CORSO DI STUDIO *Physics (LM-17)*

ANNO ACCADEMICO *2023-2024*

DENOMINAZIONE DELL'INSEGNAMENTO *Scientific Data Analysis Laboratory*

Principali informazioni sull'insegnamento	
Anno di corso	2°
Periodo di erogazione	1° Semestre: Settembre - Dicembre 2023
Crediti formativi universitari (CFU/ECTS):	6
SSD	FIS/01 – Fisica Sperimentale
Lingua di erogazione	Inglese
Modalità di frequenza	Obbligatoria

Docente	
Nome e cognome	Alexis Pompili
Indirizzo mail	alexis.pompili@ba.infn.it , alexis.pompili@uniba.it
Telefono	+39 080 5442436
Sede	Dipartimento Interateneo di Fisica (Ufficio R15) - Via G.Amendola 173, 70125 Bari
Sede virtuale	Zoom Room personale
Ricevimento	Ore di ricevimento: - Giovedì: 11-13 e 15-17 (Ufficio R15 oppure online/su zoom) - per appuntamento (Ufficio R15 oppure online/su zoom)

Organizzazione della didattica			
Ore			
Totali	Didattica frontale	Pratica (laboratorio, campo, esercitazione, altro)	Studio individuale
150	16	60	74
CFU/ECTS			
6	2	4	

Obiettivi formativi	Ci si aspetta che gli studenti raggiungano, alla fine del corso, una buona conoscenza di come applicare - in modo critico - concetti statistici avanzati a problemi presi dalla vita reale della ricerca scientifica, utilizzando metodi e tecniche statistiche appropriate utilizzate nel campo della Fisica Subnucleare e Nucleare. L'implementazione computazionale di questi metodi e tecniche deve essere effettuata all'interno dei *framework* consolidati della Fisica della Alte Energie (HEP), utilizzando dell'opportuno codice realizzato in C^{++} e/o in Python.
Prerequisiti	- corso di *Statistical Data Analysis* (1^{o} anno della Magistrale in Fisica) - corsi della Triennale in Fisica dove si studiano o sperimentano Linguaggi di programmazione (per esempio il corso di Informatica).

Metodi didattici	Facendo leva sui concetti teorici di statistica appresi nel corso di Analisi Statistica dei Dati, gli studenti sono guidati, attraverso un approccio completamente pratico, a gestire metodi e tecniche di analisi dei dati, da quelli di base a quelli avanzati, con applicazioni ed esempi tratti dal campo della Fisica delle Alte Energie. Gli studenti lavorano, per tutta la durata del corso, con il proprio *account* personale su una macchina virtuale Linux dedicata, ospitata dal Data Center

	ReCas-Bari, opportunamente configurata ed accessibile anche da casa per consentire un esercizio individuale esteso. Gli esempi/esercizi in classe sono completamente supportati da *file* realizzati in codice C/C^{++} e/o Python ed integrati da presentazioni elettroniche basate su slide in ppt/pdf. Vengono forniti tutti i file necessari; spesso vengono proposti dei compiti aggiuntivi come ulteriore sviluppo dell'esercizio iniziato nella lezione di laboratorio. Come supporto aggiuntivo, viene fornita una pagina web con *link* utili per organizzare il materiale fornito e anche documentazione aggiuntiva selezionata, manuali online, ecc. Il foglio di lavoro supplementare può essere svolto individualmente, ma si incoraggia il lavoro in piccoli gruppi.

Risultati di apprendimento previsti ***Da indicare per ciascun Descrittore di Dublino (DD)*** ***DD1* Conoscenza e capacità di comprensione** ***DD2* Conoscenza e capacità di comprensione applicate** ***DD3-5* Competenze trasversali**	**- DD1: conoscenza e capacità di comprensione** o Comprensione del metodo scientifico, della natura e delle modalità della ricerca in Fisica o Conoscenza degli strumenti informatici avanzati di uso corrente nei settori della ricerca di base ed applicata o Conoscenza delle tecniche di calcolo avanzate o conoscenza necessaria per configurare un'analisi di dati (reali o simulati) e comprendere i risultati ottenuti relativi all'estrazione o alla caratterizzazione di un segnale fisico o di una tendenza fisica, attraverso un corretto trattamento statistico; o conoscenza di diversi metodi e tecniche statistiche per il trattamento dei dati e comprensione del contesto in cui possono essere correttamente applicati e delle approssimazioni e incertezze coinvolte. **- DD2: capacità di applicare conoscenza e comprensione** o Capacità di identificare gli elementi essenziali di un fenomeno o Capacità di utilizzare lo strumento dell'analogia per applicare soluzioni conosciute a problemi nuovi (problem solving) o Capacità di utilizzo di strumenti di calcolo matematico analitico e numerico o Capacità di utilizzo delle tecnologie elettroniche e informatiche e la loro applicazione all'acquisizione dei dati sperimentali o conoscenza e capacità di impostare un'attività di analisi dei dati statisticamente corretta e computazionalmente affidabile, implement-tando un codice (realizzato in C^{++} e/o Python), opportuno allo scopo prefissato, mediante l'uso di strumenti/framework HEP consolidati; o capacità di costruire e testare modelli e ipotesi. **- DD3-5: competenze trasversali** ***-DD3** :* ● **Autonomia di giudizio** o Capacità di lavorare con crescenti gradi di autonomia, anche assumendo responsabilità nella programmazione di progetti e nella gestione di strutture o Consapevolezza dei problemi di sicurezza nell'attività di laboratorio o sviluppare la capacità di impostare ed eseguire autonomamente e correttamente un compito di analisi di dati e di valutare i risultati ottenuti, comprendendo anche come possono essere eventualmente migliorati.

	***-DD4**:* ● **Abilità comunicative** o Competenze nella comunicazione in lingua italiana e in lingua inglese nei settori avanzati della Fisica o sviluppare competenze informatiche relative all'elaborazione/analisi dei dati sperimentali e alle modalità di supporto alla loro esposizione/presentazione, utilizzando una terminologia specifica e tecnica; o sviluppare competenze comunicative e di presentazione - in lingua inglese - basate sulla terminologia specifica utilizzata nel campo dell'analisi dei dati sperimentali; inoltre, diffusione delle conoscenze con un linguaggio scientifico appropriato; o capacità generale di lavorare in gruppo e di inserirsi rapidamente ed efficacemente in un ambiente di lavoro condiviso. - ***DD5**:* ● **Capacità di apprendere in modo autonomo** o Acquisizione di strumenti conoscitivi di base per l'aggiornamento continuo delle conoscenze o capacità di apprendere e trasferire nuove tecniche e metodi di analisi dei dati sperimentali; o capacità di gestire e configurare un'analisi dei dati – una volta fornita una serie di dati di natura anche molto varia - costruendo un modello e verificandone/testandone la validità, allo scopo di estrarre una firma/caratteristica sottostante di cui valutare la significatività in termini statistici.
Contenuti di insegnamento (Programma)	I contenuti e la struttura del corso possono essere riassunti come segue. Il corso si propone di fornire, a diversi livelli di complessità dell'approccio, le conoscenze e le relative competenze necessarie per configurare, impostare e gestire, in modo appropriato dal punto di vista statistico e computazionale, l'analisi di una grande varietà di dati, reali o simulati secondo un determinato modello, e la valutazione dei risultati di questa analisi, compresa la consapevolezza delle approssimazioni implicite e la comprensione delle incertezze e delle correlazioni statistiche e sistematiche coinvolte. Viene fornita una panoramica degli strumenti software adatti e dei *framework* comunemente utilizzati nel settore HEP, insieme a un uso esteso di questi strumenti in un'ampia serie di esempi di applicazione ed esercitazioni. Il corso è strutturato in 3 moduli (A, B, C) di complessità crescente. Le ore suddivise tra questi 3 moduli e i loro titoli di riferimento sono riportati di seguito: A: 14 (2 lez. + 12 lab.) B: 46 (10 lez. + 36 lab.) C: 16 (4 lez. + 12 lab.) L'uso di strumenti software per l'analisi dei dati viene introdotto gradualmente attraverso la sequenza dei moduli: ROOT in (A), pacchetto RooFit in (B), Jupyter, "ecosistema" Pandas e RDataFrame in (C). Il programma dettagliato del corso è riportato di seguito.

	Mod. A Rassegna dei principali comandi utili di Unix/Linux. Rassegna dei principali comandi utili degli editor vi ed emacs. Introduzione al software di analisi dei dati ROOT. Gestione e rappresentazione di funzioni matematiche. Semplici applicazioni di rappresentazione dei dati di un foglio elettronico con ROOT (TGraphErrors). Semplice generazione casuale di variabili a partire da una distribuzione di campionamento, memorizzazione in strutture (istogrammi, alberi), gestione e visualizzazione dei dati generati. Gestione degli istogrammi (operazioni con gli istogrammi), confronto tra due o più istogrammi (normalizzazione relativa). Esempio di confronto dati-*Monte Carlo* (MC): tra una distribuzione di dati e un insieme di diverse componenti simulate (da comporre in un istogramma *stacked*); normalizzazione assoluta. Rapporto dati-MC; r*ebinning* in presenza di fluttuazioni rilevanti dovute a statistiche basse. Istogramma di un'efficienza ed errore binomiale. Esempio di un semplice classificatore binario per estrarre un segnale da un insieme di fondi di varia natura: variabili di selezione e valutazione del loro potere di reiezione del fondo attraverso il confronto delle loro curve *ROC*. Scelta del punto di lavoro. Mod. B Introduzione al pacchetto RooFit di ROOT. Metodo di interpolazione (*fit*) basato sul principio della massima verosimiglianza (*maximum likelihood*); implementazione di una funzione di densità di probabilità complessa e componenti di un modello di fit teorico. Massima verosimiglianza estesa. Fit *binned* e *unbinned* con esempi. Introduzione al supporto della modellazione all'interno di RooFit; librerie, area di lavoro e *factory* in RooFit. Modellazione di base: polinomi standard, polinomi di Chebyshev e Bernstein, funzione Argus, funzioni polinomiali con soglie. Funzioni sigmoidali; esempio di adattamento di una curva di efficienza con una *error function*. Bontà dell'interpolazione: chi-quadrato normalizzato, probabilità di adattamento e *pull bin-by-bin*. Le *pull* come strumento per controllare la presenza di eventuali *bias* e la sotto/sovra-stima delle incertezze. Esempi di *fit* a distribuzioni di massa invarianti; caso di una particella/risonanza con una larghezza intrinseca molto più piccola della risoluzione di massa sperimentale (*fit* a una o più Gaussiane) e caso di una particella/risonanza con una larghezza intrinseca dello stesso ordine di grandezza della risoluzione sperimentale (funzione Voigtiana, utilizzando una funzione Breit-Wigner non relativistica). Interpolazione con una convoluzione numerica esplicita di una funzione di risoluzione Breit-Wigner relativistica e di una Gaussiana. Funzioni empiriche: funzione Crystal Ball (singola e doppia) in caso di code radiative, funzione di Johnson (come migliore alternativa a più di due gaussiane), funzione di Landau. Esponenziale convoluta con una funzione di risoluzione temporale (modello RooDecay); inclusione di una incertezza temporale evento per evento.

	Matrice di covarianza/correlazione e stima delle incertezze dei parametri con i metodi Hesse (approssimazione parabolica: errori simmetrici) e Minos (errori asimmetrici). Scansione della *likelihood* e *profile likelihood* con esempi; verifica dell'equivalenza tra il metodo della *profile likelihood* e Minos. Calcolo della risoluzione effettiva in caso di più Gaussiane nel *fit*, con stima della sua incertezza mediante propagazione dell'errore tenendo in debito conto le correlazioni tra i parametri del *fit* stesso (funzione getPropagatedError). Interpolazioni simultanee (multidimensionali) in RooFit. Esempio: interpolazione 2D delle distribuzioni di massa invariante e tempo proprio (*fit* massa-vita di una particella elementare). Tecnica di interpolazione non parametrica della distribuzione di una (o più) variabili casuali. Esempi di applicazione della stima della *kernel density* (KDE) in una o due dimensioni (RooKeysPdf e Roo2DKeysPdf). RooFit come strumento di generazione di distribuzioni in base a modelli teorici scelti (*toy MC* / pseudo-esperimenti). Esempi di generazione di una distribuzione e della sua successiva inteprolazione; regolazione del seme del generatore casuale. *Closure test* di un *fitting task*. Valutazione dei tempi computazionali di un'interpolazione. Test di Kolmogorov-Smirnov. Significatività statistica di un segnale mediante un rapporto di verosimiglianza (*likelihood ratio*) nel contesto dell'applicazione del teorema di Wilks: esempio di stima approssimata. Breve accenno alla significatività statistica di un segnale nel contesto frequentista (mediante l'utilizzo di pseudo-esperimenti) ripreso nel successivo modulo C. Rassegna dei metodi di sottrazione del fondo: tecnica delle bande laterali, metodo *bin-wise* e sPlot (utilizzando il pacchetto ROOT/TMVA), con applicazioni pratiche. Reiezione del fondo ed estrazione del segnale. Esempio/esercizio di analisi multivariata con l'utilizzo di un albero decisionale *boostato* (BDT). Test di ipotesi e presenza di un segnale oltre agli eventi di fondo. Calcolo del *p-value*, e quindi della significatività statistica, per l'osservazione di un segnale; definizione di due modelli (modello nullo per l'ipotesi di solo fondo e modello alternativo per l'ipotesi di segnale+fondo) e, forniti i dati osservati, due approcci di calcolo: 1) *frequentista*, con una statistica di prova di tipo *one-side Profile likelihood*, attraverso la generazione di pseudo-esperimenti (da eseguire in ore fuori classe, essendo computazionalmente intensivo), e 2) *asintotico* (AsymptoticCalculator) con la stessa statistica di test.

	Un esercizio aggiuntivo/facoltativo prevede il calcolo del *p-value* in funzione di una caratteristica del segnale (ad esempio la massa del segnale), utilizzando il primo approccio. Mod. C Introduzione a **Python**: caratteristiche di base, controllo di flusso, funzioni. Librerie per il calcolo scientifico: **numPy**, **sciPy, matplotlib, uproot (**"ecosistema" **Pandas**). Introduzione al *framework* **Jupyter** ed esempi of moderna analisi dei dati con estrazione di un segnale nel campo della Fisica delle Alte Energie (estrazione di un signale associato al decadimento di una particella per mezzo di una selezione classica basata su tagli e configurata ed implementata mediante un **Jupyter** *notebook*).
Testi di riferimento	Il materiale fornito dal docente copre l'intero corso; consiste in un insieme di diversi file pdf inseriti in una pagina web dedicata al corso. Questa pagina web contiene anche link utili su documentazione online e tutorial aggiuntivi selezionati dal docente e liberamente disponibili su Internet. Se necessario per l'argomento, questi file contengono una breve e specifica introduzione/richiamo teorico. Come ulteriore riferimento lo studente può prendere in considerazione il materiale didattico fornito e i libri di riferimento suggeriti nel corso di Analisi statistica dei dati, in particolare i noti libri di testo di G. Cowan (*) e L. Lista (**). Il riferimento suggerito per approfondire i concetti di base del Machine Learning, alla fine del modulo C, è il libro di testo di I. Goodfellow et al. (***). Gli esempi e gli esercizi svolti nelle ore di laboratorio vengono eseguiti su una macchina virtuale ReCas dedicata al corso, dove gli studenti ricevono un *account* personale. I file da utilizzare per gli esempi o per ispirare (ulteriori) esercizi sono illustrati nel materiale didattico fornito. Vengono utilizzati gli *open data* dell'esperimento CMS. (*) G. Cowan, Statistical Data Analysis, 1998, Clarendon Press; disponibile in biblioteca e anche in formato cartaceo su richiesta del docente. (**) L. Lista, Statistical methods for Data Analysis in Particle Physics, 2020 (2a o 3a edizione), Springer Verlag, disponibile in biblioteca e anche in formato cartaceo su richiesta del docente. (***) I. Goodfellow et al., Deep Learning, 2016, MIT Press.
Note ai testi di riferimento	I libri proposti sono molto più ricchi del contenuto concettuale del corso. Possono essere utilizzati come guida di riferimento per i concetti di base, i metodi e le tecniche utilizzate nel corso.
Materiali didattici	https://web2.ba.infn.it/~pompili/teaching.html

Valutazione	

Modalità di verifica dell'apprendimento	- Elaborazione degli esercizi di laboratorio assegnati durante il corso (20%) - Esame di Laboratorio (80%)
Criteri di valutazione	Ci si aspetta che lo studente abbia appreso: - la conoscenza concettuale e pratica/tecnica dei metodi e delle tecniche di analisi dei dati e la consapevolezza del corretto trattamento statistico e degli aspetti computazionali coinvolti nell'analisi; - la capacità di progettare, configurare e impostare un compito di analisi dei dati; - la conoscenza e la comprensione dell'interpretazione statisticamente appropriata dei risultati dell'analisi, comprese le incertezze sistematiche, le eventuali approssimazioni e i possibili miglioramenti.
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	Criteri di valutazione adottati: - Conoscenza e comprensione: 30% - Applicazione delle conoscenze e della comprensione: 45% - Autonomia di giudizio: 10% - Abilità nel comunicare conoscenza e comprensione: 10% - Capacità di continuare ad apprendere: 5%
Altro	
	I tre moduli sono incentrati sulle applicazioni di laboratorio, con esempi ed esercizi, dei concetti statistici introdotti e approfonditi nel corso di Analisi Statistica dei Dati. Le ore in classe saranno dedicate a richiamare/rinnovare queste conoscenze, ma con riferimento concreto alle funzioni, ai metodi e agli algoritmi coinvolti e disponibili negli strumenti e *framework* software da utilizzare. La fine del terzo modulo prevede ore in classe per introdurre una panoramica coerente e mirata/guidata dei concetti e dei metodi di apprendimento automatico (leggermente menzionati nel corso di Analisi statistica dei dati) nel contesto specifico dell'HEP. Si presume implicitamente che un'introduzione teorica più ampia e una più vasta gamma di applicazioni in contesti diversi dall'HEP siano fornite in un corso dedicato di Machine Learning tra quelli lasciati a libera scelta degli studenti.