Dr. AI

Machine Learning for Cancer Prediction

Avinash Lal, Raahil Sha, Julio Mendez Cabrera

Background

Cancer

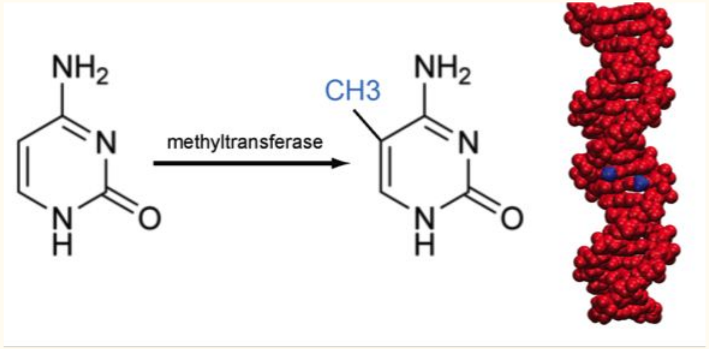
- General term for uncontrolled cell growth
- Not one disease - a category over a broad range of diseases
- Cancer cells have very diverse origins, making cancer highly heterogeneous at the population level
- Tumor Heterogeneity - different tumor cells can show very distinct genetic, epigenetic, and phenotypic profiles
- Research into cancer heterogeneity could help inform treatment strategies to improve survivability

Data

- Because cancer occurs at a late developmental stage, it is highly linked to developmental control mechanisms and epigenetics
- The Cancer Genome Atlas (TCGA) - NIH Project to catalogue bioinformatics data on a broad range of cancers (currently ~35 different cancers)
- Two Main Epigenetic Factors
    - Methylation
    - RNA

Methylation

- Methylation is the addition of a methyl group to a cytosine base
- Excessive methylation can cause transcription to fail (key proteins don’t bind or fall off of the DNA strand)
- Methylation of certain tumor suppressor genes, like p16 and p57, has been shown to play a role in cancer growth



- Profiled using Illumina 450k technology, which measures methylation percent of 450,000 locii in the DNA
- TCGA has 9,756 samples (*1.38 terabytes*) - too large to quickly download so we didn’t study methylation

Ribonucleic Acid (RNA)

- When a cell creates proteins, it first takes DNA and transcribes it into RNA
- The RNA then exits the nucleus and get translated into proteins
- Profiled using RNA-Seq technology, which relates RNA content of a cell for around 60,000 locii
- Measured using FPKM (Fragments per Kilobase of Transcript per Million Mapped Reads) - relative, normalized RNA expression level
- TCGA has 11,574 samples (6.2 gigabytes compressed) - data is a lot less to download and easier to work with and manipulate
- We chose to use RNA-Seq data in our project

Question

- ***Our Project:*** **Given an epigenetic profile of a cell, is it possible to predict if the cell has cancer? Going further, is it possible to predict the specific type of cancer that the cell has?**

Data and Methods

RNA-Seq Dataset (Post-process)

Transcript	e6bb1330-c761-43a5-9d17-16da727232f1.FPKM.txt	1a13663e-9015-4eeb-ab6a-8f40a8bbf403.FPKM.txt	50f3b3d3-dba8-4d48-bf8d-16d7ebaa18fd.FPKM.txt	b1558748-bec9-4bea-a35b-040ca9a1f4cd.FPKM.txt
ENSG00000242268.2	0	0.05955585	0.369784719	0
ENSG00000270112.3	0	0.008263563	0	0.005879084
ENSG00000167578.15	5.301041137	1.498275119	10.0475245	4.855961155

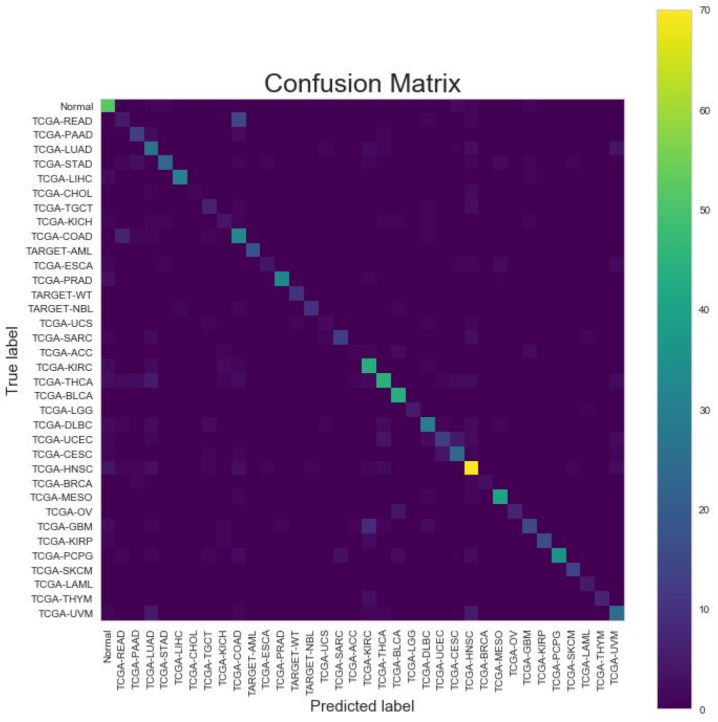
Sample Label File

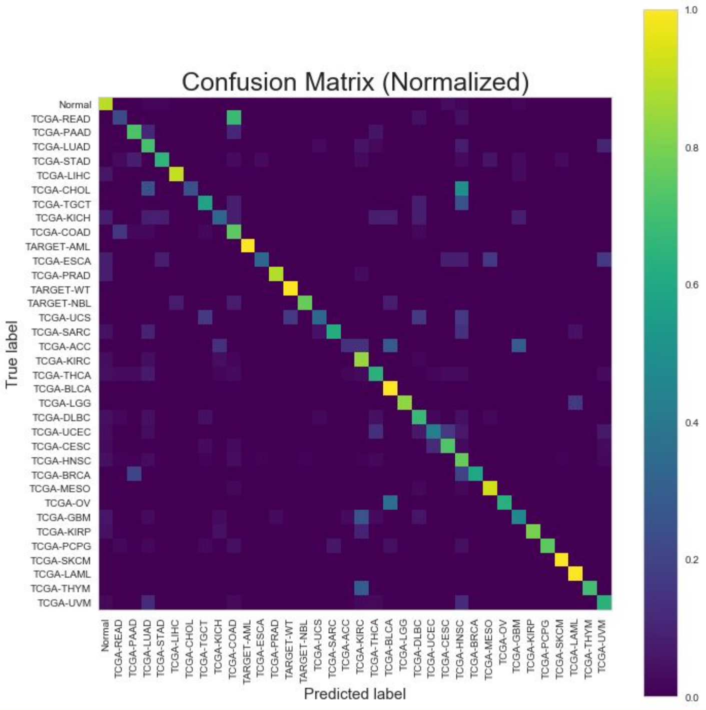
file_name	cases.0.samples.0.sample_type	cases.0.project.project_id
7a494c60-48a3-486a-83c2-aefb4c160a2c.FPKM.txt	Primary Tumor	TCGA-BRCA
33fef46f-c248-4d58-bb6e-3a4a55272334.FPKM.txt	Primary Tumor	TCGA-BRCA
0d44ebee-ebdf-442b-82ce-ce78b8b0afb4.FPKM.txt	Primary Tumor	TCGA-UCEC
d65a7cb6-4b1f-4976-990e-c4f47ae11986.FPKM.txt	Primary Tumor	TCGA-COAD
e328b8e9-2ffa-4cca-af04-5a96f09b9e04.FPKM.txt	Blood Derived Normal	TCGA-ESCA
89105180-9922-43bc-81a9-7b43b11e5e20.FPKM.txt	Primary Tumor	TCGA-LUAD

Estimator Evaluation

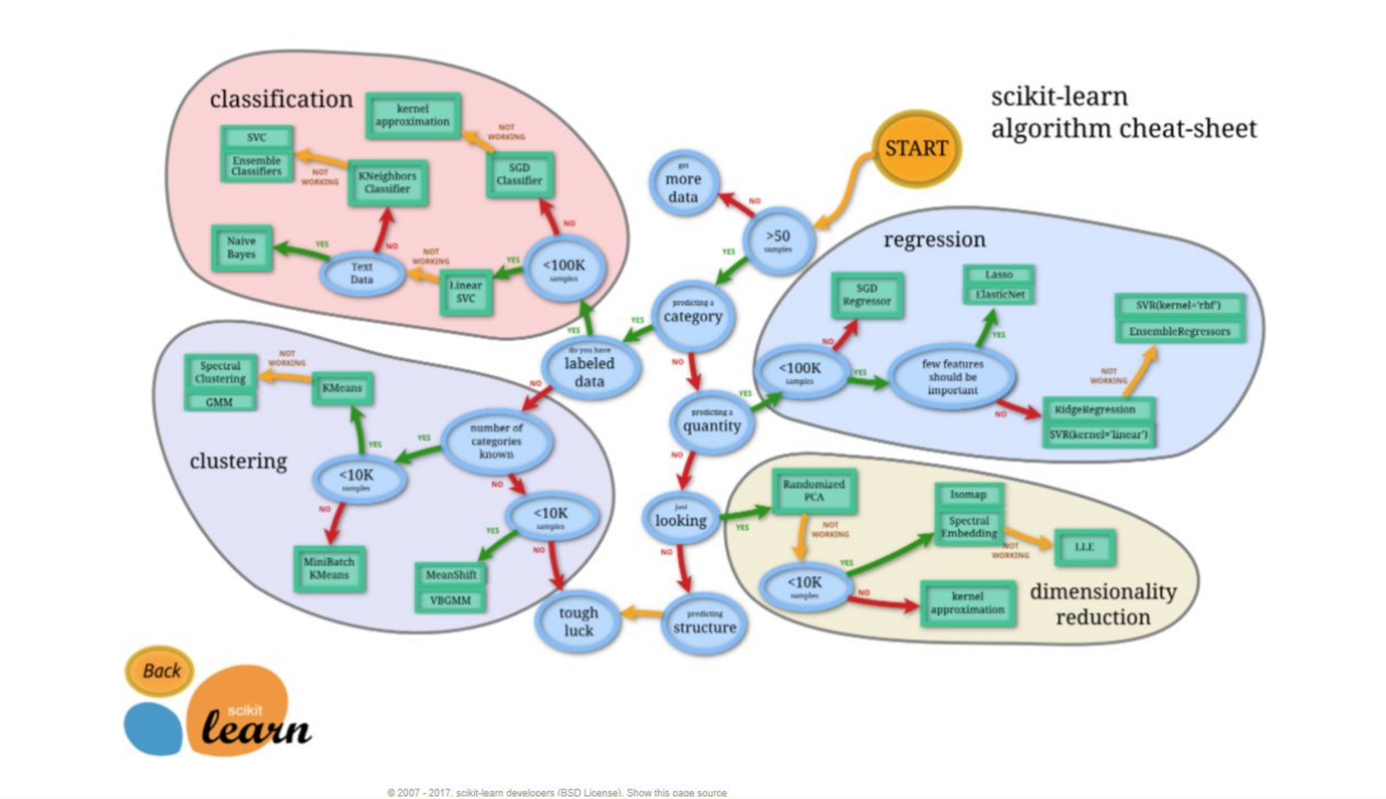
- Accuracy - How many predictions were accurate?
- Precision - How many predictions that are in-class were accurate?
    - True Positive / (True Positive + False Positive)
- Recall - How accurate was prediction on in-class data?
    - True Positive / (True Positive + False Negative)
- F1 Score - Combination of precision and recall into one metric
- All of these can be evaluated either per-class or over all classes

Confusion Matrix (Accuracy Matrix)





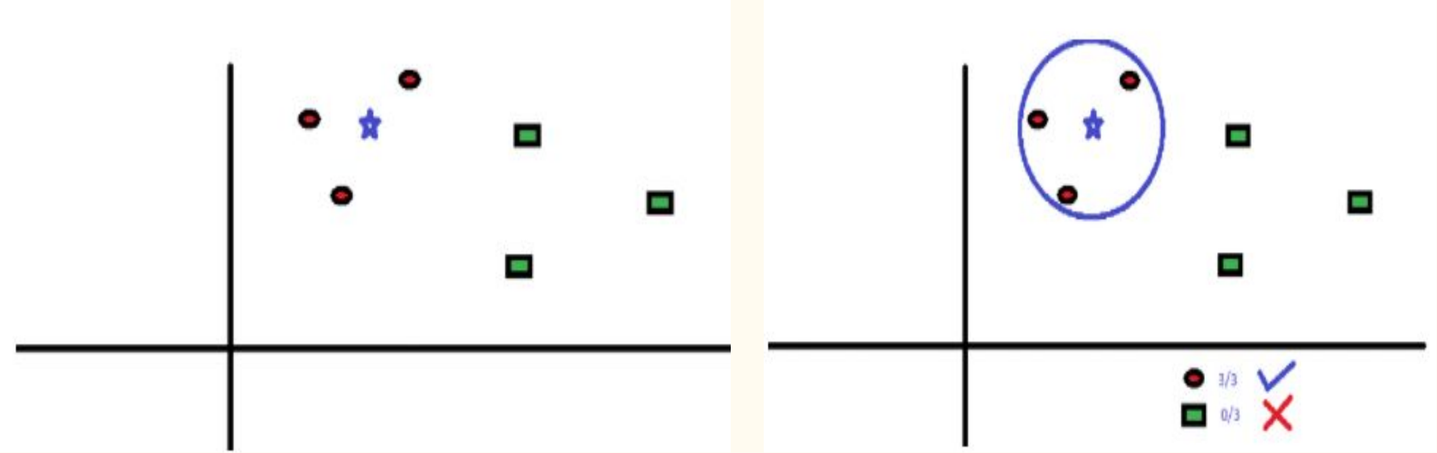
Estimator Roadmap



k-Nearest Neighbor Classification

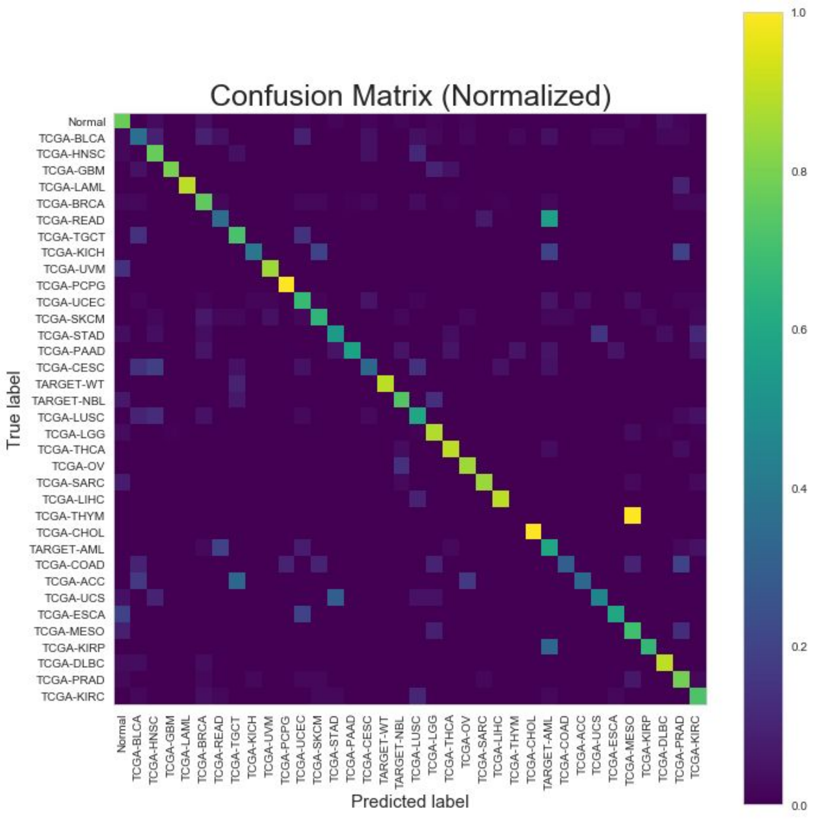
kNN Algorithm

Simple example:



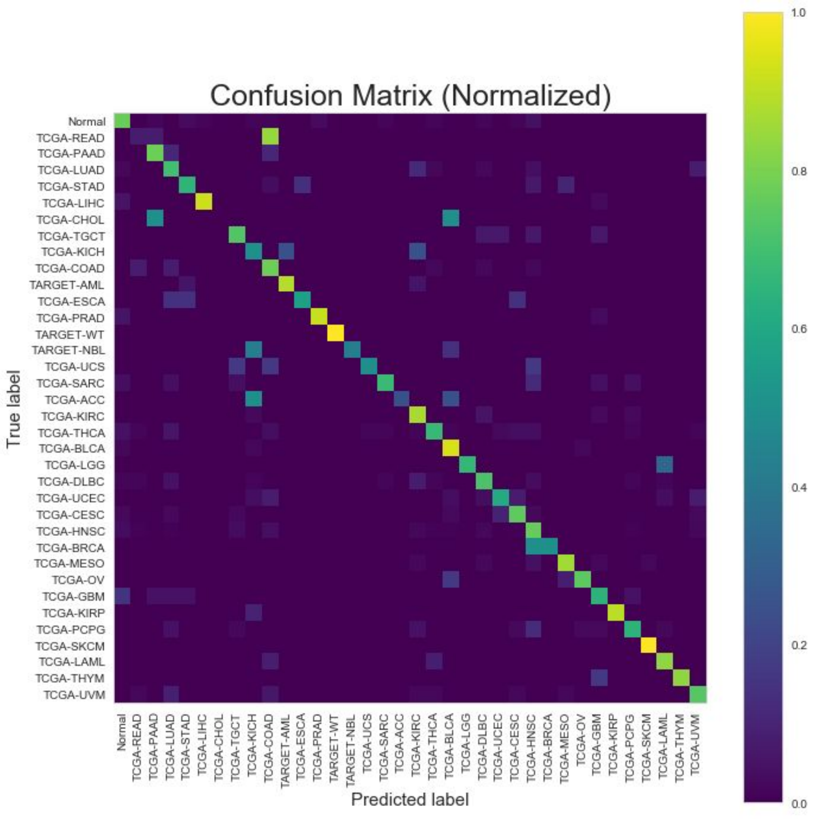
Results (kNN, n = 1)

- Accuracy: 0.717
- Recall: 0.717
- Precision: 0.725120707634



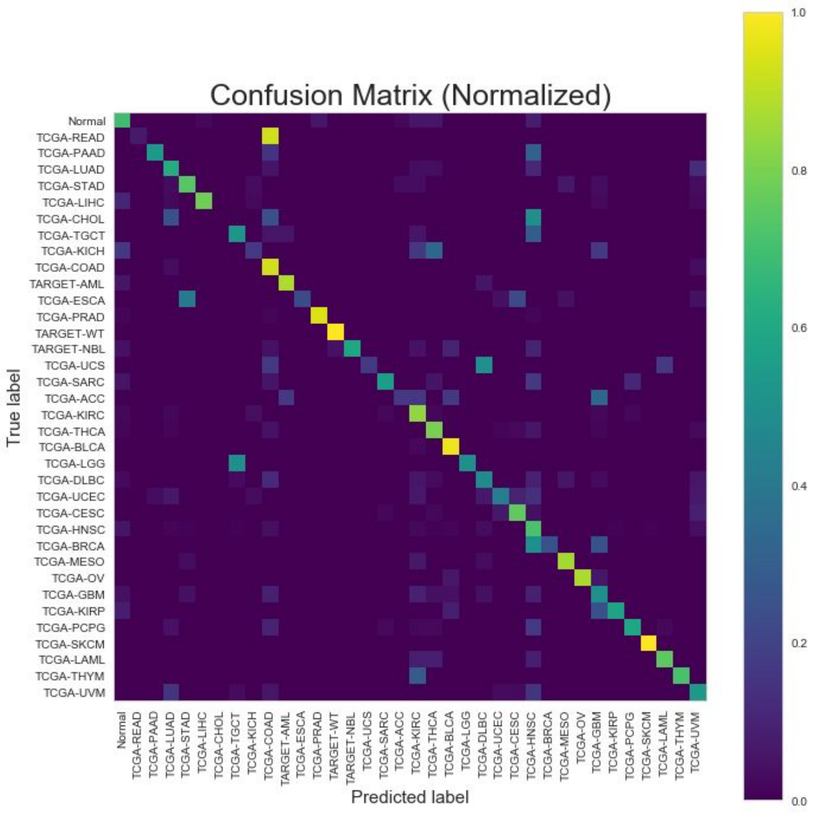
Results (kNN, n = 10)

- Accuracy: 0.752
- Recall: 0.752
- Precision: 0.76809307558

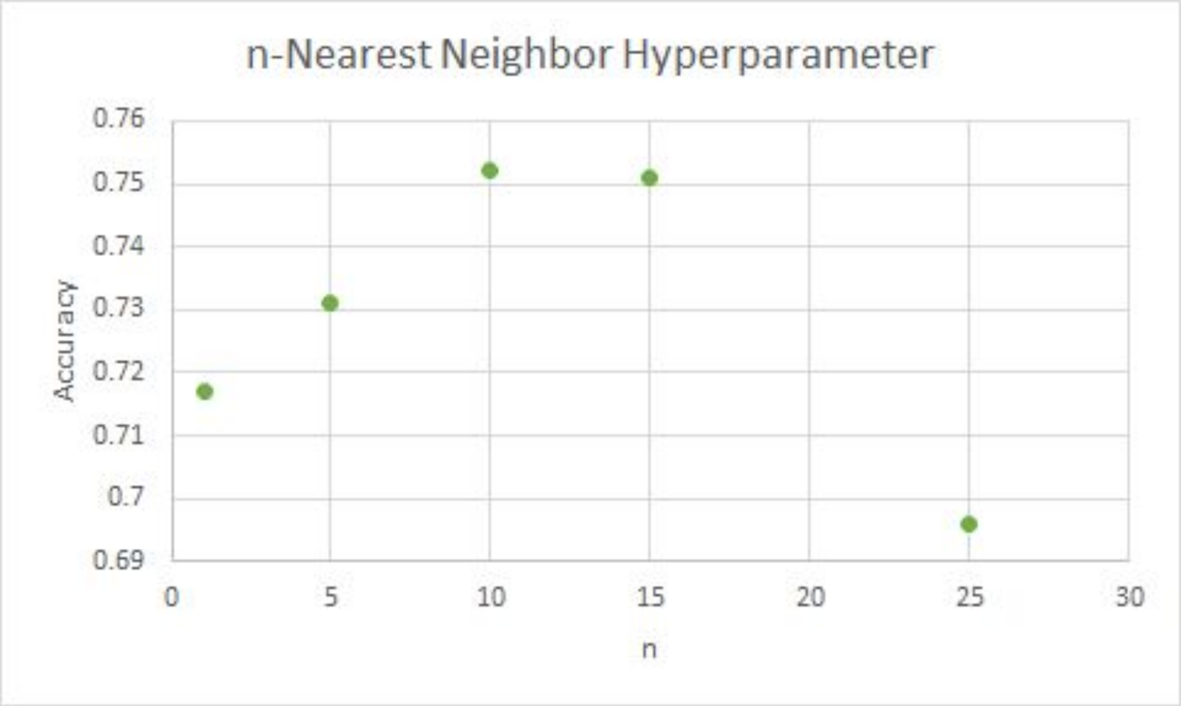


Results (kNN, n = 25)

- Accuracy: 0.696
- Recall: 0.696
- Precision: 0.734013882285

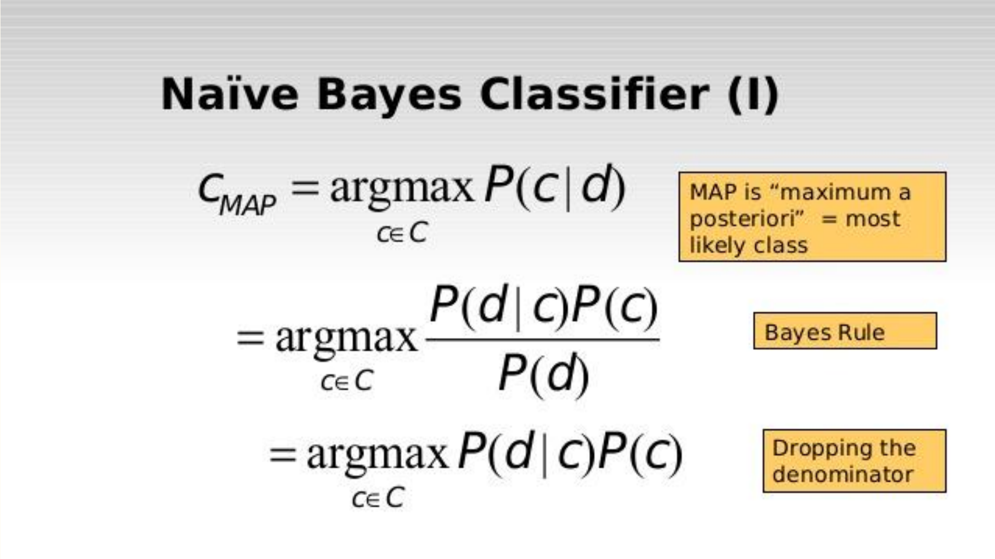


kNN Hyperparameter Tuning



Naive Bayes Classification

Naive Bayes Algorithm



Naïve Bayes Classifier (I)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

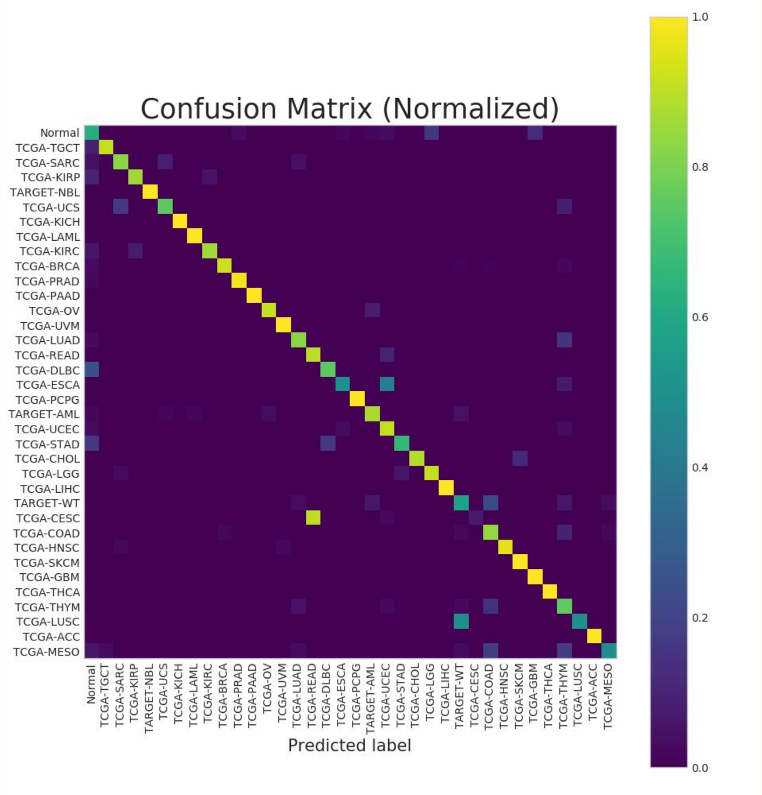
Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

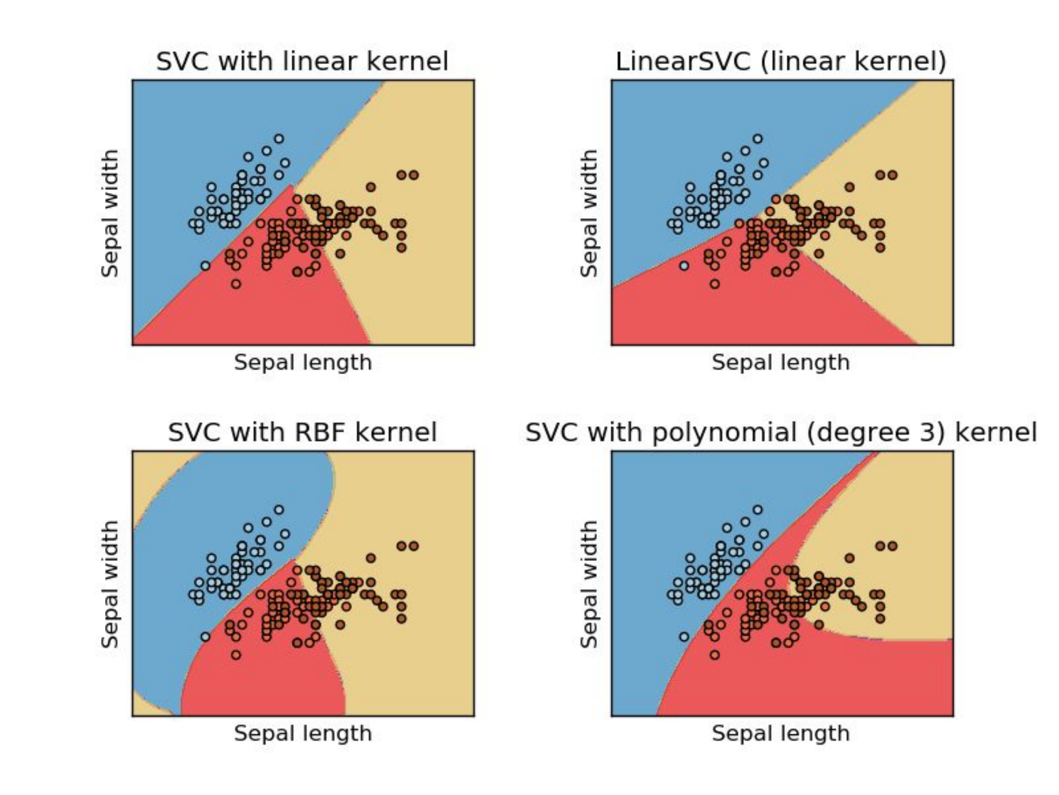
Gaussian Naive Bayes Results

- Accuracy: 0.837
- Recall: 0.837
- Precision: 0.871



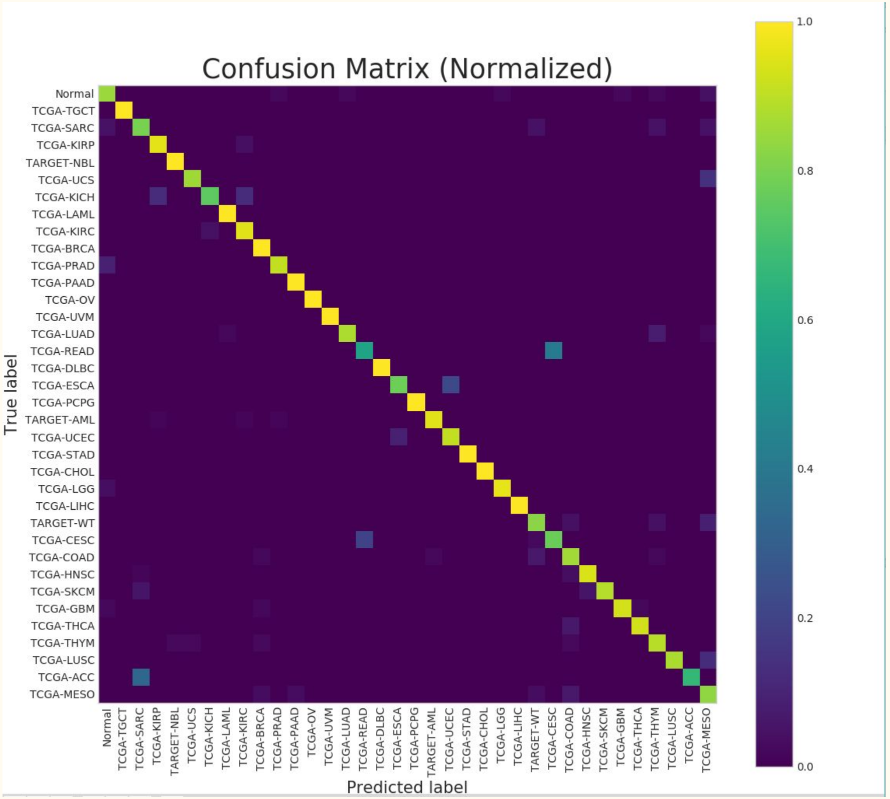
Support Vector Classification

SVC Algorithm



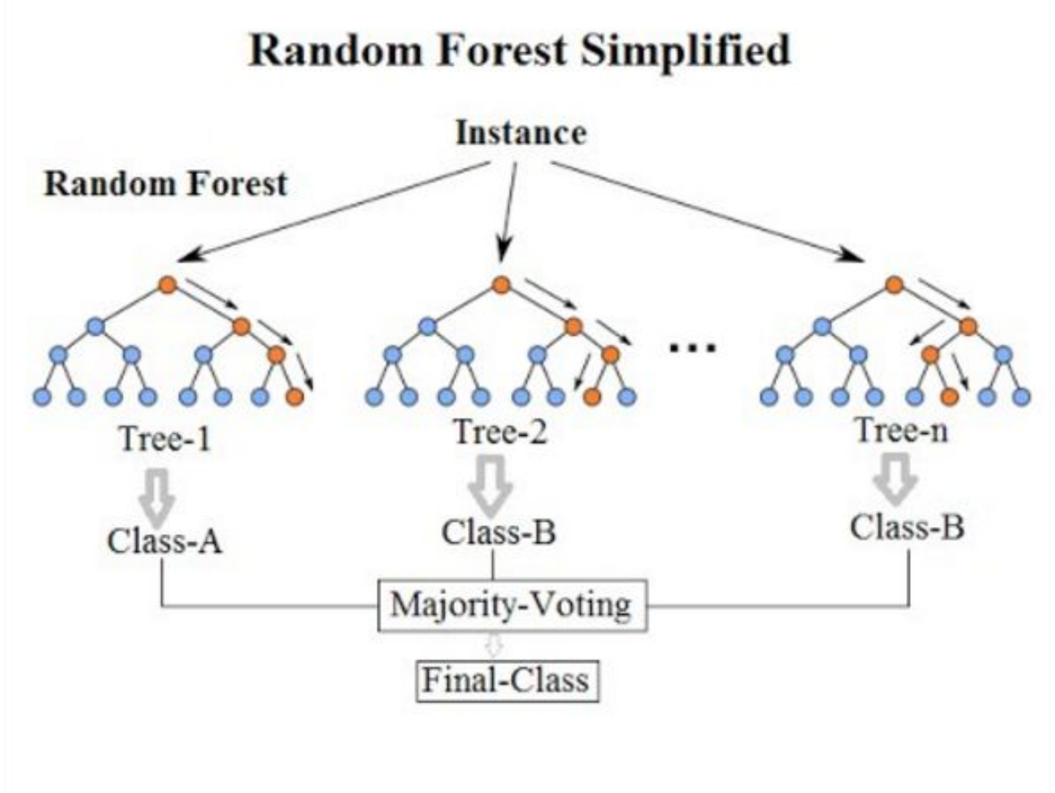
SVC Results

- Accuracy: 0.926
- Recall: 0.926
- Precision: 0.9265



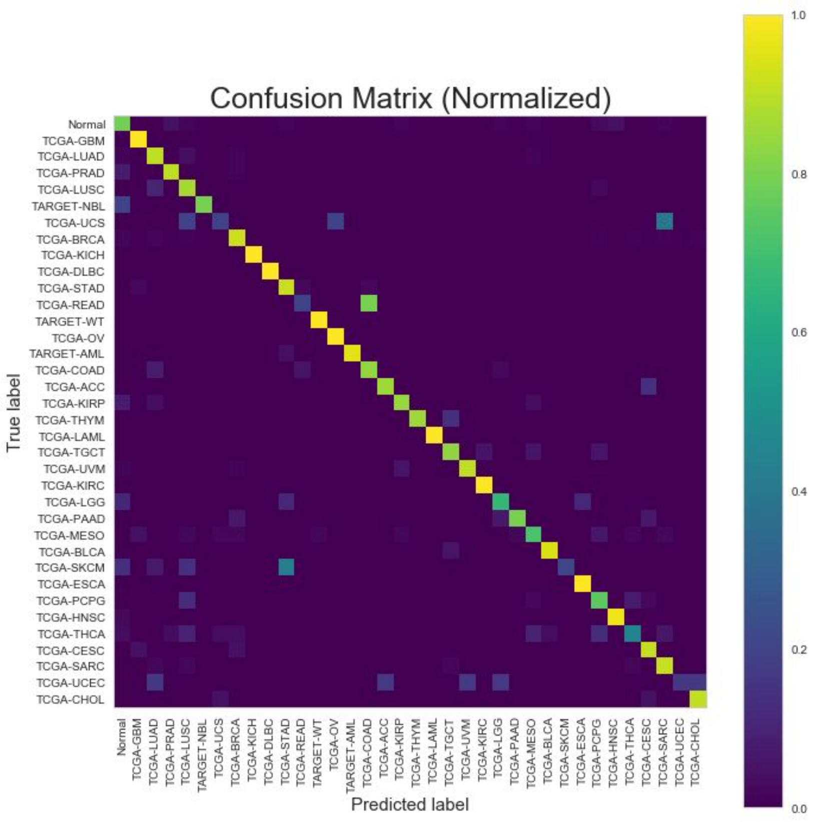
Random Forest Classification

Random Forest Algorithm



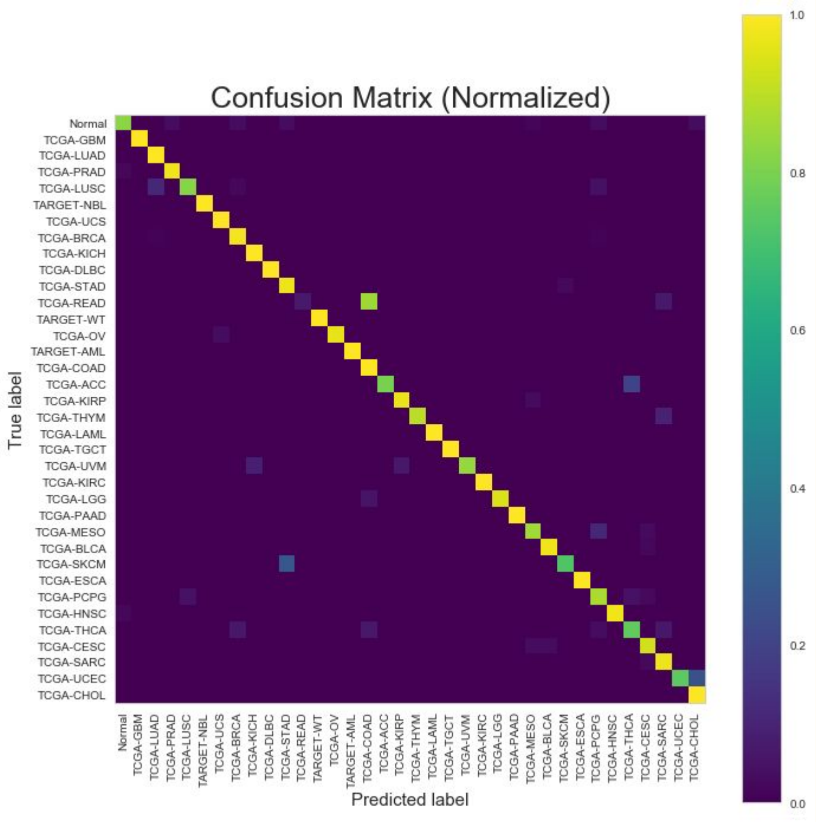
Results (Random Forest, n = 5)

- Accuracy: 0.848
- Recall: 0.848
- Precision: 0.852776

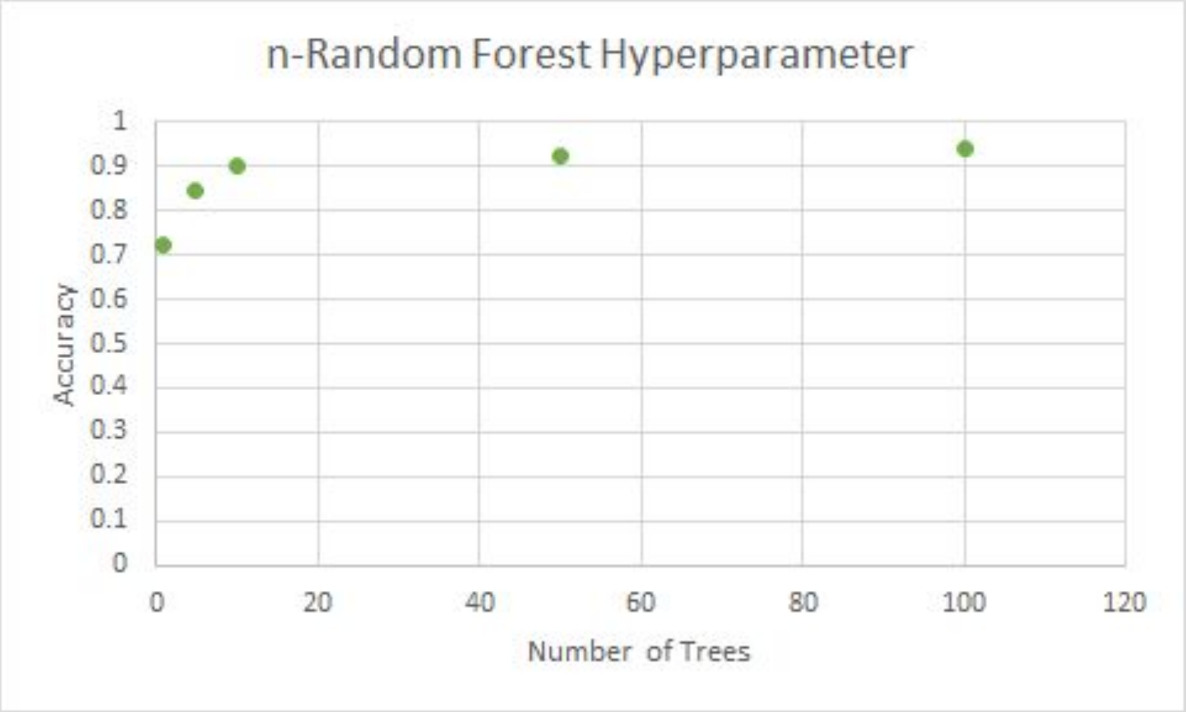


Results (Random Forest, n = 50)

- Accuracy: 0.926
- Recall: 0.926
- Precision: 0.93328



Choosing Number of Trees



Analysis and Conclusions

Classifier Model Comparison

Model	Accuracy	Time to Train	Time to Predict
k-Nearest Neighbors	75.2%	75 seconds	277 seconds
Naive Bayes	83.7%	98 seconds	20 seconds
Support Vector	92.6%	1758 seconds	315 seconds
Random Forest (50)	92.6%	33 seconds	0.32 seconds
Random Forest (100)	94%	62 seconds	0.36 seconds
Ensemble	N/A	Broke our computer	---

Future Work

- Higher-level Ensemble Methods
  - Have a group of 4-5 classifiers “vote” on the final classification to try to overcome individual classifier error
- Alternative Classification Models
  - Deep learning, neural networks?
- Evaluation of Performance on other data types
  - Methylation - Dataset was too large so we didn’t use methylation data
  - DNA - Can we predict if a person has cancer from DNA of a non-cancerous cell?
- Estimation of Cancer Dangerousness
  - How long does a person have to live given their epigenome?
  - Will a group of stationary cancer cells metastasize?