

Planificación de la Asignatura: Tópicos Especiales en Áreas Complementarias: Herramientas Informáticas para Análisis de Datos

Fecha: 23/10/2024 13:02

Código: OP003-5

Carrera: Licenciatura en Bioinformática

Departamento Académico: Informática

Docente a cargo:

Correo del docente a cargo: gmerino@ingenieria.uner.edu.ar

Régimen de Dictado: Cuatrimestral 1º Cuatrimestre

Carga Horaria Semanal: 5 horas semanales

Carga Horaria Total: 70 horas

Contenidos Mínimos:

Entorno R. Tipos de datos y estructuras propias de R. Uso de funciones y paquetes estadísticos R. Uso de herramientas informáticas para generar gráficas y analíticas de estadística descriptiva. Ajuste e interpretación de modelos lineales generales y mixtos. Herramientas para el análisis de datos multivariados y la aplicación de técnicas no paramétricas.

Competencias Genéricas:

Competencias Específicas:

Argumentación de aportes marcados en la matriz de competencias:

Correlativas Regulares para cursar:

Tercer año

Correlativas Aprobadas para cursar:

Segundo año

Correlativas Aprobadas para promocionar o rendir el examen final:

Segundo año

Insercion de la Asignatura en el plan de Estudios:

Según la Res. 847/05 del Ministerio de Educación, Ciencia y Tecnología de la Nación, los alcances del título de Licenciado/a en Bioinformática implican un número de actividades directamente vinculadas con el análisis informático de datos biológicos. En este sentido, la mayoría de los datos a analizar se obtienen de una población bajo estudio, como los estudios epidemiológicos o de expresión génica. Es por ello que contar con formación sólida en análisis estadístico se vuelve una necesidad para los futuros profesionales.

Adicionalmente, el software estadístico R es uno de los más utilizados en el análisis bioinformático, por lo que se considera fundamental que los alumnos adquieran habilidades en su uso y manipulación.

Los contenidos principales de la materia están vinculados al uso de herramientas estadísticas en el análisis de grandes bases de datos. En el contexto del plan de estudios, los contenidos aportados por todas las asignaturas resultan fundamentales para el éxito del aprendizaje de esta materia. Específicamente, los principales aportes según las asignaturas son:

"Probabilidad y Estadística" y "Métodos Estadísticos": Introducen al alumno las nociones básicas para comprender los procesos, modelos aleatorios y los métodos estadísticos que serán utilizados para analizar los datos biológicos.

"Informática Básica", "Fundamentos de Programación", "Programación Avanzada", "Algoritmos y Estructura de datos", "Computación de Alto Rendimiento", "Bases de Datos", "Ingeniería de Software I y II": Forman al alumno en el área de informática, proveyendo la base necesaria para el uso de software de análisis así como también para la implementación computacional de modelos o técnicas estadísticas de análisis aplicadas sobre bases de datos.

"Cálculo en una Variable", "Álgebra Lineal y Geometría Analítica", "Cálculo Vectorial" y "Ecuaciones Diferenciales", "Matemática Discreta": Proveen las herramientas formales para la comprensión de las funciones y ecuaciones matemáticas asociadas a las distribuciones estadísticas así como también a los métodos que se utilizan en los métodos y algoritmos estadísticos.

"Bioquímica", "Biología Celular y Molecular" y "Química orgánica y Biológica", "Genética", "Estructuras Biomoleculares", "Análisis y Alineamiento de Secuencias": Aportan el conocimiento teórico necesario para comprender los fenómenos que subyacen los datos biológicos a analizar.

“Metodología de la Investigación Científica” y “Epistemología”: Proporcionan herramientas del método científico y la investigación, útiles para la ejecución de una tarea compleja y estructurada como lo es el análisis de grandes volúmenes de datos (big data) cuyo objetivo es la extracción de información biológica útil.

“Inteligencia Artificial”, “Señales y Sistemas”, “Modelos de Sistemas Biológicos”, “Modelos y Simulación de Macromoléculas”: Proveen herramientas clásicas de análisis matemático y estadístico y útiles para la modelización de distintos fenómenos biológicos.

"Laboratorio de Inglés I a IV": Son fundamentales ya que gran parte de la bibliografía, más aún la de las temáticas de vanguardia, se encuentra disponible sólo en inglés.

Por todos los motivos anteriormente expuestos, este curso del ciclo superior se considera de vital importancia en la formación del alumno y futuro profesional de la Bioinformática.

Objetivo General:

Que el alumno logre:

1. Utilizar el programa estadístico R como herramienta para el análisis de datos representativos de fenómenos biológicos.
2. Aplicar los conceptos y las estrategias básicas para el análisis estadístico de datos biológicos mediante herramientas informáticas.
3. Utilizar los conceptos, métodos y estrategias de análisis a problemas reales del ámbito biomédico, biotecnológico y multi-ómico.
4. Interpretar los resultados de la aplicación de modelos estadísticos en el contexto de problemas reales con el objetivo de elegir la mejor solución.
5. Fortalecer el uso del marco conceptual-teórico en la capacidad del alumno para plantear modelos o herramientas estadísticas según el problema bajo estudio.
6. Reforzar la capacidad para adquirir conocimientos en forma autónoma así como también para trabajar en forma grupal.

Objetivos Particulares:

Que el alumno logre:

1. Adquirir habilidades en el uso del software estadístico R.
2. Conocer y aprovechar todas las ventajas que ofrece R y la gran diversidad de paquetes libremente disponibles.
3. Utilizar las herramientas previamente estudiadas de análisis estadístico en R.
4. Aprender métodos estadísticos avanzados para el análisis de datos bioinformáticos.
5. Analizar datos biológicos de diversa índole.
6. Adquirir habilidades críticas para la evaluación de los modelos y los resultados estadísticos.
7. Interpretar correctamente los resultados obtenidos por distintas estrategias y en el contexto del problema bajo estudio.
8. Diseñar sus propias estrategias de análisis estadístico para grandes bases de datos biológicos.
9. Utilizar con conocimiento y juicio crítico las herramientas computacionales disponibles.
10. Implementar sus propias funciones en R.
11. Reforzar y adquirir lenguaje estadístico útil para la descripción de las tareas de análisis y la comunicación de los resultados obtenidos.

Programa Analítico:

- Unidad Temática 1: Introducción a R. Introducción a R y RStudio. Instalación de paquetes específicos. Tipos de datos y estructuras propias de R. Manipulación de datos y uso de funciones. Espacio de trabajo en RStudio y uso del ayuda de R.
- Unidad Temática 2: Herramientas del análisis exploratorio. Estadística descriptiva en R. Obtención e implementación de gráficos exploratorios y confirmatorios en R. La librería ggplot2.
- Unidad Temática 3: Uso de modelos lineales. Definición de los modelos lineales en R. Regresión lineal y Análisis de la Varianza con el paquete stats. Implementación del método de mínimos cuadrados. Análisis e interpretación de los coeficientes. Gráficas básicas de diagnóstico en R.
- Unidad Temática 4: Modelos lineales avanzados. Teoría de los modelos lineales avanzados. Ajuste de modelos lineales avanzados. Técnicas de diagnóstico en R.
- Unidad Temática 5: Análisis Multivariado en R. Discriminante Lineal. Análisis de componentes principales. Análisis en conglomerados. Árboles de decisión.
- Unidad Temática 6: Estadística No Paramétrica en R. Estimación de la función de distribución acumulada. Análisis de correlación e independencia en R. Uso e implementación de pruebas no paramétricas.

Metodología Didáctica:

La asignatura está orientada al estudio y aplicación de las herramientas informáticas para el análisis estadístico de grandes bases de datos biológicos. Dada la diversidad de las áreas de estudio en las que éstas pueden ser utilizadas, el principal aprendizaje que los alumnos adquirirán será la experiencia al trabajar con diferentes ejemplos y creando sus propias estrategias de análisis. Los conjuntos de datos bioinformáticos suelen ser voluminosos y de compleja estructura, por lo que es fundamental que los alumnos adquieran además habilidades informáticas de manera que puedan aplicar los conocimientos adquiridos. Es por ello que esta asignatura propone una metodología basada en clases teórico-prácticas.

Las clases consisten en una instancia inicial, en donde se presentarán los contenidos teóricos mínimos de la unidad correspondiente, una instancia media, en donde se analizarán ejemplos básicos de análisis de datos, y una instancia final en donde se resolverán problemas de aplicación. Toda la actividad es coordinada por el docente a cargo.

Para cada unidad, se proveerá a los alumnos con material teórico-práctico digital, diseñado exclusivamente para este curso, el cuál estará además complementado de bibliografía sugerida para cada unidad temática. Allí, se dispondrán los contenidos teóricos mínimos así como también diversos ejemplos de análisis, cuyo código en R estará listo para que los alumnos puedan, en simultáneo, replicar los ejemplos mostrados. Durante las clases teórico-prácticas, el alumno tendrá participación activa, debatiendo los contenidos presentados y planteando inquietudes que serán debidamente atendidas. Los problemas de aplicación que deberán ser resueltos por los alumnos, requerirán de la formación de grupos de trabajo de 2 a 3 miembros. Esto fomentará tanto el trabajo en grupo como un espacio para el debate de las posibles soluciones que pueden plantearse para un mismo problema. Durante la resolución de los problemas, los alumnos recibirán el apoyo necesario, aunque el objetivo principal de esta instancia es que desarrollen habilidades para trabajar en forma autónoma.

La integración de los contenidos se logra mediante la resolución de problemas integradores. Para ello, se proveerá a los alumnos de un conjunto de problemas de diversas áreas temáticas, entre los cuales ellos deberán seleccionar dos para su resolución. Cada grupo de alumnos deberá resolver dichos problemas y redactar un informe describiendo la estrategia utilizada y los resultados obtenidos. Toda esta información deberá estar fundamentada y complementada mediante gráficos y tablas adecuados. La evaluación de los mismos será parte de la nota final de los alumnos.

Se prevé además dedicar un tiempo a la búsqueda, estudio e implementación de trabajos científicos que

apliquen métodos estadísticos concretos. Esta tarea está dirigida a que los alumnos tomen contacto con el ámbito científico y comprendan la importancia de la correcta divulgación de los resultados. Los alumnos deberán desarrollar un trabajo final, en grupos de 2 a 3 miembros, replicando parcial o totalmente el análisis descrito en el trabajo científico. El informe de su trabajo deberá ser presentado en forma impresa y será defendido en forma oral ante toda la clase. El objetivo de este trabajo es lograr que los alumnos tengan una visión integradora de todos los contenidos desarrollados en la asignatura y que puedan aplicar el conocimiento adquirido en un ejemplo real.

También se dispondrá de un horario de consulta semanal destinado a evacuar todas aquellas dudas que los alumnos no hayan podido resolver durante las clases teórico-prácticas.

Formación Práctica:

La asignatura propone una formación práctica sólida. Es por ello que la primer clase, introduce al alumno directamente en el uso del software estadístico R. Los conocimientos teóricos, serán reforzados con ejemplos prácticos que deberán ser ejecutados por los alumnos, para adquirir, en forma paulatina, las habilidades necesarias para el uso de este programa informático. La segunda clase estará destinada a reforzar los conocimientos estadísticos adquiridos en las asignaturas anteriores. Al mismo tiempo, se presentará a los alumnos las distintas alternativas para implementar o utilizar estos conceptos dentro del software estadístico y aplicarlos a conjuntos de datos reales. Las clases subsiguientes estarán destinadas al desarrollo de contenido en forma teórica y complementada con actividades prácticas a desarrollar en el software. En cada unidad, se analizarán ejemplos aplicados a datos biológicos en donde se muestre la utilidad de las herramientas estudiadas así como también del uso del programa estadístico y sus funcionalidades.

La formación práctica de la asignatura será complementada con la resolución de problemas planteados en torno a situaciones reales e hipotéticas, cuya solución requiere la aplicación de conocimientos adquiridos en asignaturas del ciclo básico y complementarias a la presente materia. Dado que los problemas serán del tipo abiertos, será necesario que los alumnos formen grupos de trabajo, con el fin de generar el debate entre sus miembros y consensuar entre todos una única resolución. La actividad práctica se vuelve un desafío para el alumno, ya que no sólo deberán resolver un problema sino además deberán presentar un informe donde describan tanto las tareas desarrolladas como los resultados obtenidos, fundamentando la metodología utilizada. Adicionalmente, el trabajo final que debe presentar el alumno, que requiere de la implementación de la solución práctica utilizada en un trabajo científico, se presenta como una oportunidad valuable para trabajar en un problema real, como el que podría enfrentarse en su futuro profesional. Es por eso que, la aprobación de este trabajo consiste tanto en la redacción de un informe como en una defensa oral del mismo.

Listado de Actividades de Formación Práctica:

Los trabajos prácticos (TP) que se desarrollarán a lo largo de la asignatura se corresponden con sus seis unidades temáticas:

TP1: Introducción a R.

TP2: Estadística descriptiva en R.

TP3: Modelos lineales en R.

TP4: Modelos lineales avanzados.

TP5: Herramientas para el análisis multivariado.

TP6: Estadística no paramétrica en R.

Adicionalmente, se trabajará con dos guías de problemas integradores (GPI) entre los cuales cada grupo de alumnos deberá seleccionar un problema, que deberá resolver utilizando el software estadístico y las herramientas vistas en las unidades abarcadas por cada uno de ellos. La GPI1 abarca los contenidos de las primeras tres unidades mientras que la GPI2, los de las últimas unidades.

Intensidad de la formación práctica

Detalle de la carga horaria total prevista para cada una de las siguientes actividades:

Actividades prácticas que aportan a las competencias específicas en el Nivel de dominio 1: 0 horas

Actividades prácticas que aportan a las competencias específicas en el Nivel de dominio 2: 0 horas

Actividades prácticas que aportan a las competencias específicas en el Nivel de dominio 3: 0 horas

Horas totales de actividades de formación práctica: 0 horas

Metodología de Evaluación Durante el cursado:

La asignatura cuenta con dos instancias distintas de evaluación: a) Resolución de problemas integradores y b) Trabajo Final.

a) Resolución de problemas integradores: las evaluaciones de los informes de resolución de los problemas integradores planteados en las guías de problemas integradores (GPIs) estarán destinadas a acreditar el aprendizaje, la comprensión y la capacidad de relacionar los conceptos desarrollados en las clases e incluidos en la guía de trabajos prácticos (GTP). La evaluación se realizará a partir del informe escrito presentado por cada grupo de alumnos (dos a tres miembros), el cual deberá, al menos, contener las siguientes secciones:

- Introducción: presentación del problema a resolver
- Métodos: descripción de la estrategia y herramientas de análisis utilizada, justificando la elección.
- Resultados: conteniendo los resultados obtenidos en forma de tablas y/o gráficos, incluyendo, cuando fuese necesario, fragmentos del código utilizado en el análisis y su correspondiente salida.
- Discusión: debatiendo acerca de los resultados obtenidos y/o métodos utilizados.

Cada informe será evaluado en una escala del 0 al 100. La evaluación considerará tanto la elección e implementación de la estrategia de análisis como la calidad de la presentación del informe. El examen se considerará aprobado cuando se haya conseguido un puntaje de 60 o superior.

b) Trabajo final: La evaluación del trabajo final consiste de dos partes. Por un lado, se evaluará el informe escrito presentado por el grupo de alumnos y por otro, la defensa oral de dicho informe. Durante el cursado, los alumnos deberán realizar una búsqueda bibliográfica con el objetivo de encontrar un trabajo científico, que sea de su interés, donde se aplique alguna de las metodologías estadísticas estudiadas en la asignatura. La temática y el alcance del trabajo deberán acordarse con el docente responsable antes de comenzar el trabajo. En base a la publicación seleccionada, deberán replicar el trabajo o parte de él que involucra los conceptos de interés. Una vez hecho esto, deberán presentar un informe, con la misma estructura de los informes de resolución de los problemas integradores. Posteriormente, deberán defender su trabajo delante de toda la clase, con una presentación de 15 minutos. Dado que la defensa es grupal, los alumnos deberán organizarse previamente de modo de asegurar la participación de todos los miembros del equipo en la defensa. Tanto el docente como los alumnos participarán luego en una instancia de preguntas/discusión del trabajo (5 minutos), la cual también será evaluada. El trabajo final será evaluado en una escala del 0 al 100 en forma individual. El 60% de los puntos se asignarán en función del informe escrito, por lo que todos los alumnos de un mismo grupo tendrán el mismo puntaje. El 40% restante será de

carácter individual y se obtendrá en función de la participación y calidad de la defensa del informe (30%) y en función de la participación del alumno en el debate de los otros trabajos (10%).

Metodología de Evaluación en Exámenes Finales:

Evaluación del alumno regular: La evaluación del alumno regular será de carácter individual y se basará en los dos trabajos integradores desarrollados durante el cursado de la asignatura y en el trabajo final. Cada alumno deberá presentar impreso los informes de los dos trabajos integradores y del trabajo final corregidos acorde a las sugerencias que el docente haya hecho en las evaluaciones correspondientes. Luego, los alumnos deberán defender oralmente sus trabajos, respondiendo las preguntas que el docente considere pertinentes respecto de la metodología utilizada. Para aprobar la asignatura el alumno deberá lograr al menos 60 puntos sobre un total de 100.

Evaluación del alumno libre: Todo aquel alumno que cuente con las materias correlativas aprobadas, podrá rendir de forma libre la asignatura. Para ello, deberá comunicarse de forma fehaciente a la cátedra con al menos 15 días de anticipación su intención de rendir con esta modalidad. La cátedra preparará un examen integrador de todas las unidades, cuya resolución deberá realizarse en computadora, utilizando el programa estadístico R. Los resultados obtenidos deberán organizarse en un informe que deberá entregarse, junto con el código R utilizado para el procesamiento de los datos, en el correspondiente turno de examen. En dicha instancia también deberá defender el trabajo en forma oral ante el tribunal evaluador, con una presentación de 15 minutos, seguida de una instancia de preguntas de 5 minutos. Adicionalmente, se le presentará un conjunto de ejercicios que el alumno deberá resolver en esa instancia para demostrar sus habilidades tanto respecto de los contenidos como del uso del software estadístico. El contenido y presentación del informe, la metodología utilizada, la defensa oral y la resolución de los ejercicios presentados durante la mesa de exámen serán conjuntamente evaluados para obtener una calificación global en una escala del 0 al 100. Para aprobar la asignatura el alumno deberá lograr al menos 60 puntos.

Condiciones de Regularidad :

En términos generales, la condición de promoción de esta asignatura implica que el alumno ha alcanzado un nivel de conocimientos, destrezas y habilidades suficientes, asociadas al objetivo de la asignatura. El alumno regular será aquel cuyos conocimientos adquiridos son mínimos, de modo que le permitirán preparar, sin mayores dificultades, el examen final de esta asignatura. Teniendo en cuenta esto, al final del cursado, un alumno al final del cursado de la asignatura puede lograr cuatro condiciones (Regular, Promocionado, Libre por evaluación y Libre por inasistencia) que se describen a continuación:

a) Regular.

Para acceder a la condición de regularidad, el alumno deberá cumplir la totalidad de los siguientes requisitos:

- Asistir al menos al 75% de las clases.
- Aprobar la evaluación de los problemas integradores. Logrará esta condición cuando, junto con su grupo de trabajo, haya presentado los dos informes correspondiente, habiendo alcanzado una calificación igual o superior a 60 puntos en ambos informes. En caso de haber desaprobado uno de los dos informes, el grupo podrá acceder a una instancia de recuperatorio donde deberán reevaluar/rehacer el informe siguiendo las correcciones del docente encargado.
- Aprobar la implementación y presentación del trabajo final. Logrará esta condición cuando, junto con su grupo de trabajo, haya presentado el informe correspondiente, habiendo alcanzado una calificación igual o superior a 60 puntos.

b) Promoción.

Accederá a la condición de promovido aquel alumno que además de haber alcanzado los requisitos exigidos para la regularidad, haya cumplido con las siguientes condiciones:

Obtener un promedio de 80 puntos en las evaluaciones de problemas integradores.

Obtener una calificación del trabajo final superior a 80 puntos.

c) Libre por evaluación.

Se considera libre por evaluación a aquel alumno que:

- acredite asistencia al menos al 75% de las clases.
- obtuvo calificación menor a 60 puntos en alguno de los informes de resolución de los problemas integradores, habiendo utilizado las instancias de recuperatorio, o en la implementación y presentación del trabajo final.

d) Libre por inasistencia.

Se considera libre por inasistencia a aquel alumno que:

- No acredita un mínimo de asistencia a clases del 75%.
- No presentó alguno de los informes de resolución de los problemas integradores, incluso en las instancias de recuperatorio, o del informe del trabajo final, o no participó de la instancia de defensa del trabajo final.



Cronograma de parciales durante el primer Cuatrimestre:

Cronograma de parciales durante el segundo Cuatrimestre:

Bibliografía Principal:

Cabe destacar que no existe un único libro que cubra por completo los contenidos que involucra la asignatura. Es por ello que se torna necesario consultar a varios títulos que cubren aspectos específicos para cada unidad temática. El contenido teórico debe completarse con aplicaciones y ejemplos que pueden ser tomados tanto de libros como de artículos científicos.

EVERITT, Brian; HOTHORN, Torsten. An introduction to applied multivariate analysis with R. Springer Science & Business Media, 2011.

FISCHETTI, Tony. Data Analysis with R. Packt Publishing Ltd, 2015.

HOTHORN, Torsten; EVERITT, Brian S. A handbook of statistical analyses using R. Chapman and Hall/CRC, 2009.

UGARTE, Maria Dolores; MILITINO, Ana F.; ARNHOLT, Alan T. Probability and Statistics with R. CRC Press, 2008.

VENABLES, W. N.; SMITH, D. M. An Introduction to R: Notes on R, A Programming Environment for Data Analysis and Graphics, v. 3.5. 0. Network Theory, Bristol, UK, 2018.

WEHRENS, Ron. Chemometrics with R: multivariate data analysis in the natural sciences and life sciences. Springer Science & Business Media, 2011.

WEST, Brady T.; WELCH, Kathleen B.; GALECKI, Andrzej T. Linear mixed models: a practical guide using statistical software. Chapman and Hall/CRC, 2014.

ZELTERMAN, Daniel. Applied multivariate statistics with R. Cham: Springer, 2015.

Bibliografía Complementaria:

BERRIDGE, Damon Mark; CROUCHLEY, Robert. Multivariate generalized linear mixed models using R. CRC Press, 2011.

FARAWAY, Julian J. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC press, 2016.

Equipo de Cátedra:

Único personal responsable: Profesor Titular: Dra. Mg. Bioing. Gabriela Alejandra Merino (cargo interino, Res. "C.D." 215/18)

La Profesora responsable se encargará de elaborar todo el material didáctico necesario para el dictado de la asignatura. El material consistirá de:

- Clases en formato digital, en donde se desarrollarán los contenidos mínimos así como también ejemplos básicos y ejercicios para que los alumnos desarrollen durante la clase.
- Guías de Trabajos Prácticos, donde se propondrán ejercicios adicionales a los vistos en clase para que los alumnos refuercen los conceptos desarrollados.
- Guías de Problemas Integradores, que incluirán problemas integradores con datos biológicos que los alumnos deberán resolver para las instancias evaluatorias anteriormente especificadas.

Actividades de Investigación Gestión y Extensión:

La estructura actual del equipo de cátedra (una única persona) no permiten llevar a cabo actividades de investigación y extensión de la forma que se considera adecuada. Sin embargo, la Profesora responsable continuará trabajando en su línea de Investigación, la cual involucra el desarrollo de métodos estadísticos y bioinformáticos para el análisis de datos ómicos. Adicionalmente, se trabajará en la formulación de un proyecto de investigación y desarrollo dirigido al análisis estadístico de datos ómicos para la identificación de marcadores moleculares en cáncer colorrectal. La Profesora responsable también participará, junto con docentes de otras cátedras, en otro proyecto de investigación y desarrollo destinado al desarrollo de herramientas cuantitativas para soporte en la toma de decisiones de procesos de rehabilitación de patologías motrices.

La Profesora responsable también participará en el dictado de un curso de posgrado del Doctorado en Ingeniería.

Requisitos de admisión para alumnos oyentes:

Infraestructura, equipamiento y recursos necesarios:

Laboratorio de Computación con proyector y pizarrón, con computadoras con R y RStudio instalado, con acceso a internet.

Otros: