**ชื่อหัวข้อที่เลือกทำ (Term project Title)**

**Cancer Prediction from Genomic Analysis with Machine Learning**

**ชื่อสมาชิกในกลุ่ม (Team members)**

เตชิต กุลกัลยากรกมล 6531318421

วีรภัทร ล้ำศิริเจริญโชค 6531343021

จิรพนธ์ จิระธนานันต์ 6531307521

## Progress Report #1

**ที่มาและความสำคัญ (Background and Motivation / Rationale)**

โรคมะเร็งเป็นสาเหตุของการเสียชีวิตอันดับต้น ๆ ทั่วโลก และเป็นปัญหาทางการแพทย์ที่มีความซับซ้อน การวินิจฉัยมะเร็งในระยะเริ่มต้นสามารถเพิ่มโอกาสในการรักษาและการอยู่รอดของผู้ป่วยได้ ด้วยความก้าวหน้าทาง bioinformatics และการเข้าถึงข้อมูลทาง genome ที่มีอยู่มากขึ้น การศึกษาและการวิเคราะห์ genome ของผู้ป่วยจึงเป็นแนวทางที่สามารถใช้ในการพัฒนา model การทำนายที่จะช่วยให้สามารถวินิจฉัยและทำนายโอกาสในการเกิดมะเร็งได้

การใช้ machine learning ในโครงการนี้ช่วยในการประมวลผลและวิเคราะห์ข้อมูลที่ซับซ้อนและมีขนาดใหญ่ของ genome มนุษย์ โดยสามารถหาความสัมพันธ์ระหว่างลำดับ genome และโอกาสในการเกิดมะเร็ง ซึ่งวิธีการนี้อาจช่วยให้แพทย์มีเครื่องมือที่มีประสิทธิภาพในการวินิจฉัยและตัดสินใจทางการรักษาที่แม่นยำมากขึ้น อีกทั้งยังสามารถนำไปใช้เป็นพื้นฐานในการวิจัยและพัฒนา model การทำนายอื่น ๆ ที่เกี่ยวข้องกับโรคที่มีการแสดงออกของ gene

**งานวิจัยที่เกี่ยวข้อง (อย่างน้อย 3 เรื่อง) (Related Work)**

งานวิจัยที่ 1:Cancer Prediction from Genomic Analysis with Machine Learning

1) เป็นงานวิจัยเกี่ยวกับอะไร ทำไมถึงเป็นปัญหา (ให้สรุปมา) (Background and Motivation of this research)

งานนี้มุ่งเน้นการทำนายโรคมะเร็งจากการวิเคราะห์ genome ด้วยการใช้การเรียนรู้ของเครื่อง เพื่อแก้ปัญหาการวิเคราะห์ข้อมูลทางพันธุกรรมที่ซับซ้อนในการคาดการณ์มะเร็ง วิธีนี้มีศักยภาพในการช่วยแพทย์เข้าใจปัจจัยเสี่ยงและสัญญาณของโรค

2) สิ่งที่งานวิจัยนี้นำเสนอ และวิธีการที่ใช้ (Proposed methodology)

การศึกษาใช้ Machine Learning model เช่น Random Forest, SVM, และ Deep Learning ในการวิเคราะห์และตรวจจับรูปแบบของข้อมูล genome ที่เชื่อมโยงกับมะเร็ง

3) ข้อมูลที่ใช้ (Data)

ข้อมูล genome ที่รวบรวมมาจากกลุ่มผู้ป่วยโรคมะเร็งหลากหลายชนิดและจากกลุ่มควบคุมที่ไม่มีมะเร็ง ข้อมูลนี้ประกอบด้วยลำดับ DNA ที่ช่วยในการระบุความเสี่ยงและพยากรณ์โรค

4) ผลการทดลอง (Experimental results)

model ต่าง ๆ สามารถช่วยเพิ่มความแม่นยำในการทำนาย โดยเฉพาะการใช้การเรียนรู้เชิงลึกซึ่งมีประสิทธิภาพในการวิเคราะห์รูปแบบที่ซับซ้อนในจีโนมมากขึ้น ทำให้การทำนายมีความแม่นยำสูงกว่าเทคนิคดั้งเดิม

5) อภิปรายผลของงานวิจัยนี้ (Discussion)

งานวิจัยนี้ชี้ให้เห็นถึงศักยภาพของ Machine Learning ในการคาดการณ์และทำความเข้าใจความเสี่ยงต่อมะเร็ง ข้อมูลนี้สามารถนำไปใช้พัฒนาการแพทย์เฉพาะบุคคล (Personalized Medicine) โดยช่วยให้สามารถออกแบบการรักษาที่เหมาะสมกับลักษณะทางพันธุกรรมของผู้ป่วยแต่ละรายได้

งานวิจัยที่ 2: A Study on the Prediction of Cancer Using Whole-Genome Data and Deep Learning

1) เป็นงานวิจัยเกี่ยวกับอะไร ทำไมถึงเป็นปัญหา (ให้สรุปมา) (Background and Motivation of this research)

งานวิจัยนี้สำรวจวิธีพยากรณ์มะเร็งด้วยข้อมูล genome ทั้งหมด เพื่อแก้ปัญหาความไม่แม่นยำในการพยากรณ์จากความซับซ้อนและความหลากหลายของข้อมูลทางพันธุกรรม การพัฒนาวิธีที่แม่นยำจะเป็นประโยชน์อย่างมากต่อการรักษาและการป้องกันมะเร็ง

2) สิ่งที่งานวิจัยนี้นำเสนอ และวิธีการที่ใช้ (Proposed methodology)

ใช้การเรียนรู้เชิงลึกสร้าง model ที่สามารถตรวจจับความแตกต่างทางพันธุกรรมได้อย่างละเอียด ออกแบบให้จำแนกลักษณะพันธุกรรมที่บ่งบอกถึงความเสี่ยงต่อมะเร็งได้ดียิ่งขึ้น

3) ข้อมูลที่ใช้ (Data)

ข้อมูล genome ทั้งหมดของผู้ป่วยมะเร็งและคนปกติ ซึ่งมีขนาดใหญ่และครอบคลุมความหลากหลายของลักษณะทางพันธุกรรม

4) ผลการทดลอง (Experimental results)

การใช้ model นี้ช่วยเพิ่มความแม่นยำและประสิทธิภาพในการพยากรณ์ความเสี่ยงต่อมะเร็งเมื่อเปรียบเทียบกับวิธีการที่เคยใช้ โดยเฉพาะอย่างยิ่งในการตรวจจับความแตกต่างที่เล็กมากในข้อมูลพันธุกรรม

5) อภิปรายผลของงานวิจัยนี้ (Discussion)

การทดลองนี้ชี้ให้เห็นว่าเทคโนโลยี Deep Learning สามารถสนับสนุนการพยากรณ์โรคมะเร็งที่ดีกว่าเดิมได้ในอนาคต โดยแสดงให้เห็นถึงการใช้ข้อมูลพันธุกรรมในระดับที่ลึกขึ้น ซึ่งมีความเป็นไปได้ที่จะนำไปใช้จริงเพื่อเพิ่มความแม่นยำในการคัดกรองและป้องกันโรคมะเร็งในกลุ่มเสี่ยง

งานวิจัยที่ 3: Assessment of deep learning and transfer learning for cancer prediction based on gene expression data

1) เป็นงานวิจัยเกี่ยวกับอะไร ทำไมถึงเป็นปัญหา (ให้สรุปมา) (Background and Motivation of this research)

งานวิจัยนี้มุ่งเน้นที่การปรับปรุงประสิทธิภาพของโมเดล machine learning ในบริบทของข้อมูลที่มีมิติสูงแต่มีตัวอย่างข้อมูลน้อย ซึ่งเป็นปัญหาที่พบได้บ่อยในการวิเคราะห์ bioinformatics เช่น การศึกษาทางพันธุศาสตร์และการแพทย์ โดยเฉพาะในงานวิจัยที่ต้องใช้ข้อมูลทางจีโนมและการแสดงออกของยีน ข้อมูลประเภทนี้มักมีฟีเจอร์จำนวนมาก แต่จำนวนตัวอย่างมีจำกัด ซึ่งทำให้โมเดล machine learning มีความเสี่ยงต่อการเกิด overfitting และไม่สามารถทำนายได้แม่นยำกับข้อมูลใหม่ ดังนั้น งานวิจัยนี้มีแรงบันดาลใจในการนำเสนอแนวทางที่ช่วยแก้ปัญหานี้ เพื่อปรับปรุงประสิทธิภาพของโมเดลโดยใช้เทคนิคการคัดเลือกฟีเจอร์และการปรับแต่งพารามิเตอร์ที่เหมาะสม

2) สิ่งที่งานวิจัยนี้นำเสนอ และวิธีการที่ใช้ (Proposed methodology)

งานวิจัยนี้นำเสนอการผสมผสานของสองแนวทางหลักคือ การคัดเลือกฟีเจอร์ (Feature Selection) และ การปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) เพื่อปรับปรุงประสิทธิภาพของโมเดลแมชชีนเลิร์นนิงในข้อมูลที่มีมิติสูงและตัวอย่างน้อย แนวทางที่ใช้ประกอบด้วย:

- การเลือกใช้ เทคนิคการคัดเลือกฟีเจอร์ เช่นการใช้สถิติต่าง ๆ เพื่อคัดเลือกเฉพาะฟีเจอร์ที่มีความสำคัญและสัมพันธ์กับข้อมูลเป้าหมาย เพื่อลดมิติของข้อมูล
- การใช้ เทคนิคการปรับแต่งพารามิเตอร์ เช่น การทำ Grid Search หรือ Random Search ในการหาเซ็ตพารามิเตอร์ที่เหมาะสมสำหรับโมเดล โดยวิธีนี้ช่วยให้โมเดลปรับตัวกับข้อมูลที่มีมิติสูงได้ดีขึ้น

3) ข้อมูลที่ใช้ (Data)

งานวิจัยนี้ใช้ ข้อมูลจีโนมและการแสดงออกของยีน ที่มีมิติสูงและจำนวนตัวอย่างน้อย โดยข้อมูลถูกนำมาใช้ในการทดสอบประสิทธิภาพของโมเดล machine learning ที่หลากหลายเพื่อแสดงให้เห็นถึงประโยชน์ของการคัดเลือกฟีเจอร์และการปรับแต่งพารามิเตอร์ในข้อมูลประเภทนี้ ข้อมูลเหล่านี้มักประกอบด้วยหลายฟีเจอร์ที่แสดงถึงลักษณะของยีนและความสัมพันธ์กับโรคทางการแพทย์ต่าง ๆ เช่น การเกิดโรคมะเร็ง

4) ผลการทดลอง (Experimental results)

ผลการทดลองแสดงให้เห็นว่าเทคนิคที่งานวิจัยนี้นำเสนอสามารถปรับปรุงประสิทธิภาพของโมเดลได้อย่างชัดเจน โดยการใช้เทคนิคการคัดเลือกฟีเจอร์ช่วยลดจำนวนฟีเจอร์ที่ไม่สำคัญลง ส่งผลให้โมเดลมีความแม่นยำสูงขึ้นและประมวลผลได้รวดเร็วขึ้น นอกจากนี้ การปรับแต่งพารามิเตอร์ยังช่วยเพิ่มความแม่นยำและความน่าเชื่อถือของโมเดลในข้อมูลใหม่ ผลลัพธ์ของการทดลองจึงสนับสนุนว่าวิธีที่นำเสนอสามารถปรับปรุงประสิทธิภาพของโมเดลในกรณีที่ข้อมูลมีมิติสูงและตัวอย่างน้อยได้อย่างมีนัยสำคัญ

5) อภิปรายผลของงานวิจัยนี้ (Discussion)

งานวิจัยนี้ชี้ให้เห็นถึงความสำคัญของการใช้เทคนิคการคัดเลือกฟีเจอร์และการปรับแต่งพารามิเตอร์ในการวิเคราะห์ข้อมูลชีวสารสนเทศ โดยเฉพาะในกรณีที่ข้อมูลมีฟีเจอร์มากแต่จำนวนตัวอย่างน้อย ซึ่งเป็นเงื่อนไขที่ทำให้โมเดลแมชชีนเลิร์นนิงประสบปัญหาการทำนายที่ไม่แม่นยำ งานวิจัยเสนอว่าการคัดเลือกฟีเจอร์ช่วยลดความซับซ้อนของข้อมูล ในขณะที่การปรับแต่งพารามิเตอร์ช่วยเพิ่มประสิทธิภาพของโมเดลในแง่ของการทำนาย และแสดงให้เห็นถึงความเป็นไปได้ในการประยุกต์ใช้แนวทางนี้กับข้อมูลประเภทอื่น ๆ ที่มีลักษณะคล้ายกัน

## Progress Report #2

ต้องการเปรียบเทียบผลการทำงานของ Classification model อื่นๆ นอกเหนือจาก Random forest และ Extreme Gradient Boosting (XGBoost) เช่น Support Vector Machine, Light Gradient Boosting Machine, Logistic Regression และ Deeplearning Model เช่น Multi Layer Perceptron และ Autoencoder

โปรเจคต้นฉบับ "Cancer Prediction from Genomic Analysis with Machine Learning" มีจุดมุ่งหมายในการสร้าง machine learning model เพื่อทำนายความเป็นไปได้ของการเกิดโรคมะเร็งจากข้อมูลทาง genome แต่ปัญหาที่พบได้ในโปรเจคนี้มีหลายประการ เช่น

1. ข้อมูลที่ใช้ในโปรเจคนี้ มีเพียง 50% ของข้อมูลทั้งหมดจาก the Cancer Genome Atlas (TCGA) project conducted by the U.S. National Institutes of Health (NIH). ทำให้การวิเคราะห์และผลลัพธ์อาจมีข้อจำกัดในด้านความครอบคลุมและความแม่นยำของข้อมูล อย่างไรก็ตาม การใช้ข้อมูลที่มีอยู่ทั้งหมดอาจสามารถช่วยให้เห็นแนวโน้มและทำความเข้าใจลักษณะสำคัญบางประการที่เกี่ยวข้องกับโรคมะเร็งได้
2. ข้อมูลมีขนาดใหญ่และมี dimension สูง: ข้อมูล genome มีฟีเจอร์จำนวนมากที่อาจไม่สัมพันธ์กับการทำนายโดยตรง ทำให้การประมวลผลช้าลงและอาจส่งผลให้ model ขาดความแม่นยำ การทำ feature engineering เองอาจจะไม่เพียงพอในการลดจำนวนฟีเจอร์หรือเลือกฟีเจอร์ที่สำคัญได้อย่างมีประสิทธิภาพ อาจจำเป็นต้องใช้เทคนิคการลด dimension เช่น PCA หรือการใช้โมเดลที่สามารถทำ feature selection อัตโนมัติเพื่อปรับปรุงประสิทธิภาพและความแม่นยำในการทำนาย
3. ประสิทธิภาพของ model อาจไม่สูงเพียงพอ: แม้จะใช้โมเดลที่มีประสิทธิภาพสูงอย่าง Random Forest, Gradient Boosting อยู่แล้ว แต่ความท้าทายอาจเกิดจากการที่ข้อมูลมีความซับซ้อนสูงหรือมีฟีเจอร์ที่ไม่สัมพันธ์กับการทำนายโดยตรง ซึ่งอาจทำให้โมเดลเกิด overfitting หรือเรียนรู้ไม่เต็มที่ การปรับแต่ง hyperparameter, การเลือกฟีเจอร์ (feature selection) อย่างรอบคอบ หรือการใช้เทคนิค preprocessing ข้อมูลเพิ่มเติม เช่น normalization หรือการจัดการกับค่าที่หายไป อาจช่วยเพิ่มประสิทธิภาพของโมเดลได้มากขึ้น นอกจากนี้ การใช้เทคนิค Ensemble หรือการผสมผสานโมเดลหลายประเภทอาจเป็นอีกแนวทางหนึ่งในการเพิ่มประสิทธิภาพการทำนาย.

### ความแตกต่างในการพัฒนาจากงานวิจัยต้นฉบับ

เพื่อแก้ไขปัญหาดังกล่าว การพัฒนาโปรเจคนี้เพิ่มเติมโดยใช้แนวทางที่เรียบง่ายแต่มีประสิทธิภาพสามารถช่วยเพิ่มความแม่นยำและลดความซับซ้อนของการประมวลผลได้ ความแตกต่างหลัก ๆ มีดังนี้:

- การคัดเลือกฟีเจอร์อย่างเป็นระบบ: คัดเลือกเฉพาะฟีเจอร์ที่มีความสัมพันธ์สูงกับการเกิดมะเร็ง เช่น การใช้ SelectKBest เพื่อช่วยลดจำนวนฟีเจอร์และเพิ่มประสิทธิภาพของ model
- การเลือก model ที่เหมาะสมและใช้งานง่ายกว่า: เช่น CatBoost หรือ LightGBM ซึ่งใช้งานง่ายและเหมาะกับข้อมูลจำแนก โดยไม่ต้องการการปรับแต่งซับซ้อนแต่สามารถให้ความแม่นยำสูงขึ้น
- การประเมินโมเดลอย่างละเอียดขึ้น: ใช้ Cross-Validation เพื่อเพิ่มความน่าเชื่อถือของ model และประเมินผลลัพธ์ผ่าน Accuracy, Sensitivity, Specificity และ F1-Score เพื่อเห็นภาพรวมของประสิทธิภาพที่ชัดเจนขึ้น

### ขั้นตอนการพัฒนา

ขั้นตอนการพัฒนาแบบง่ายเพื่อแก้ไขปัญหาและพัฒนาต่อจากงานวิจัยต้นฉบับประกอบด้วย:

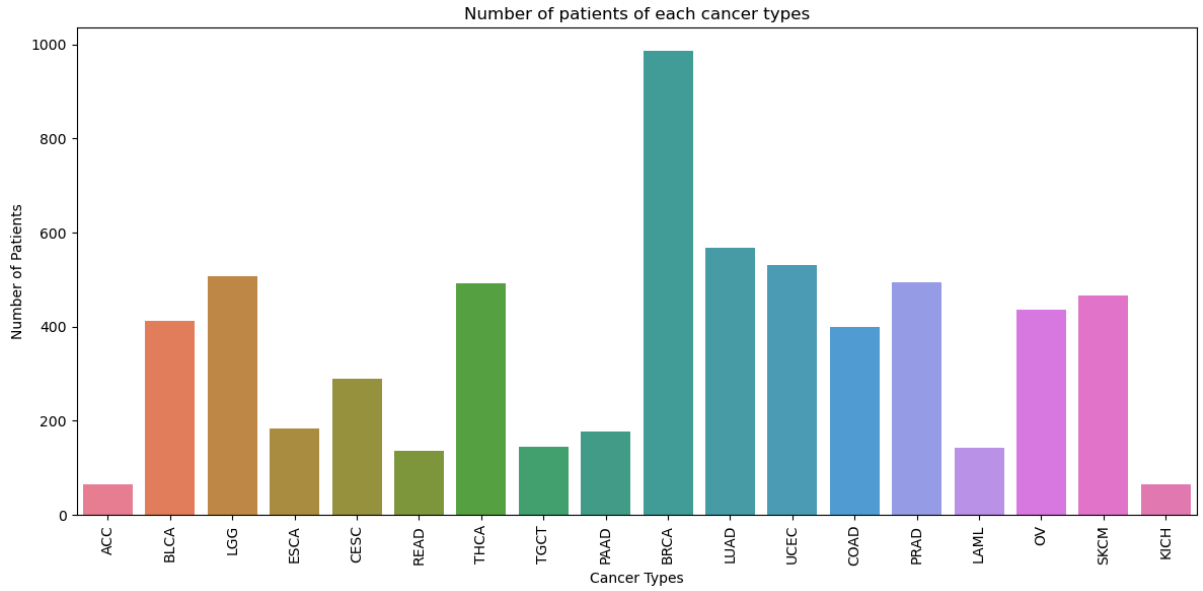
1. การเตรียมข้อมูลและคัดเลือกฟีเจอร์:

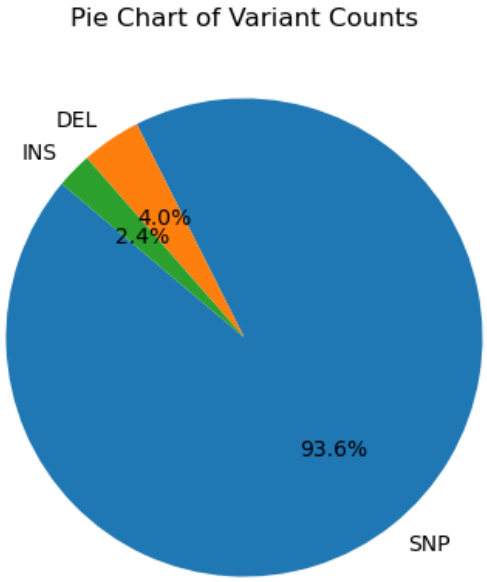
ใช้การทำ Scaling ข้อมูล โดยใช้ **StandardScaler** และจัดการข้อมูลที่ขาดหายเพื่อลดอัตราความคลาดเคลื่อน นอกจากนี้ การคัดเลือกฟีเจอร์อัตโนมัติโดยใช้ **SelectKBest** เพื่อเลือกเฉพาะฟีเจอร์ที่มีความสัมพันธ์สูงที่สุด 10 ฟีเจอร์กับมะเร็ง จะช่วยลด dimension และเพิ่มประสิทธิภาพของโมเดลได้

ข้อมูล จาก The Cancer Genome Atlas (TCGA) project conducted by the U.S. National Institutes of Health (NIH). ซึ่งประกอบไปด้วย ข้อมูล 36 คอลลัมน์ ได้แก่ #"chrom", chromStart, chromEnd, name, score, strand, thickStart, thickEnd, reserved, blockCount, blockSizes, chromStarts, sampleCount, freq, Hugo_Symbol, Entrez_Gene_Id, Variant_Classification, Variant_Type, Reference_Allele, Tumor_Seq_Allele1, Tumor_Seq_Allele2, dbSNP_RS, dbSNP_Val_Status, days_to_death, cigarettes_per_day, weight, alcohol_history, alcohol_intensity, bmi, years_smoked, height, gender, project_id, ethnicity, Tumor_Sample_Barcode, Matched_Norm_Sample_Barcode, case_id

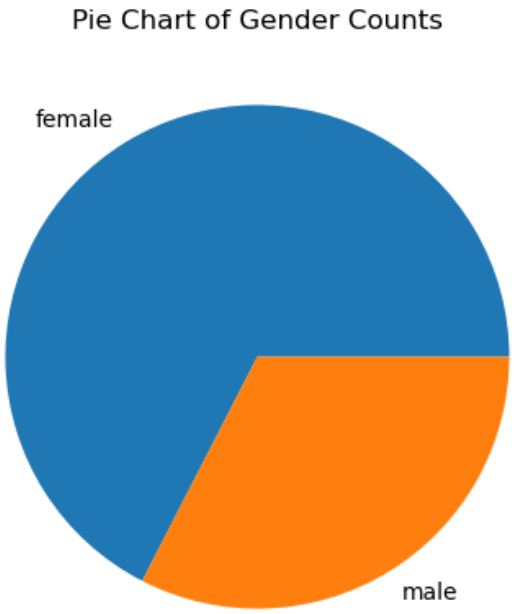
กราฟแสดง จำนวน ข้อมูลของมะเร็งในแต่ละประเภท ซึ่งประกอบด้วย

1. KICH - Kidney Chromophobe
2. ACC - Adrenocortical Carcinoma
3. BLCA - Bladder Urothelial Carcinoma
4. BRCA - Breast Invasive Carcinoma
5. CESC - Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
6. ESCA - Esophageal Carcinoma
7. LAML - Acute Myeloid Leukemia
8. LGG - Brain Lower Grade Glioma
9. OV - Ovarian Serous Cystadenocarcinoma
10. PAAD - Pancreatic Adenocarcinoma
11. PRAD - Prostate Adenocarcinoma
12. READ - Rectum Adenocarcinoma
13. TGCT - Testicular Germ Cell Tumors
14. THCA - Thyroid Carcinoma
15. LUAD - Lung Adenocarcinoma
16. UCEC - Uterine Corpus Endometrial Carcinoma
17. COAD - Colon Adenocarcinoma
18. SKCM - Skin Cutaneous Melanoma

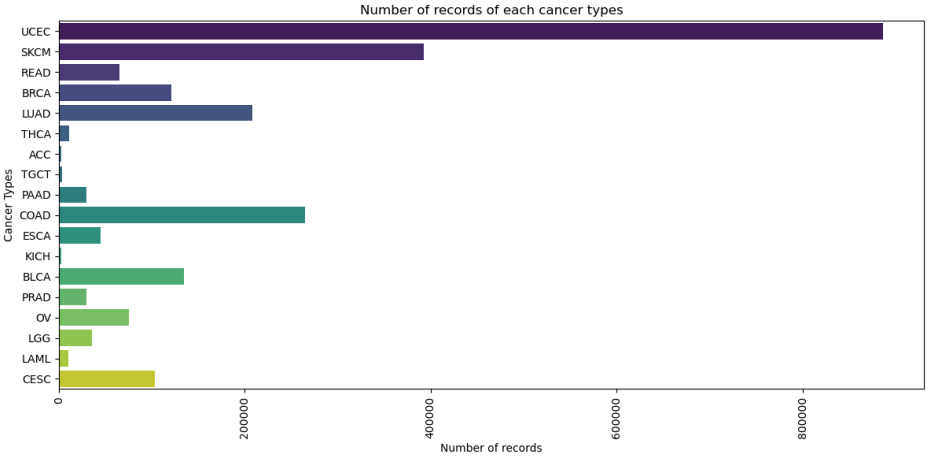





จากกราฟวงกลม เราสามารถสรุปได้ว่า **ส่วนใหญ่ของ Variant ที่พบในข้อมูลชุดนี้เป็นประเภท SNP** หรือการเปลี่ยนแปลงของเบสเพียงตัวเดียว ขณะที่ Variant ประเภท DEL และ INS มีสัดส่วนที่น้อยกว่ามาก



จากกราฟวงกลม จะเห็นได้ว่าข้อมูลที่เก็บมาจะมีจำนวนเพศหญิงมากกว่าผู้หญิง โดยสัดส่วนของผู้หญิงเกือบ ⅔ ของประชากรทั้งหมดที่สำรวจ




จากกราฟแท่ง ได้ว่า มะเร็งที่มีผู้ป่วยมากที่สุดคือ มะเร็งผิวหนังชนิดเมลาโนมา (SKCM) มีจำนวนผู้ป่วยสูงที่สุดและ มะเร็งที่มีผู้ป่วยน้อยที่สุดคือ มะเร็งปากมดลูก (CESC) มีจำนวนผู้ป่วยน้อยที่สุดเมื่อเทียบกับชนิดมะเร็งอื่นๆ

1. การฝึก machine learning model

   โดย model ที่ใช้ฝึกได้แก่

   1. Random Forest
   2. XGBoost
   3. LightGBM
   4. CatBoost
   5. Multi-Layer Perceptron

2. การประเมินประสิทธิภาพของโมเดล:

   ใช้ตัวชี้วัดได้แก่ **Accuracy, Precision, Recall และ F1-Score** เพื่อเปรียบเทียบความแม่นยำและความสามารถของโมเดลในการทำนายผลลัพธ์

## Progress Report #3

ในขั้นตอนนี้ของโครงการ เราได้ขยายขนาดของชุดข้อมูลเพื่อปรับปรุงประสิทธิภาพและความแม่นยำของโมเดล จากเดิมที่เราจำกัดการวิเคราะห์ข้อมูลเพียง 50% ของข้อมูลจากโครงการ the Cancer Genome Atlas (TCGA) project conducted by the U.S. National Institutes of Health (NIH) โดยในรายงานนี้เราจะสรุปผลลัพธ์ที่ได้หลังจากขยายขนาดข้อมูลเป็น 100% ซึ่งรวมทั้งหมดประมาณ 2,400,000 record

วัตถุประสงค์หลักคือการประเมินว่าการเพิ่มขนาดชุดข้อมูลจะช่วยให้ประสิทธิภาพของโมเดลต่างจากเดิมไหมในแง่ของ Accuracy, Precision, Recall และค่า F1-score หรือไม่ โดยเราได้ทำการทดลองใช้ model ต้นแบบ ได้แก่ Random Forest และ XGBoost

| model | accuracy | precision | recall | f1-score |
|---|---|---|---|---|
| Random Forest | 0.9844 | 0.8025 | 0.8427 | 0.8172 |
| XGBoost | 0.9953 | 0.8883 | 0.9005 | 0.8896 |

**การเปรียบเทียบโมเดล**:

- **Random Forest**: โมเดลนี้มีค่าความแม่นยำอยู่ที่ 98.44% โดยมีการปรับปรุงในด้าน Precision (80.25%), Recall (84.27%) และ F1-score (81.72%)
- **XGBoost**: โมเดลนี้มีประสิทธิภาพสูงกว่าทุกด้าน โดยมีค่าความแม่นยำอยู่ที่ 99.53% และค่า Precision (88.83%), Recall (90.05%), และ F1-score (88.96%) ที่สูงกว่า Random Forest อย่างมาก ซึ่งแสดงให้เห็นว่า XGBoost สามารถใช้ประโยชน์จากข้อมูลที่มากขึ้นได้อย่างมีประสิทธิภาพ

## Final Report

จาก Progress Report #3, เราได้เพิ่มการทดลองด้วยโมเดล Machine Learning อีก 3 โมเดล ประกอบด้วย MLP, LightGBM, Catboost ได้ผลลัพธ์ดังตาราง

| Final Results | | | | |
|---|---|---|---|---|
| model | accuracy | precision | recall | f1-score |
| Random Forest | 0.9844 | 0.8025 | 0.8427 | 0.8172 |
| XGBoost | 0.9953 | 0.8883 | 0.9005 | 0.8896 |
| MLP | 0.9565 | 0.9576 | 0.9565 | 0.9558 |
| LightGBM | 0.9984 | 0.9351 | 0.9369 | 0.9350 |
| Catboost | 1 | 0.93 | 0.93 | 0.93 |

Key Insights:

- LightGBM มีความแม่นยำสูงสุด (0.9984) แสดงให้เห็นว่ามีประสิทธิภาพสูงสุดในการทำนายเมื่อเทียบกับโมเดลอื่น ๆ
- MLP มีค่า precision สูงสุด (0.9576) และ recall สูงสุด (0.9565) ซึ่งแสดงถึงประสิทธิภาพในการระบุกรณีบวกได้อย่างถูกต้อง
- XGBoost ทำได้ดีในทุกตัวชี้วัด โดยมีค่า f1-score ค่อนข้างสูง (0.8896)
- Random Forest มีค่า f1-score ต่ำสุด (0.8172) แสดงถึงการแลกเปลี่ยนระหว่าง precision และ recall

โดยสรุป LightGBM ดูเหมือนจะเป็นโมเดลที่ทำงานได้ดีที่สุดโดยรวม โดยเฉพาะในแง่ของความแม่นยำและประสิทธิภาพที่สมดุลในทุกตัวชี้วัด ส่วน MLP โดดเด่นในด้าน precision และ recall และ XGBoost มีความสมดุลในทุกตัวชี้วัด

## อภิปรายและสรุปผล

โครงการนี้มีเป้าหมายในการพัฒนาโมเดลทำนายความเป็นไปได้ในการเกิดโรคมะเร็งจากข้อมูลจีโนม (genomic data) โดยอาศัยเทคนิค Machine Learning ซึ่งมีการใช้งานที่หลากหลาย โมเดลต่าง ๆ เช่น Random Forest, XGBoost, MLP (Multi-Layer Perceptron), LightGBM, และ CatBoost ได้รับการนำมาทดลองใช้เพื่อเปรียบเทียบประสิทธิภาพในการทำนายโรคมะเร็ง โดยเน้นการวัดประสิทธิภาพผ่านค่าต่าง ๆ ได้แก่ Accuracy, Precision, Recall และ F1-score

### ขั้นตอนและผลการทดลอง

1. การเตรียมข้อมูล

ข้อมูลที่ใช้มาจาก The Cancer Genome Atlas (TCGA) ซึ่งรวบรวมข้อมูลจีโนมจากกลุ่มผู้ป่วยมะเร็งหลากหลายชนิดและกลุ่มควบคุม (กลุ่มที่ไม่เป็นมะเร็ง) ข้อมูลมีขนาดใหญ่และมีจำนวนคุณลักษณะสูง ซึ่งทำให้ต้องใช้การเลือกฟีเจอร์ (feature selection) เพื่อลดจำนวนฟีเจอร์ที่ไม่จำเป็น เช่น การใช้เทคนิค SelectKBest เพื่อลดจำนวนฟีเจอร์และเพิ่มประสิทธิภาพของโมเดล

2. การฝึกโมเดล Machine Learning

โมเดลที่ใช้ในการทดสอบ ได้แก่ Random Forest, XGBoost, LightGBM, CatBoost, และ Multi-Layer Perceptron (MLP) แต่ละโมเดลถูกฝึกฝนเพื่อทำการทำนายข้อมูลจีโนมว่าผู้ป่วยมีแนวโน้มเป็นมะเร็งหรือไม่ โดยมีการปรับพารามิเตอร์ของโมเดลและใช้เทคนิค Cross-Validation เพื่อเพิ่มความน่าเชื่อถือ

3. การประเมินผลลัพธ์

ผลการทดลองสรุปในตารางที่มีการเปรียบเทียบค่าของ Accuracy, Precision, Recall และ F1-score ของโมเดลต่าง ๆ พบว่า:

- LightGBM มีค่าความแม่นยำสูงสุดที่ 99.84% ซึ่งแสดงถึงความสามารถในการทำนายโรคมะเร็งได้แม่นยำที่สุดเมื่อเทียบกับโมเดลอื่น ๆ
- MLP มีค่า Precision สูงสุดที่ 95.76% และ Recall ที่ 95.65% ซึ่งแสดงให้เห็นว่าโมเดลนี้เหมาะสำหรับการทำนายกรณีบวก (เช่น กรณีที่พบว่ามีความเสี่ยงสูงต่อการเป็นมะเร็ง)
- XGBoost มีค่า F1-score สูงและมีความสมดุลในตัวชี้วัดทั้งหมด ทำให้เป็นโมเดลที่สามารถทำนายได้ดีในหลากหลายบริบทและให้ผลการทำนายที่สมดุล
- Random Forest มีค่า F1-score ต่ำที่สุดที่ 81.72% ซึ่งแสดงให้เห็นถึงการแลกเปลี่ยนระหว่าง Precision และ Recall ในการทำนาย

**การอภิปรายและข้อสรุป**

จากการทดสอบโมเดลต่าง ๆ พบว่า LightGBM เป็นโมเดลที่มีความแม่นยำสูงสุด ทำให้เป็นตัวเลือกที่ดีที่สุดสำหรับการทำนายมะเร็งจากข้อมูลจีโนมโดยรวม อย่างไรก็ตาม MLP มีความโดดเด่นในเรื่องของ Precision และ Recall ซึ่งเหมาะสมสำหรับการทำนายกรณีบวก ส่วน XGBoost แสดงความสามารถที่สมดุลในทุกตัวชี้วัด ทำให้เหมาะสำหรับการใช้งานทั่วไป

การศึกษาแสดงให้เห็นถึงศักยภาพของการใช้ Machine Learning ในการวิเคราะห์ข้อมูลจีโนมสำหรับการทำนายโรคมะเร็ง ซึ่งสามารถนำไปพัฒนาเป็นเครื่องมือช่วยแพทย์ในการวินิจฉัยและการตัดสินใจในการรักษาโรคเฉพาะบุคคล

## เอกสารอ้างอิง


1. โปรเจคต้นแบบ

https://medium.com/@shuv.sdr/cancer-prediction-from-genomic-analysis-with-machine-learning-2c957b579f05

2. ฐานข้อมูลที่นำมาใช้ในการ train model

https://genome.ucsc.edu/cgi-bin/hgTables