**بررسی رابطه بین ژنتیک میکروبیوم روده و اضافه وزن (چاقی) با استفاده از الگوریتم های یادگیری ماشینی**

**محقق: دانش مرادی گراوند**

**انستیتو سرطان و ژنومیک دانشگاه بیرمنگام انگلستان**

**استاد میزبان: دکتر بابک حسین خلج**

**گروه سیستمهای اطلاعات و علوم داده- پژوهشکده جامع علوم و فناوری های همگرا**

**دانشگاه صنعتی شریف**

**پروژه جایگزین خدمت سربازی**

**تیر ۱۴۰۲**

**خلاصه:**

روده انسان جمعیت متنوعی از میکروارگانیسم‌ها را در خود جای داده است که مجموعاً به عنوان میکروبیوتای روده شناخته می‌شوند. این جمعیت ها نقشی اساسی در جنبه‌های مختلف سلامت انسان، متابولیسم و ​​عملکرد سیستم ایمنی ایفا می‌کند. تحقیقات اخیر پیوندهای بالقوه بین تغییرات در میکروبیوتای روده و ایجاد اختلالات متابولیک از جمله چاقی را نشان می دهند . این پروژه به رابطه ی بین میکروبیوتای روده و چاقی می پردازد و تلاش دارد چاقی را بر اساس نشانه های میکروبیوم پیش بینی کند. شواهد در حال ظهور ارتباط بین تغییرات در ترکیب میکروبیوتای روده و چاقی را اثبات می‌کند، بنابراین نشان می‌دهد که ترکیب میکروبیوتای روده می‌تواند اطلاعات پیش‌بینی‌کننده قابل‌توجهی برای تشخیص زودهنگام و پیش‌آگهی چاقی داشته باشد. در این پروژه، ما از قدرت الگوریتم‌های یادگیری ماشین برای پیش‌بینی وضعیت چاقی با استفاده از گونه‌ها و اطلاعات عملکردی مشتق‌شده از میکروبیوتای روده ۲۴۹ نمونه انسانی استفاده کردیم. یافته‌های ما دقت پیش‌بینی بالایی را نشان داد (AUC-ROC = 0.95) و امکان شناسایی گونه‌های کلیدی و مسیرهای متابولیکی دخیل در چاقی را فراهم کرد. برای اینکه مدل پیش‌بینی خود را در دسترس جامعه پژوهشی قرار دهیم، یک نرم افزار را توسعه دادیم که هم از طریق پایانه دستورمحور و هم فضای گرافیکی دارد.

**معرفی:**

**>>> میکروبیوتا و نقش آن در بیماری متابولیک**

میکروبیوتای روده از تریلیون ها میکروارگانیسم از جمله باکتری ها، ویروس ها، قارچ ها و سایر میکروب ها تشکیل شده است. این میکروارگانیسم ها با یکدیگر و با بدن میزبان تعامل دارند و بر فرآیندهای فیزیولوژیکی مختلف تأثیر می گذارند. یکی از نقش های کلیدی میکروبیوتای روده دخالت آن در متابولیسم است[1-3]. میکروبیوتای روده به تجزیه و جذب مواد مغذی غذایی کمک می کند، متابولیت ها را تولید می کند و بر تعادل انرژی تأثیر می گذارد. بنابراین اختلال در تنظیم میکروبیوتای روده می تواند این عملکردها را مختل کند و منجر به عدم تعادل متابولیک شود و به توسعه بیماری های متابولیک کمک کند[4].

**>>> تعریف چاقی مفرط**

چاقی یک بیماری پیچیده و چند عاملی است که با تجمع بیش از حد چربی در بدن تشخیص داده می شود. چاقی مفرط معمولاً بر اساس شاخص توده بدنی (BMI) تعریف می شود که به صورت وزن بر حسب کیلوگرم تقسیم بر مجذور قد بر حسب متر محاسبه می شود. با این حال، چاقی تنها با وزن بدن فرد تعیین نمی شود و معلول یک تعامل پیچیده از عوامل ژنتیکی، محیطی و سبک زندگی است. چاقی با طیف وسیعی از عوارض سلامتی از جمله دیابت نوع 2، بیماری های قلبی عروقی و انواع خاصی از سرطان همراه است. خاورمیانه از جمله ایران با چالش مهم و رو به رشدی در قالب چاقی مواجه است. در سال‌های اخیر، افزایش قابل توجهی در نرخ چاقی در سراسر منطقه وجود داشته است که ایران به‌ویژه تحت تأثیر قرار گرفته است[5, 6]. ترکیبی از شهرنشینی سریع، سبک زندگی بی تحرک، الگوهای غذایی ناسالم و عوامل فرهنگی به افزایش هشدار دهنده شیوع چاقی کمک کرده است. چاقی در ایران یک مشکل چندوجهی با پیامدهای گسترده برای سلامت عمومی است [7, 8]. پرداختن به این موضوع نیازمند رویکردی جامع است که شامل روش های تشخیص زودهنگام چاقی و عوامل موثر در آن است.

**>>> پیوندها و شواهد بین چاقی و میکروبیوتا**

شواهد در حال افزایش نشان می دهد که تغییرات در ترکیب و عملکرد میکروبیوتای روده با چاقی مرتبط است. برای مثال، مطالعات نشان داده است که افراد چاق اغلب در مقایسه با افراد لاغر، میکروبیوتای روده تغییر یافته ای دارند که با کاهش تنوع میکروبی و نمایش بیش از حد گونه های خاص باکتری مشخص می شود[9, 10]. علاوه بر این، مطالعات حیوانی نشان داده‌اند که پیوند میکروبیوتای روده از افراد چاق به موش‌های بدون میکروب می‌تواند باعث افزایش وزن و ناهنجاری‌های متابولیک شود و نقش بالقوه میکروبیوتای روده را در ایجاد چاقی برجسته کند [9, 10].

**>>> فن آوری های توالی یابی ژنتیکی و یادگیری ماشین**

ترکیب میکروبیوتای روده را می توان با استفاده از فناوری های توالی یابی پیشرفته ژنتیکی، که اطلاعات ارزشمندی را در مورد ساختار و تنوع جامعه میکروبی ارائه می دهد، تجزیه و تحلیل کرد . براساس اطلاعات که شامل داده‌های ژنومی، مانند انواع ژنتیکی و پروفایل‌های بیان ژن هستند، محققان می‌توانند برهمکنش‌های بالقوه بین میزبان و میکروبیوتا را که بر صفات پیچیده تأثیر می‌گذارد، کشف کنند. علاوه بر این، تجزیه و تحلیل مسیر امکان بررسی فرآیندهای بیولوژیکی و مسیرهای مولکولی را فراهم می کند که از طریق آن میکروبیوتای روده ممکن است بر فیزیولوژی و بیماری انسان تأثیر بگذارد. در این مطالعه، ما یک پایگاه داده انتخاب شده حاوی اطلاعات در مورد گونه‌ها و فراوانی مسیر برای گروه‌های چاق و غیرچاق افراد سالم ایجاد کردیم. با استفاده از این مجموعه داده برچسب‌گذاری شده (labeled)، ما مدل‌های یادگیری ماشینی را برای پیش‌بینی برچسب‌های افراد چاق در مقابل افراد غیرچاق آموزش دادیم. پس از آن، ما از ویژگی‌های تفسیرپذیری (interpretability) برای شناسایی تأثیرگذارترین ویژگی‌ها برای پیش‌بینی چاقی استفاده کردیم.

**مواد و روش ها**

**داده های ورودی**

در این پروژه از پایگاه داده استاندارد HumanMetagenomDB استفاده کردیم[11]. هدف اصلی HumanMetagenomDB ساده‌سازی شناسایی و استفاده از متاژنوم‌های یعنی محتوای ژنتیکی در میکروبیوم است. HumanMetagenomDB نسخه 1.0 حاوی ابرداده 69822 متاژنوم است. این پایگاه شامل 203 ویژگی بر اساس توصیف ویژگی های میزبان (مانند جنس، سن و شاخص توده بدن)، اطلاعات تشخیصی (مانند سرطان، بیماری کرون و پارکینسون)، مکان (به عنوان مثال کشور، طول جغرافیایی و عرض جغرافیایی)، محل نمونه برداری (مثلاً). روده، ریه و پوست) و ویژگی های توالی (مانند پلت فرم توالی، طول متوسط ​​و کیفیت توالی) است. HumanMetagenomDB نسخه 1.0 شامل 58 کشور، 9 سایت نمونه اصلی (یعنی اعضای بدن)، 58 تشخیص و سنین متعدد، از تازه متولد شده تا 91 سالگی است. ما توالی کوتاه (short reads) را برای ۲۵۰ نفر، متشکل از ۱۲۵ فرد چاق و ۱۲۵ فرد غیر چاق، به دست آوردیم. گروه غیر چاق شامل گروه های نرمال و کم وزن بر اساس شاخص توده بدنی بود. داده ها با (<https://hpc.nih.gov/apps/sratoolkit.html>) sratoolkit با استفاده از شماره دسترسی واکشی شدند. به طور متوسط ​​هر نمونه 457 مگابایت اندازه داشت. نمونه ها از افراد از کشورها و گروه های سنی مختلف برای حذف هر گونه اثر مخدوش کننده بودند. ما همچنین نمونه های توالی یابی شده بر روی پلت فرم های Illumina را برداشتیم تا تاثیر شیوه توالی یابی را از میان ببریم. شماره دسترسی و مشخصات نمونه ها در دایرکتوری GitHub پروژه ارائه شده است (<https://github.com/DaneshMoradigaravand/ObesityML>).

**خط تحلیل متاژنومیک**

پس از دریافت توالی‌های کوتاه خوانده شده، آنها را برای شناسایی فراوانی گونه‌ها و مسیرها تحت خط تحلیل قرار دادیم. خط تحلیل bioBakery، همراه با ابزارهای MetaPhlAn و HUMANN، مجموعه قدرتمندی از ابزارهای محاسباتی طراحی شده برای تجزیه و تحلیل داده های متاژنومی و کشف عملکرد جوامع میکروبی را نشان می دهد [12]. خط تحلیل با MetaPhlAn آغاز می‌شود که با کمی کردن دقیق فراوانی میکروارگانیسم‌ها در سطح گونه بر اساس ژن‌های نشانگر، پروفایل‌های طبقه‌بندی را فراهم می‌کند[13]. MetaPhlAn از مجموعه منحصربفردی از ژن‌های نشانگر اختصاصی استفاده می‌کند، که پروفایل‌های با وضوح بالا را حتی برای گونه‌های با فراوانی کم ممکن می‌سازد.

به دنبال پروفایل گونه ای، HUMANN (شبکه تجزیه و تحلیل متابولیک یکپارچه HMP) برای پیش بینی پروفایل های عملکردی (functional profiling) جمعیت میکروبی استفاده می شود. HUMANN تخمین فراوانی طبقه بندی MetaPhlAn را با پایگاه داده مرجع جامع خانواده های ژن میکروبی ادغام می کند. HUMANN با تراز کردن قرائت‌های متاژنومی با این پایگاه داده، می‌تواند فراوانی خانواده‌های ژن عملکردی و ماژول‌های مسیر را تعییت کند. این رویکرد به محققان اجازه می دهد تا اطلاعاتی در مورد قابلیت های متابولیکی بالقوه و تنوع عملکردی جامعه میکروبی به دست آورند. ما خروجی دو ابزار بالا را به عنوان داده های جدولی برای خط لوله یادگیری ماشین انتخاب کردیم. در داده های گونه ای ما سطوح تاکسونومی شاخه (phylum) و کلاس‌(class) را از گونه‌ها برای پیش‌بینی در نظر گرفتیم. فایل های خروجی به دایرکتوری GitHub بار گزاری شده است (<https://github.com/DaneshMoradigaravand/ObesityML> ) .

**خط تحلیل یادگیری ماشین**

در مطالعه خود، ما از یک رویکرد استاندارد برای تقسیم داده‌ها برای یادگیری ماشین به 80٪ مجموعه آموزشی و اعتبار سنجی (training and validation) و 20٪ مجموعه تست (test)استفاده کردیم. این تقسیم‌بندی به ما امکان داد تا عملکرد و توانایی تعمیم مدل‌های خود را به‌طور مؤثر ارزیابی کنیم. علاوه بر این، ما از اعتبارسنجی متقابل (cross-validation) 5 برابری (5-fold) در طول مرحله آموزش برای ارزیابی و اعتبارسنجی عملکرد مدل‌ها استفاده کردیم و احتمال بیش برازش (overfitting)را کاهش دادیم. این رویکرد به ما کمک کرد تا مدل های خود را بهینه کنیم و بهترین عملکرد را برای تجزیه و تحلیل بیشتر انتخاب کنیم.

برای آماده‌سازی داده‌های ورودی برای یادگیری ماشین، خصوصیات طبقه بندی شده حضور و یا عدم حضور مسیر متابولیکی را به نمایش عددی تبدیل کردیم که می‌تواند توسط مدل‌ها پردازش شود. این فرآیند شامل رمزگذاری متغیرهای طبقه بندی و مقیاس بندی متغیرهای عددی از دست رفته بود.

به عنوان مدل‌های پایه، ما از طبقه‌بندی‌کننده رگرسیون لجستیک (logistic regression)، یک روش پرکاربرد و قابل تفسیر برای وظایف طبقه‌بندی دوگانه (binary) استفاده کردیم. برای مقایسه عملکرد آن، ما همچنین از یک مدل درخت تصمیم افزایش یافته با گرادیان (gradient boosted decision tree) با استفاده از کتابخانه scikit-learn استفاده کردیم. ما یک جستجوی شبکه ابرپارامتری (hyperparameter grid search) برای تنظیم عمق درخت و تعداد تکرارها در مدل تقویت‌شده گرادیان و عبارت جریمه منظم‌سازی در رگرسیون لجستیک انجام دادیم. این جستجو به ما امکان داد ترکیب بهینه ای از فراپارامترها را پیدا کنیم که سطح زیر منحنی مشخصه عملکرد گیرنده (AUC-ROC) را به حداکثر می رساند. این شاخص برای ارزیابی کیفیت طبقه بندی دوگانه استفاده شد.

پس از انتخاب بهترین مدل بر اساس AUC-ROC، تجزیه و تحلیل اهمیت ویژگی (SHapley Additive Explanations) را انجام دادیم. SHAP یک تکنیک مدل-آگنوستیک است که سهم هر ویژگی را در پیش بینی های مدل تعیین می کند. با تجزیه و تحلیل مقادیر SHAP، بینشی در مورد اهمیت و تأثیر هر ویژگی، یعنی گونه‌ها و مسیرهای متابولیکی، در فرآیند طبقه بندی مدل به دست آوردیم. این تجزیه و تحلیل به ما کمک کرد تا مکانیسم‌ها و محرک‌های عملکرد پیش‌بینی‌کننده را شناسایی کنیم . گونه ها و مسیر های متابویلیکی مهم برای پیش بینی چاقی را شناسایی کنیم.

**استقرار یادگیری ماشینی**

برای افزایش قابلیت استفاده و دسترسی به مدل یادگیری ماشین خود، آن را به یک برنامه خط فرمان (command line) و یک برنامه وب مبتنی بر Streamlit تبدیل کردیم که می تواند خروجی را برای هر ورودی دلخواه پیش بینی کند. با ایجاد یک برنامه خط فرمان، کاربران می توانند با ارائه داده های ورودی و دریافت خروجی پیش بینی شده، مستقیماً از ترمینال با مدل تعامل داشته باشند. کند.

علاوه بر این، ما یک برنامه Streamlit، یک چارچوب وب کاربر پسند، ایجاد کردیم تا یک فضای بصری برای کاربران فراهم کنیم تا داده‌های خود را وارد کنند و فوراً پیش‌بینی چاقی را انجام بدهند. برنامه Streamlit به کاربران اجازه می دهد تا به صورت دستی داده ها را از طریق یک رابط کاربری وارد کنند، و سپس خروجی پیش بینی شده را همراه با گونه های مهم یا اطلاعات تولید شده توسط مدل نمایش می دهد.این برنامه‌ها به کاربران اجازه می‌دهند بدون توجه به تخصص فنی یا آشنایی با کدنویسی، با استفاده از مدل یادگیری ماشینی بهینه سازی شده به راحتی پیش‌بینی کنند.

**کد و در دسترس بودن داده ها**

تمام کدها و مجموعه داده های ردیف و پاک شده به دایرکتوری GitHub این پروژه بارگزاری و قابل دسترس است ((<https://github.com/DaneshMoradigaravand/ObesityML>.

**نتایج و بحث:**

**>>> فراوانی برچسب در داده ورودی برای پیش بینی**

ما برچسب‌های چاق و غیرچاق را از داده‌های میکروبیوتا برای ۲۴۹ نفر پیش‌بینی کردیم. مجموعه داده ورودی شامل ۴۰۴ویژگی، از جمله ۲۷ ویژگی مربوط به فراوانی گونه‌ها و ویژگی‌های باقی‌مانده نشان‌دهنده مسیرها بود (نمودار ۱). نمودار A1 متداول ترین گونه در مجموعه داده را نشان می دهد. در سطح شاخه، اکثر سویه ها متعلق به Firmicutes، Bacteroidetes و Actinobacteria بودند. در سطح طبقه بندی بعدی، کلاس، کلستریدیا و باکتریایی غالب بودند، همانطور که برای میکروبیوتای انسانی انتظار می رفت [14, 15]. این یافته دقت و قابل اتکا بودن خط تحلیل در شناسایی گونه و مسیر متابولیکی را تأیید می کند. با توجه به فراوانی مسیر، ما مسیرهای متعددی را با فراوانی متفاوت شناسایی کردیم. فرکانس‌های مختلف این مسیرها اطلاعات پیش‌بینی ارزشمندی را ارائه می‌کنند که می‌توان از آن برای پیش‌بینی‌های مبتنی بر یادگیری ماشین استفاده کرد.

**>>> بررسی عملکرد مدل های یادگیری ماشین**

ما مدل یادگیری ماشین را بر روی فراوانی گونه‌ها و وجود/عدم وجود مسیرهای متابولیکی در نمونه‌ها آموزش دادیم. مجموعه داده به مجموعه داده‌های آموزشی (training) و آزمایشی (test) تقسیم شد و دو مدل آموزش داده شدند. عملکرد رگرسیون لجستیک پایه و درختان تصمیم تقویت شده با گرادیان(gradient boosted decision trees) نشان می دهد که هر دو مدل می توانند برچسب چاقی را به دقت پیش بینی کنند (نمودار ۲). درختان تصمیم تقویت شده با گرادیان عملکرد بهتری از رگرسیون لجستیک داشتند، با مقدار AUC-ROC 0.90 در مقایسه با 0.83 (نمودار 2A). با این حال، برای سایر معیارهای پیش‌بینی، به نظر می‌رسد هر دو مدل بیش برازش دارند، همانطور که با معیارهای سنجش طبقه بندی مختلف (F1، دقت، و بازیابی) برای مجموعه داده‌های آموزشی و آزمایشی نشان داده شده است. به طور متوسط، این تفاوت 0.15 0.18 بود. مدل رگرسیون لجستیک بازیابی (recall) بالاتری را نشان داد، در حالی که درخت تصمیم تقویت شده با گرادیان دقت (precision)بهتری را به همراه داشت (نمودار 2B). برای بررسی بیشتر این معیارها، ما آنها را از ماتریس‌های در هم ریختگی (confusion matrix) برای مدل‌های رگرسیون لجستیک و تقویت‌شده گرادیان استخراج کردیم، که دقت قابل مقایسه را نشان داد. با این حال، دقت و یادآوری برچسب چاق برای مدل تقویت‌شده گرادیان بالاتر بود (شکل 3). برعکس، ارزش بازیابی برای برچسب غیرچاق برای مدل رگرسیون لجستیک بالاتر بود. این نتایج به طور کلی نشان می دهد که مدل پیچیده تر عملکرد بهتری دارد و انتخاب یک مدل پیچیده تر را توجیه می کند. دو مدل سطحی از بیش برازش را نشان دادند، و این به دلیل تعداد بیشتر ویژگی های پیش بینی نسبت به اندازه نمونه است. برای رفع این مشکل یک مجموعه داده بزرگتر احتمالا مورد نیاز است. دقت بالای پیش‌بینی‌ها بر روی مجموعه داده‌های کوچک، دلالت بر سیگنال‌های پیش‌بینی قوی در داده‌های میکروبیوتا دارند که به طور موثر توسط مدل‌ها استفاده شده اند.

**>>> تجزیه و تحلیل اهمیت پیش بینی ویژگی های گونه ای و مسیرهای متابولیکی**

با توجه به عملکرد بهتر، ما با مدل تصمیم گیری تقویت شده با گرادیان پیچیده تر ادامه دادیم. با وجود دقت بالاتر پیش بینی، این مدل ها درک مکانیکی از ویژگی ها ارائه نمی دهند. برای به دست آوردن بینش و کمی کردن اهمیت ویژگی‌ها در داده‌های فراوانی تاکسون‌ها و مسیرهای متابولیکی، از سنجه ی SHAP استفاده کردیم (نمودار ۴). مقادیر SHAP سهم هر ویژگی را در پیش‌بینی نهایی برای هر نمونه، همانطور که در نمودار ۴ با نقاط مشخص شده است، نشان می‌دهد. این مدل ها عدم وجود برهم کنش بین ویژگی ها را فرض می کند (نمودار 4A). علاوه بر این، این مقادیر جهت اثر را نشان می‌دهند، یعنی آنکه ویژگی پیش‌بینی‌کننده برچسب چاق یا غیرچاق را بیشتر پیش‌بینی می کند. در میان ۲۰ ویژگی پیش‌بینی‌کننده برتر برای برچسب‌های چاقی و غیرچاقی، ۱۱ ویژگی مربوط به گونه و ۹ مورد در مورد مسیر متابولیک بودند. از ۱۱ ویژگی مربوط به گونه، ۶ مورد در سطح تاکسونومی کلاس و ۵ مورد در سطح شاخه بودند که در نمودار ۴ به ترتیب با پیشوند "c" برای کلاس ها و "p" برای شاخه تعیین شده اند (شکل 4A).

**>>> برترین ویژگی های پیش بینی گونه**

ما در مرحله بعد پنج تاکسون را بررسی کردیم که بیشترین پیشبینی را برای تمایز برچسب های چاقی و غیر چاقی فراهم می آوردند و سپس آنها را با شواهد موجود مورد بررسی قرار دادیم. نتایج نمودار ۴ نشان می دهند که پیش‌بینی‌کننده‌ترین سویه گروه باسیل است که حضور آن به شدت با صفت چاق همبستگی دارد (نمودار 4A و 4B). مطالعه پیشین سطوح بالاتری از گونه‌های خاص باسیلوس را در افراد چاق در مقایسه با افراد دارای وزن سالم مشاهده کرده‌اند که نشان‌دهنده یک همبستگی مثبت بالقوه بین فراوانی باسیل‌ها و چاقی است[16]. باسیل ها گونه های مختلفی را در بر می گیرند که به دلیل قابلیت های متابولیکی و تأثیر بالقوه بر فیزیولوژی میزبان شناخته شده اند. تحقیقات نشان می دهد که فراوانی و تنوع باسیل ها در میکروبیوتای روده ممکن است در ایجاد چاقی و اختلال در تنظیم متابولیک نقش داشته باشد[16]. به صورت مشخص Bacillus subtilis و سایر گونه‌های باسیلی اسیدهای چرب با زنجیره کوتاه (SCFAs) تولید می‌کنند و بر تولید هورمون‌ها و آنزیم‌های دخیل در ذخیره‌سازی و متابولیسم چربی تأثیر می‌گذارند. بنابراین تغییرات در فراوانی و فعالیت باسیل ها می تواند تولید و سیگنال دهی SCFA را تغییر دهد و منجر به اختلال در تعادل انرژی و تجمع چربی شود.

**>>> برترین ویژگی های پیش بینی گونه**

برخلاف باسیل ها، دومین گونه پیش بینی کننده، یعنی کلستریدیا، ارتباط منفی با چاقی نشان داد (نمودار 4A). کلستریدیا باکتری های بی هوازی هستند که چندین گونه را در بر می گیرند که به دلیل قابلیت های متابولیکی متنوع خود شناخته شده اند [17, 18]. آنها نقش مهمی در تخمیر فیبرهای غذایی، تولید SCFAs و تعدیل پاسخ های ایمنی میزبان دارند. مطالعات به طور مداوم نشان می دهد که فراوانی گونه های خاص کلستریدیا در افراد چاق در مقایسه با افراد دارای وزن سالم تغییر می کند[18]. به طور خاص، کاهش سطح کلستریدیاهای تولیدکننده بوتیرات، مانند [19]*Faecalibacterium prausnitzii* و *Roseburia spp*، به طور مداوم در افراد چاق مشاهده شده است. یافته های ما با این ارتباط منفی بین کلستریدیا و چاقی همخوانی دارد.

**>>> سایر گونه های پیش بینی کننده**

سه ویژگی پیش‌بینی‌کننده بعدی برای چاقی در سطح گونه ای، کلاس Negativicutes ، شاخه اکتینوباکتری‌ها و کلاس دلتاپروتئوباکتری‌ها بودند. حضور Negativicutes با برچسب چاقی همبستگی مثبت دارد، در حالی که ارتباط دلتاپروتئوباکتری ها و اکتینوباکتری ها منفی یا نامعلوم به نظر می رسد (شکل 4A، 4B). Negativicutes ، دسته ای از باکتری ها، در افراد چاق نسبت به افراد دارای وزن سالم افزایش یافته است. این باکتری ها به تولید اندوتوکسین هایی مانند لیپوپلی ساکاریدها (LPS) معروف هستند که می توانند باعث التهاب و اختلال عملکرد متابولیک شوند[20]. فراوانی بیشتر Negativicutes در چاقی، نقش بالقوه ای را در ترویج التهاب مزمن درجه پایین، نشانه اختلالات متابولیک مرتبط با چاقی، نشان می دهد. از سوی دیگر، دلتاپروتئوباکتری ها، دسته دیگری از باکتری ها، در افراد چاق کاهش یافته است [21]. در مقایسه با افراد لاغر این باکتری ها عملکردهای متابولیکی متنوعی از جمله کاهش سولفات دارند و در حفظ هموستاز روده نقش دارند. کاهش فراوانی دلتاپروتئوباکتری ها در چاقی ممکن است متابولیسم سولفات و عملکرد سد روده را مختل کند و به اختلال در تنظیم متابولیک کمک کند. برای Actinobacteria، یک پیوند پیچیده بین وجود تنها گونه‌های خاص در افراد چاق گزارش شده است که آن را به یک نشانگر زیستی کمتر قوی برای چاقی تبدیل می‌کند [22]. نتایج بالا به طور کلی نشان می دهد که پیش‌بینی‌کننده‌ترین گونه های برچسب چاقی با گزارش‌های موجود درباره نقش این سویه ها همسو هستند و نشان می‌دهند که این گونه‌ها می‌توانند به عنوان نشانگرها پ دیش بینی کننده های زیستی بالقوه برای چاقی عمل کنند.

**>>> مسیرهای متابولیکی پیش بینی کننده برای چاقی**

مشابه سویه های پیش بینی کننده، برخی از ویژگی‌های مرتبط با فراوانی مسیرهای متابولیکی پیش‌بینی‌کننده‌های قوی چاقی هستند. با این حال، پیوند دادن این ویژگی‌ها به چاقی ممکن است به اندازه گونه‌ها ساده نباشد، زیرا فراوانی فقط ممکن است منعکس کننده فراوانی بیش از حد یک گونه خاص باشد که مسیرهای خاصی را دارد. با این وجود، پیش‌بینی‌کننده‌ترین مسیرهای متابولیکی با نقش‌های متابولیکی تأییدشده‌شان در مطالعات پیشین همسو هستند. در این میان، پیش‌بینی‌کننده‌ترین ویژگی مسیر چاقی PWY-7198 بود (نمودار 4A و 4B)، که مربوط به پیریمیدین دئوکسی ریبونوکلئوتیدهای deoxyribonucleotides de novo biosynthesis IV است [23]. این مسیر متابولیک خاص، که در سنتز بلوک‌های سازنده DNA نقش دارد، در چاقی و اختلال عملکرد متابولیک نقش دارد. این مسیر نقش مهمی در تامین اجزای لازم برای تکثیر DNA و تکثیر سلولی ایفا می کند. اختلال در تنظیم بیوسنتز IV پیریمیدین دئوکسی ریبونوکلئوتید در افراد چاق مشاهده شده است که نشان دهنده دخالت بالقوه در ایجاد این بیماری است [23].

بر خلاف PWY-7198، مسیر PWY-5189 یک پیش بینی کننده قوی برای صفت غیرچاقی است و همبستگی منفی با چاقی دارد (نمودار 4A و 4B). بیوسنتز تتراپیرول II، به طور خاص از گلیسین، یک مسیر متابولیکی است که نقش حیاتی در تولید هم، کلروفیل و سایر ترکیبات ضروری دخیل در فرآیندهای سلولی مختلف دارد. شواهد جدید حاکی از ارتباط بالقوه بین اختلالات در بیوسنتز تتراپیرول II و چاقی است. اختلال در این مسیر می‌تواند سنتز Heme را که جزء حیاتی هموگلوبین و آنزیم‌های متعدد دخیل در متابولیسم انرژی است، مختل کند [24]. Hemeهمچنین در تنظیم تمایز سلول های چربی و متابولیسم لیپیدها نقش دارد. تغییرات در بیوسنتز تتراپیرول II ممکن است منجر به عدم تعادل در سطوح هم شود که بر عملکرد بافت چربی و مدیریت لیپیدها تأثیر می گذارد. نتایج ما نشان می دهد که عدم وجود این مسیر در افراد چاق شایع است. با این حال، مکانیسم دقیق اعمال این اثر نیازمند تجزیه و تحلیل بیشتر در سطح متابولوم است. مشابه یافته‌های مربوط به سویه ها، یافته های ما درباره ی قدرت پیش‌بینی مسیرهای متابولیکی برای چاقی همسویی با مطالعات و شواهد موجود را نشان می‌دهند. این یافته ها می‌توانند نشانگرهای زیستی برای پیش بینی چاقی را در سطح سلولی بالاتر متابولیکی ارائه دهند. تجزیه و تحلیل بیشتر علائم متابولومیک از داده های متابولومیکس می تواند مکانیسم های دقیق عمل را روشن سازد.

**>>> انتشار مدل در قالب نرم افزاری**

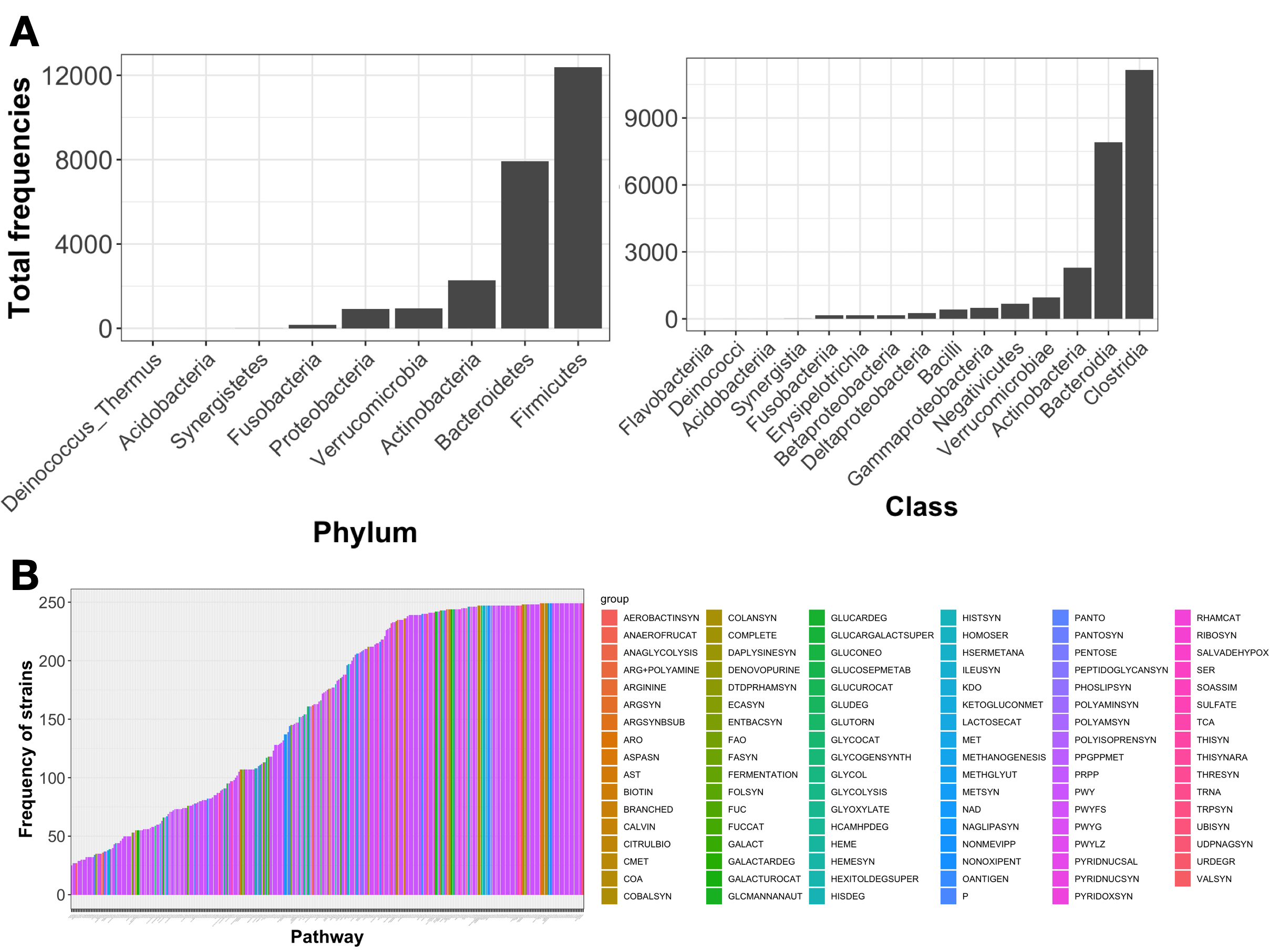
ما مدل خود را به عنوان یک راه حل نمونه اولیه علم داده برای کاربران نهایی ارایه نمودیم. بهترین مدل هم خط فرمان و هم رابط گرافیکی خروجی های خط لوله زیستی را می پذیرند و برچسب چاقی مرتبط را پیش بینی می کنند و بین افراد چاق و غیر چاق تمایز قائل می شوند. نمودار ۵ سه قابلیت در نرم افزار گرافیکی را نشان می دهد. هنگامی که کاربر فایل‌های خروجی مربوط به فراوانی گونه ها و مسیرهای متابولیکی از خط تحلیل متاژنومیک را آپلود کرد، داده به مدل تقویت‌شده گرادیان با بهترین عملکرد وارد می‌شوند و پیشبینی برچسب چاقی و غیرچاقی برآن ها صورت می گیرد. این ابزار علاوه بر ارائه پیش‌بینی‌ها، اهمیت ویژگی‌ها را برای پیش‌بینی برچسب با استفاده از نمودارهای اهمیت محلی SHAP نشان می دهند (نمودار 5B). ما همچنین به کاربران اجازه می‌دهیم که داده‌های جدولی را به دست آورند که مقادیر ارزش پیش بینی ویژگی های ورودی با همبستگی مثبت و منفی با چاقی را نشان می‌دهد. این مدل کاربردی بودن مدل های یادگیری ماشین را در تسهیل تشخیص سریع بیماری های پیچیده انسانی نشان می دهد.

**نتیجه گیری**

در این پروژه، ما آموزش و اعتبارسنجی مدل‌های یادگیری ماشینی را برای پیش‌بینی چاقی بر اساس مشخصات ژنومی میکروبیوم روده انجام دادیم. ارتباط بین میکروبیوتای روده و چاقی توسط چندین مکانیسم پیشنهادی تبیین می شود. این مکانیسم ها شامل افزایش استخراج انرژی از رژیم غذایی، افزایش ذخیره چربی، تغییر عملکرد سد روده، التهاب مزمن با درجه پایین و تعدیل هورمون های تنظیم کننده اشتها است. اگرچه مکانیسم های دقیق هنوز در دست بررسی هستند، اما بدیهی است که میکروبیوتای روده نقش مهمی در متابولیسم انرژی ایفا می کند و ممکن است در ایجاد و پیشرفت چاقی نقش داشته باشد [9, 10]. مطالعه ما گونه‌ها و مسیرهای متابویلیکی را به عنوان پیش‌بینی‌کننده‌های قابل اطمینان چاقی پیشنهاد می کند.

در دو حوزه می توان بهبودهای آتی برای تقویت این مطالعه در آینده انجام داد. اولاً، اضافه کردن داده های ژنومی بیشتر در پیش بینی کننده ها مفید خواهد بود. در حال حاضر، مجموعه داده های استاندارد شده در این زمینه به دلیل ناهمگونی داده های تولید شده در شرایط مختلف، محدود است. نیاز به چنین داده هایی به طور گسترده ای شناخته شده است و تقاضای فزاینده ای برای آن وجود دارد [25]. با مجموعه داده های بزرگتر، دقت مدل ها به طور بالقوه می تواند به سطحی برسد که امکان استفاده از این ابزارها را به عنوان کمک های تشخیصی در تنظیمات بالینی فراهم کند. ثانیا، ادغام سایر داده های omics به عنوان پیش بینی در مدل بسیار مهم است. این امر درک جامعی از محیط پیچیده اکوسیستم میکروبی روده را ممکن می سازد. بنابراین، پرداختن به چالش‌های مرتبط با مجموعه داده‌های بزرگ و متنوع، و همچنین ادغام داده‌های چند omics، یک اولویت برای پیشبرد این زمینه است.

مطالعه میکروبیوتای روده و تاثیر آن بر سلامت انسان در سال‌های اخیر توجه قابل توجهی را به خود جلب کرده است. محققان به طور فزاینده ای در حال بررسی پتانسیل استفاده از ترکیب میکروبیوتای روده، اطلاعات ژنومی و تجزیه و تحلیل مسیر برای پیش بینی صفات پیچیده انسانی هستند. این صفات از جمله چاقی، دیابت، بیماری التهابی روده و اختلالات سلامت روان ناشی از ترکیبی از عوامل ژنتیکی، محیطی و میکروبی هستند. با ادغام داده‌های ترکیب میکروبیوتای روده، ژنومیک و تجزیه و تحلیل مسیر، ما بینش‌هایی را در مورد روابط پیچیده بین این عوامل به دست می‌آوریم و مدل‌های پیش‌بینی‌کننده را برای صفات پیچیده مختلف توسعه می‌دهیم. این بیماری‌ها چالش‌های حادی در ایران ایجاد می‌کنند و تشخیص و توسعه روش‌های درمانی جدید می‌تواند کمک زیادی به ارتقای سلامت عمومی در کشور کند.



**نمودار 1** ترکیب مجموعه داده ورودی را نشان می دهد که شامل مسیرهای متابولیکی و سویه ها می باشد. در پانل A، توزیع گونه ها در هر دو سطح شاخه و کلاس ارائه شده است. محور y جمع فرکانس های نسبی مشاهده شده در تمام نمونه ها را نشان می دهد. پانل B توزیع مسیرها را با رنگ‌هایی که کلاس‌های مسیر مربوطه را نشان می‌دهند، نشان می‌دهد. محور y فرکانس تعداد سویه هایی را که شامل هر مسیر هستند را نشان می دهد.

A picture containing plot, line, diagram, text

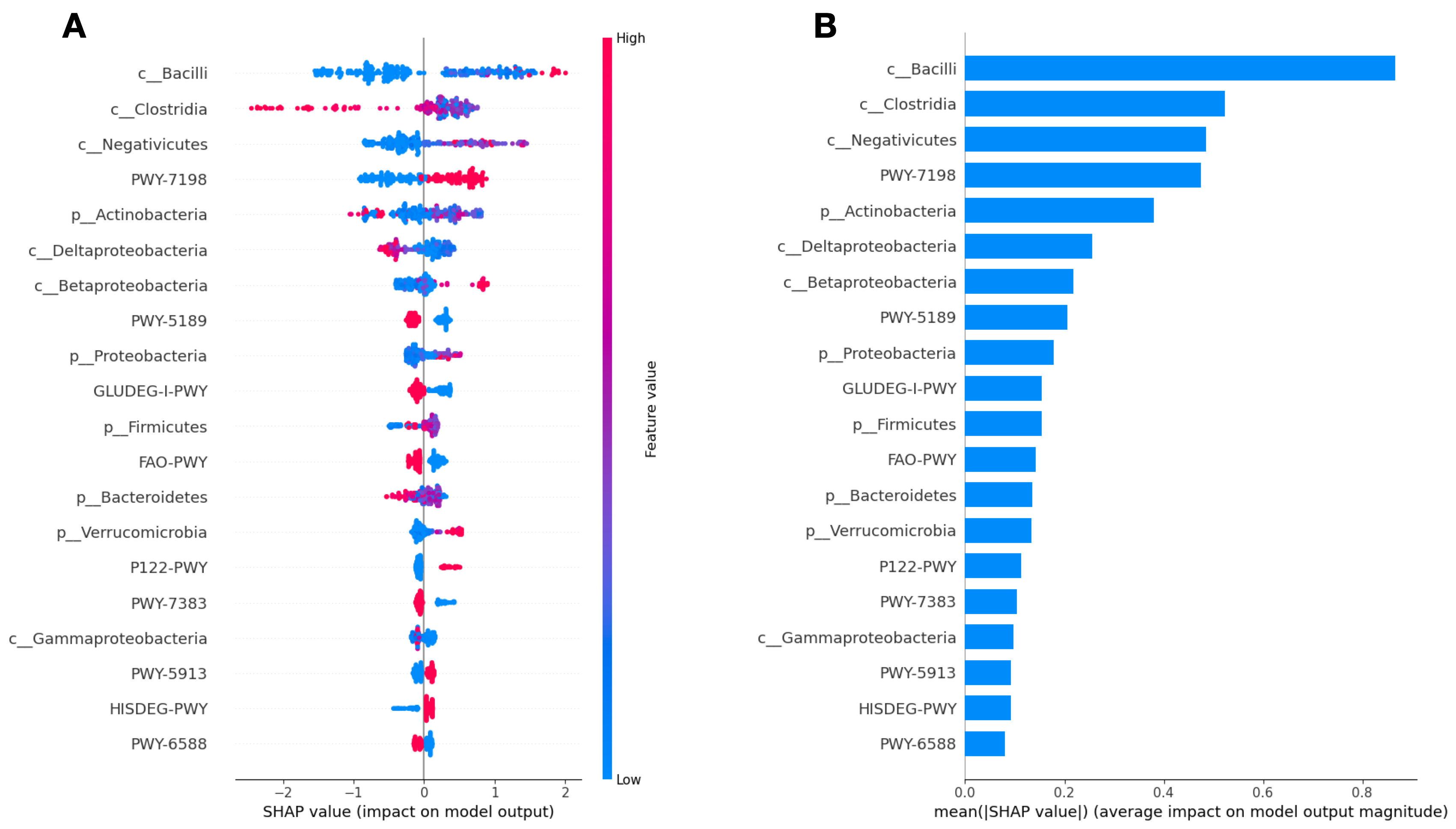
Description automatically generated

**نمودار 2** عملکرد رگرسیون لجستیک و درختان تصمیم تقویت شده با گرادیان را نشان می دهد. پانل A منحنی های ROC را به نمایش می گذارد، در حالی که پانل B عملکرد طبقه بندی دو مدل را در مجموعه داده های آزمایشی و آموزشی ارائه می دهد. نوارهای خطا در پانل B نشان‌دهنده فاصله اطمینان 95 درصدی است که از ده پیاده‌سازی تصادفی مدل‌ها در ده مجموعه داده آموزشی و آزمایشی به‌طور تصادفی حاصل شده است.

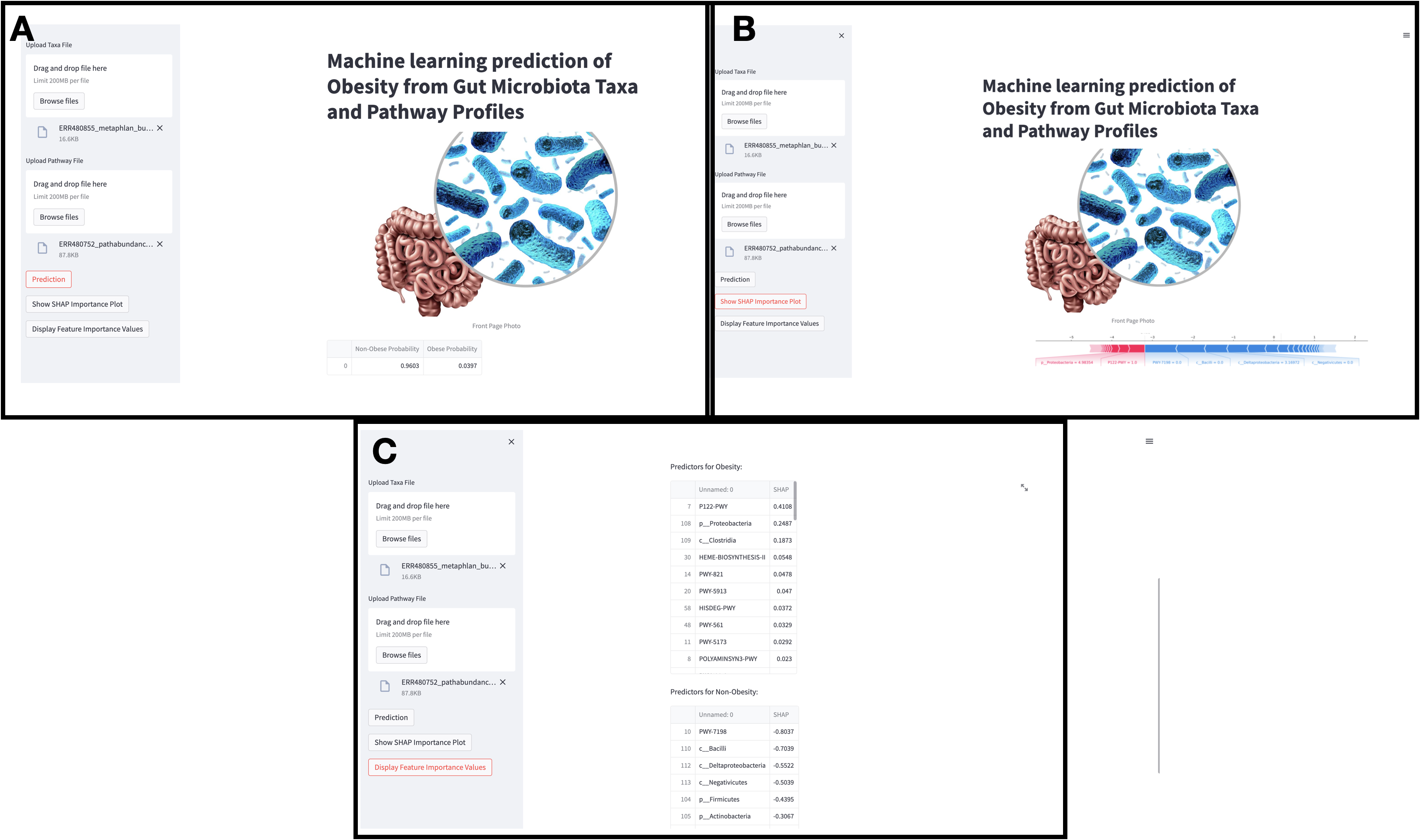
A white background with black text

Description automatically generated with low confidence

**نمودار 3** سنجه های پیش بینی کننده بر اساس ماتریس در همریختگی برای نمونه از داده‌های آموزشی و آزمایشی، با مقایسه درخت‌های تصمیم تقویت‌شده با گرادیان و طبقه‌بندی‌کننده‌های رگرسیون لجستیک نشان می‌دهد.



**نمودار 4** نتایج حاصل از تحلیل اهمیت ویژگی را نشان می دهد. A) توزیع مقادیر SHAP را با هر نقطه نشان دهنده یک نمونه نشان می دهد. رنگ های قرمز و آبی به ترتیب نشان دهنده گرایش به وضعیت های چاق و غیر چاق هستند. B) توزیع مقادیر متوسط ​​SHAP را در تمام نمونه ها نمایش می دهد.



**شکل 5** نمایشی از ویژگی های تعاملی پیاده سازی شده برای کاربر از طریق برنامه Streamlit. A) برنامه نتیجه احتمالی پیش‌بینی را بر اساس فایل‌های ورودی نشان می‌دهد، که احتمال وضعیت چاق و غیرچاق را نشان می‌دهد. B) تجزیه و تحلیل اهمیت ویژگی برای داده های ورودی ارائه شده است که رنگ های قرمز و آبی به ترتیب نشان دهنده تمایل به وضعیت چاق و غیر چاق هستند. C) جدول مهم‌ترین ویژگی‌های پیش‌بینی‌های چاق و غیرچاق را که از مقادیر SHAP استنباط می‌شود، نشان می‌دهد.

**منابع:**

1. Park, M., et al., *Effect of black ginseng and silkworm supplementation on obesity, the transcriptome, and the gut microbiome of diet-induced overweight dogs.* Sci Rep, 2021. **11**(1): p. 16334.

2. Cuevas-Sierra, A., et al., *Diet, Gut Microbiota, and Obesity: Links with Host Genetics and Epigenetics and Potential Applications.* Adv Nutr, 2019. **10**(suppl\_1): p. S17-S30.

3. Maruvada, P., et al., *The Human Microbiome and Obesity: Moving beyond Associations.* Cell Host Microbe, 2017. **22**(5): p. 589-599.

4. Li, S.Z., S.L. Zeng, and E.H. Liu, *Anti-obesity natural products and gut microbiota.* Food Res Int, 2022. **151**: p. 110819.

5. Asadi, A., et al., *Obesity and gut-microbiota-brain axis: A narrative review.* J Clin Lab Anal, 2022. **36**(5): p. e24420.

6. Kassaian, N., et al., *The effects of probiotic and synbiotic supplementation on metabolic syndrome indices in adults at risk of type 2 diabetes: study protocol for a randomized controlled trial.* Trials, 2017. **18**(1): p. 148.

7. Sadeghi, T., et al., *The prevalence and associated factors of overweight/obesity and abdominal obesity in South-eastern of Iran: a cross-sectional study based on Rafsanjan cohort study.* BMC Public Health, 2023. **23**(1): p. 861.

8. Zoghi, G., et al., *Prevalence of metabolically unhealthy obesity, overweight, and normal weight and the associated risk factors in a southern coastal region, Iran (the PERSIAN cohort study): a cross-sectional study.* BMC Public Health, 2021. **21**(1): p. 2011.

9. Liu, B.N., et al., *Gut microbiota in obesity.* World J Gastroenterol, 2021. **27**(25): p. 3837-3850.

10. Green, M., K. Arora, and S. Prakash, *Microbial Medicine: Prebiotic and Probiotic Functional Foods to Target Obesity and Metabolic Syndrome.* Int J Mol Sci, 2020. **21**(8).

11. Kasmanas, J.C., et al., *HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes.* Nucleic Acids Res, 2021. **49**(D1): p. D743-D750.

12. Beghini, F., et al., *Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.* Elife, 2021. **10**.

13. Manghi, P., et al., *MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice.* Cell Rep, 2023. **42**(5): p. 112464.

14. den Besten, G., et al., *The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism.* J Lipid Res, 2013. **54**(9): p. 2325-40.

15. Petriz, B.A., et al., *Exercise induction of gut microbiota modifications in obese, non-obese and hypertensive rats.* BMC Genomics, 2014. **15**(1): p. 511.

16. Kim, B., et al., *Protective effects of Bacillus probiotics against high-fat diet-induced metabolic disorders in mice.* PLoS One, 2018. **13**(12): p. e0210120.

17. Leung, J., et al., *Possible association between obesity and Clostridium difficile infection.* Emerg Infect Dis, 2013. **19**(11): p. 1791-8.

18. Obanda, D.N., et al., *Abundance of the species Clostridium butyricum in the gut microbiota contributes to differences in obesity phenotype in outbred Sprague-Dawley CD rats.* Nutrition, 2020. **78**: p. 110893.

19. Faria, A.M.C., B.S. Reis, and D. Mucida, *Tissue adaptation: Implications for gut immunity and tolerance.* J Exp Med, 2017. **214**(5): p. 1211-1226.

20. Hu, J., et al., *Gut Microbiota Signature of Obese Adults Across Different Classifications.* Diabetes Metab Syndr Obes, 2022. **15**: p. 3933-3947.

21. Shin, N.R., T.W. Whon, and J.W. Bae, *Proteobacteria: microbial signature of dysbiosis in gut microbiota.* Trends Biotechnol, 2015. **33**(9): p. 496-503.

22. Turnbaugh, P.J., et al., *A core gut microbiome in obese and lean twins.* Nature, 2009. **457**(7228): p. 480-4.

23. Khamzina, L., et al., *Increased activation of the mammalian target of rapamycin pathway in liver and skeletal muscle of obese rats: possible involvement in obesity-linked insulin resistance.* Endocrinology, 2005. **146**(3): p. 1473-81.

24. Yu, D., et al., *Long-term Diet Quality and Gut Microbiome Functionality: A Prospective, Shotgun Metagenomic Study among Urban Chinese Adults.* Curr Dev Nutr, 2021. **5**(4): p. nzab026.

25. Yilmaz, P., et al., *Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.* Nat Biotechnol, 2011. **29**(5): p. 415-20.