta a una valutazione peggiore delle silhouette.

6.1.3 Indice di Rand

L'indice di Rand 5.1.2 conferma quanto osservato nelle matrici di confusione. Sui dati originali, l'indice è molto basso in linea con la matrice di confusione 5.1, che risulta essere molto confusa. Al contrario, sui dati trasformati tramite PCA, l'indice di Rand è significativamente più alto, ed infatti la matrice di confusione 5.2 appare più ordinata e le classi sono separabili più facilmente.

6.2 SVM

Le **Support Vector Machine** ottengono risultati eccellenti, con metriche molto alte e una matrice di confusione quasi perfetta 5.3.1.

Come previsto dalla figura 2.1, le classi con il minor numero di vettori presentano statistiche inferiori rispetto alle altre classi.

Le metriche 5.3.2 mostrano anche che **SVM** performa meglio sui dati originali senza l'applicazione di **PCA**, sebbene questo comporti un costo maggiore in termini di tempo di addestramento.

Anche le **curve ROC** 5.3.3 confermano che il modello SVM ottiene prestazioni leggermente superiori sui dati originali rispetto a quando viene allenato sui dati generati grazie alla PCA.

6.3 Reti Neurali

Dai risultati degli esperimenti con le **reti neurali**, si osserva un leggero miglioramento nel caso del dataset originale, in termini di **matrici di confusione** 5.4.1, **metriche** 5.4.2 e grafici delle **curve ROC** 5.4.3, rispetto al modello allenato sui dati trasformati tramite PCA. Tuttavia, a differenza dei modelli SVM, in questo caso il tempo complessivamente risparmiato è notevole. Infatti, sui dati originali l'addestramento richiede circa una sessantina di epoche, mentre con la PCA ne bastano solo una decina. Inoltre, il tempo di training delle reti neurali sui dati originali è circa tre volte maggiore rispetto a quello necessario per le istanze trasformate tramite PCA.

In aggiunta, si può notare che i modelli convergono rapidamente. Come mostrato dai grafici 5.13, 5.14, 5.15 e 5.16, i valori di **loss** e **accuracy** si stabilizzano già dopo poche epoche di addestramento.

7. Conclusioni

Il dataset analizzato presenta sfide significative per i modelli non supervisionati, come **K-Means**, che non riescono a separare efficacemente le classi. Questo è particolarmente evidente quando si applica la **PCA**, in quanto dai grafici prodotti si nota che le istanze sono vicine nello spazio e le classi non sono ben separate. Tuttavia, modelli supervisionati come **SVM** e **reti neurali** ottengono ottimi risultati, separando le classi in modo più preciso grazie alla loro capacità di gestire meglio la complessità dei dati e di apprendere funzioni anche non lineari.

L'applicazione della **PCA** ha rivelato che le classi nel dataset, purtroppo, non sono completamente separabili anche dopo la riduzione dimensionale. Tuttavia, i modelli supervisionati riescono comunque a ottenere buone performance, sfruttando le informazioni contenute nelle prime tre componenti della **PCA**, che descrivono la maggior parte della variabilità nei dati. Questo approccio ha permesso di ridurre significativamente la complessità del modello, mantenendo prestazioni elevate, e di ridurre i tempi di addestramento rispetto all'utilizzo del dataset completo.

Inoltre, l'analisi delle **curve ROC**, delle **matrici di confusione** e delle **metriche di valutazione** ha mostrato che i modelli supervisionati hanno appreso correttamente la funzione di separazione, anche grazie all'uso della **PCA** che ha semplificato la struttura dei dati. **SVM** ha fornito i risultati più robusti, con una separazione netta delle classi e una bassa confusione tra le categorie più simili.

In generale, i risultati suggeriscono che l'uso di modelli supervisionati come **SVM** e reti neurali, combinato con tecniche di riduzione dimensionale come la **PCA**, può portare a modelli efficienti e rapidi, che performano molto bene anche in presenza di un dataset complesso e parzialmente sovrapposto.